UNIVERSITY OF CALIFORNIA
Los Angeles

Novel Techniques for Automated Medical Image Segmentation Spanning Diverse Anatomical Structures and Imaging Modalities

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Noor Nakhaei

2025

ABSTRACT OF THE DISSERTATION

Novel Techniques for Automated Medical Image Segmentation Spanning Diverse Anatomical Structures and Imaging Modalities

by

Noor Nakhaei
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2025
Professor Demetri Terzopoulos, Chair

First, we introduce a framework for improving spatial correlation between mammography and specimen radiography, enhancing diagnostic accuracy and surgical guidance in breast cancer treatment.

Second, we develop hybrid architectures that synergistically combine Active Contour Models (ACMs) with Convolutional Neural Networks (CNNs), demonstrating superior boundary detection performance compared to either approach used in isolation. Third, we formulate differentiable implementations of traditionally fixed ACM parameters, enabling end-to-end training within neural networks and improving adaptability across diverse imaging contexts. Fourth, we design boundary-aware transformer models with specialized attention mechanisms that prioritize accurate delineation of anatomical borders, addressing limitations in standard transformer architectures for medical segmentation tasks. Finally, we integrate human attention patterns derived from eye-tracking studies of expert radiologists into computational models, creating systems that achieve higher accuracy and generate explanations that align with clinical reasoning.

Empirical evaluations across multiple anatomical structures, imaging modalities, and clinical tasks demonstrate that these approaches significantly enhance technical performance and clinical relevance. The hybrid methods show improved generalization across different institutions and imaging protocols, while the human-aligned attention mechanisms facilitate

interpretability and clinical acceptance.

Our research advances the field toward medical image analysis systems that serve as reliable partners in clinical decision-making by combining the mathematical rigor of traditional approaches, the representational power of deep learning, and the domain expertise of clinical practitioners.

The dissertation of Noor Nakhaei is approved.

Bolei Zhou

Achuta Kadambi

William Hsu

Demetri Terzopoulos, Committee Chair

University of California, Los Angeles

2025

To my mom, Giti Amidpour, who gave me all she had,

and to my dad, Mohammad Nakhaei, who has always been my greatest champion.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply grateful to my dissertation committee for their guidance and support:
— Professor Demetri Terzopoulos, thank you for your unwavering support throughout these years.
— Professor William Hsu, thank you for your mentorship and the countless hours you invested in my growth. — Professors Achuta Kadambi and Bolei Zhou, thank you for your suggestions and insights for the thesis.

I also thank Professor Anne Hoyt for her insights and for teaching me about patients and the real-world problems we aim to solve.

Tomis, thank you for your support and mentorship during my first two years. Your patience, knowledge, and guidance were invaluable as I found my footing in this program.

I want to acknowledge the amazing students with whom I have had the privilege to work: Elaine Steinberg, Allen Luo, and Olivia Zhang, Tanmay Sanjay Hukkeri, Shaira Alam, and Vaishnavi Manthena ; and my wonderful labmates Yannan Lin, Ruiwen (Rina) Ding, Luoting (Lottie) Zhuang, Yunzheng Zhu, Kimaya Kulkarni, Bing Zhu, Anil Yadav, and Stephen Park.

Thank you to my co-authors and collaborators: Chrysostomos Marasinou, Akinyinka Omigbodun, Nina Capiro, Bo Li, Jeremy Paige, Arvind Vepa, Andrew Choi, and Wonjun Lee. Your expertise and collaboration enriched this work immeasurably.

To the Computer Science graduate office, especially Joseph Brown—I could not have finished this PhD without your support. Thank you! To Juliana Alvarez, Madelen Hem, and Diana Villegas, you are the beacons of hope and support in the department.

To my closest friends who became family and made LA home away from home—Emma, Mura, Tanvi, Haein, Brooke, Ali, Ale, and Filo—I truly would not have survived without you. Your friendship sustained me through every single crazy chapter.

To Stéphane, Suhail, Zeina, Shuyang, Emily, Ana, Christina, Elia, and Madhu—you made the CS department feel like home and gave me a true sense of community. Thank you for making this journey so much brighter.

Jay, Spencer, and Jackson, you were with me through all the madness that student government brought us! Thank you for helping me survive the three mad mad years.

János, thank you for your endless patience, love, and support. You've stood by me through all my moods this past year and finishing this thesis—I couldn't have asked for a better partner by my side.

Finally, to my parents, words cannot fully express the depth of my gratitude to you. Mom, thank you for always being available when I got homesick or frustrated with my model's performance, despite the 11.5-hour time difference. Your love and support has been carrying me throughout my life.

VITA

2013–2017	BS (Computer Engineering), Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
2019–2022	MS (Computer Science), UCLA, Los Angeles, CA
2019–	PhD Candidate (Computer Science), UCLA, Los Angeles, CA
2019–Present	Graduate Student Researcher, Computer Science Department, UCLA, Los Angeles, CA
2020–2023	Teaching Fellow, Computer Science Department, UCLA, Los Angeles, CA
2023	Head Teaching Fellow and Instructor, Computer Science Department, UCLA, Los Angeles, CA

PUBLICATIONS

N. Nakhaei, T. Zhang, D. Terzopoulos, and W. Hsu, "Refining boundaries of the segment anything model in medical images using an active contour model," in *Proc. SPIE 12927, Medical Imaging 2024: Computer-Aided Diagnosis*, April 2024, Online, pgs. 749–758.

N. Konz, et al., **N. Nakhaei**, et al., and Maciej A. Mazurowski, "A Competition, Benchmark, Code, and Data for Using Artificial Intelligence to Detect Lesions in Digital Breast Tomosynthesis," *JAMA Network Open*, **6**(2):e230524, February 2023.

W. Hsu, D. S. Hippe, **N. Nakhaei**, et al., and Christoph I. Lee, "External Validation of an Ensemble Model for Automated Mammography Interpretation by Artificial Intelligence," *JAMA Network Open*, **5**(11):e2242343, November 2022.

A. Vepa, A. Choi, **N. Nakhaei**, et al., and F. Scalzo, "Weakly-Supervised Convolutional Neural Networks for Vessel Segmentation in Cerebral Angiography," in *Proc. IEEE International Conference on Applications in Computer Vision (WACV 2022)*, January 2022, Waikoloa, Hawaii, pgs. 585–594.

K. Marathe, C. Marasinou, B. Li, **N. Nakhaei**, B. Li, J. G. Elmore, L. Shapiro, and W. Hsu, "Automated quantitative assessment of amorphous calcifications: Towards improved malignancy risk stratification," *Computers in Biology and Medicine*, **146**:105504, June 2022.

N. Nakhaei, C. Marasinou, A. Omigbodun, N. Capiro, B. Li, A. Hoyt, and W. Hsu, "Spatial Matching of Magnified 2D Mammography Images and Specimen Radiographs: Towards Improved Characterization of Suspicious Microcalcifications," in *Proc. SPIE 11597, Medical Imaging 2021: Computer-Aided Diagnosis*, February 2021, pgs. 511–516.

C. Marasinou, B. Li, J. Paige, A. Omigbodun, **N. Nakhaei**, A. Hoyt, and W. Hsu, "Improving the Quantitative Analysis of Breast Microcalcifications: A Multiscale Approach," *Journal of Digital Imaging*, **36**(3):1016–1028, June 2023.

CHAPTER 1

Introduction

Medical image analysis plays an important role in diagnosis and treatment. The process of analyzing and diagnosing medical images is costly and time-consuming. In recent years, there have been numerous efforts in implementing computer-aided diagnosis (CAD) systems to assist clinicians on important downstream tasks such as segmentation, classification, and risk stratification (Litjens et al., 2017). To this end, researchers have developed a large variety of algorithms. In particular, for the purposes of medical image segmentation, algorithms range from classic Active Contour Models (ACMs) (Kass et al., 1988) to modern deep learning-based approaches (Ronneberger et al., 2015). Despite promising advances, integrating computational image analysis into routine clinical practice faces substantial challenges related to algorithm robustness, interpretability, clinical validation, regulatory approval, and implementation within complex healthcare systems. Specifically, data-driven deep learning methods are insufficient to achieve adequate medical image analysis, as they typically require large annotated datasets, which are often impossible to achieve in practice. Moreover, the difference among the different datasets makes it impossible to train machine learning models on one set of data and infer adequately on a similar dataset from another institution.

This dissertation explores novel approaches to addressing these challenges by developing explainable deep learning models for medical image segmentation and classification. We focus on test-time domain-adaptation, segmenting boundaries, and building generalizable models.

1.1 Challenges

Despite the remarkable progress in medical image analysis, several critical challenges that are most relevant to the scope of this thesis continue to impede the effective translation of computational approaches into clinical practice.

Medical diagnosis frequently requires integrating complementary information from multiple modalities (e.g., CT, MRI, X-ray), each capturing different aspects of anatomy and physiology. Effectively fusing this information necessitates sophisticated registration techniques that can account for differences in resolution, contrast mechanisms, and acquisition parameters while preserving anatomical correspondence. Fusing specimen radiograph and pre-operative mammography improves orientation reliability and tumor localization (Drozgyik et al., 2025), reduces turnaround time and resource utilization (Arudra et al., 2021), and enhances radiology-pathology correlation (Nakhaei et al., 2021).

Accurate boundary detection and segmentation remain difficult, particularly in medical images where anatomical structures often exhibit low contrast, irregular boundaries, and pathological variations. The inherent complexity and variability of biological structures make it challenging for algorithms to precisely delineate organs, lesions, and tissues—a fundamental prerequisite for quantitative analysis, characterization, and computer-aided diagnosis.

Domain adaptation represents one of the most important challenges in the clinical deployment of AI systems. A fundamental disparity exists between radiologists and AI models' ability to generalize knowledge across different settings. Radiologists are less sensitive to differences in image appearance or differences in images across different institutions. While radiologists can effectively apply their expertise across institutions, imaging protocols, and patient populations with minimal adaptation, current AI systems exhibit performance degradation when faced with images from new environments—a phenomenon known as domain shift (Zech et al., 2018; Tayebi Arasteh et al., 2023). This necessitates resource-intensive fine-tuning for each new clinical setting, significantly limiting scalability and practical utility (Hsu et al., 2022).

The significant perceptual gap — the systematic difference between what AI models

"see" and what human experts deem salient — in performance between AI systems and radiologists further complicates integration into clinical workflows. Current AI models process images as collections of pixels with statistical patterns, fundamentally different from the anatomically-informed, experience-based reasoning of clinicians. This difference manifests in the regions of interest; AI often focuses on image features that differ substantially from those radiologists consider diagnostically relevant, potentially identifying correlations lacking meaningful clinical significance. Perhaps most critically, the "black box" nature of advanced AI systems undermines trust and hampers adoption. Clinicians require accurate predictions and transparent explanations that align with their medical knowledge. Without explainability, verifying whether a model's decision process is clinically sound or based on spurious correlations becomes impossible, creating significant barriers to responsible deployment and regulatory approval.

1.2 Motivations

While promising, current approaches to medical image analysis exhibit fundamental limitations that motivate the need for novel methodologies aligned with clinical practice. The aforementioned challenges highlight the need for human-AI integration approaches that align computational methods with clinical expertise rather than attempting to replace human judgment.

Several critical gaps in existing techniques drive our research: Despite their impressive benchmark performances, deep learning models for medical imaging suffer from significant shortcomings when deployed in real-world clinical environments. Most architectures optimize for statistical performance on curated datasets rather than clinical utility, leading to models that excel in controlled settings but fail to address the complexities and variability encountered in routine practice. Conventional convolutional neural networks, while powerful for feature extraction, lack the anatomical priors and structured reasoning capabilities inherent to human diagnostic processes. Furthermore, the data-hungry nature of these models creates practical barriers in medical domains where annotated data is scarce and expensive to obtain.

The generalization failure across institutions and imaging protocols represents perhaps the most concerning limitation of current approaches. The performance degradation observed when deploying models trained at one institution to another—even when the clinical task remains identical—undermines the scalability and cost-effectiveness promised by AI systems. This problem is exacerbated by the proliferation of proprietary imaging protocols and equipment configurations across healthcare systems. While human radiologists naturally adapt their knowledge across these variations, current deep learning models essentially "overfit" to their training environment, capturing incidental correlations related to institutional factors rather than focusing on invariant disease characteristics.

Precise boundary detection, a seemingly straightforward task for human experts, remains elusive for computational approaches. Conventional segmentation algorithms struggle with the inherent ambiguity in medical image boundaries, particularly in regions where pathology alters normal tissue appearance or physiological motion creates artifacts. This limitation directly impacts downstream clinical applications that require accurate volumetric measurements, treatment planning, or longitudinal comparisons. The failure to achieve human-level precision in this fundamental task undermines confidence in more complex analytical capabilities.

Current AI systems fundamentally diverge from human cognition in their approach to image interpretation. While radiologists employ hierarchical reasoning, anatomical knowledge, and causal understanding to interpret images, deep learning models primarily detect statistical patterns without meaningful conceptual representation of anatomical structures or pathophysiological processes. This creates a "cognitive mismatch" between human and AI diagnostic approaches, where models may arrive at correct classifications through reasoning pathways that make little sense to clinicians. This divergence hampers explainability and increases the risk of unexpected failures when encountering novel presentations or rare variants.

Clinical relevance and explainability requirements cannot be treated as secondary considerations or post-hoc additions to existing models. The healthcare context demands that AI systems provide accurate predictions and support clinical decision-making through explanations that align with medical knowledge and reasoning. Current explainability methods often produce visualizations that highlight statistically significant regions without connecting them

to clinically meaningful concepts, creating a "semantic gap" between model explanations and medical reasoning. Moreover, the regulatory landscape increasingly demands transparent AI systems that can justify their outputs in human-interpretable terms.

The incorporation of human expertise into computational models represents an untapped opportunity. Rather than treating AI development as an isolated technical challenge, there is compelling motivation to design systems that explicitly leverage the complementary strengths of human and machine intelligence. Radiologists possess contextual knowledge, causal reasoning abilities, and ethical judgment that computational systems lack, while machines excel at consistent pattern recognition across large datasets. Models that can meaningfully incorporate human knowledge, mimic clinical reasoning processes, and integrate seamlessly with existing workflows represent the next frontier in medical image analysis.

These motivations collectively point toward the need for a paradigm shift in our approach to medical image analysis—moving from purely data-driven statistical models toward hybrid systems that combine the power of computational methods with the reasoning capabilities and domain knowledge of clinical experts. This dissertation explores this interdisciplinary frontier, aiming to develop systems that think more like clinicians while retaining the advantages of computational approaches.

1.3 Research Objectives

This dissertation addresses five interrelated research objectives that systematically target the challenges identified in medical image analysis. Each question corresponds to specific objectives that collectively advance the field toward more clinically relevant and human-aligned computational approaches.

1.3.1 Spatial Correlation in Multi-Modal Imaging

Research Question 1: How can we improve spatial matching between mammography and specimen radiography images to enhance diagnostic accuracy and surgical guidance?

Objectives:

1. Develop novel registration algorithms specifically tailored to the challenging task of matching 2D projections (mammograms) with their corresponding specimen radiographs acquired under different compression and orientation conditions

2. Quantify the performance improvements in lesion localization and margin assessment when enhanced spatial correlation techniques are applied

3. Evaluate the clinical impact of improved spatial matching on surgical decision-making and re-excision rates

Our research addresses the fundamental challenge of maintaining spatial correspondence across different imaging modalities and acquisition conditions, a critical capability for accurate diagnosis and treatment planning in breast cancer care.

1.3.2 Hybrid Approaches for Boundary Detection

Research Question 2: How can the strengths of traditional Active Contour Models (ACMs) and modern deep learning approaches be combined to enhance boundary detection in medical images?

Objectives:

1. Design a synergistic framework that leverages both the mathematical guarantees of ACMs and the feature representation power of deep neural networks

2. Evaluate the hybrid approach against standalone deep learning and traditional ACM methods across diverse anatomical structures and imaging modalities

3. Analyze the robustness of the hybrid approach to variations in image quality, noise, and anatomical variability

Our research aims to address the limitations of purely data-driven approaches by incorporating explicit shape constraints and prior knowledge, while maintaining the flexibility and feature extraction capabilities of deep learning.

1.3.3 Learnable Parameters in Traditional Models

Research Question 3: How can we make traditionally fixed parameters in Active Contour Models learnable within neural network architectures to enhance adaptability across different imaging modalities?

Objectives:

1. Formulate a differentiable framework that enables end-to-end training of ACM parameters within deep learning architectures
2. Develop adaptive parameter selection mechanisms that adjust to specific image characteristics and anatomical contexts
3. Demonstrate improved generalization across different imaging protocols and patient populations through learnable parameter frameworks

Our research bridges the gap between classical mathematical models and modern learning-based approaches, creating systems that combine theoretical guarantees with data-driven adaptability.

1.3.4 Boundary-Aware Transformer Models

Research Question 4: How can transformer-based architectures be enhanced with explicit boundary awareness to improve segmentation accuracy in medical imaging?

Objectives:

1. Design novel attention mechanisms that specifically emphasize boundary features within transformer architectures.
2. Incorporate boundary-focused loss functions and architectural components that guide the model to prioritize accurate delineation of anatomical borders.
3. Evaluate the impact of boundary-aware transformers on downstream clinical tasks requiring precise boundary detection.

Our research addresses the limitations of current transformer models in medical image segmentation, particularly their tendency to focus on global features at the expense of precise boundary localization.

1.3.5 Integration of Human Attention Patterns

Research Question 5: How can human attention patterns and clinical expertise be effectively integrated into medical AI systems to align computational attention with radiological expertise?

Objectives:

1. Collect and analyze eye-tracking and interaction data from radiologists during diagnostic tasks to create models of expert visual attention.

2. Develop novel mechanisms to incorporate these human attention patterns as priors or constraints within deep learning architectures.

3. Evaluate the impact of human-aligned attention on model explainability, diagnostic accuracy, and clinical acceptance.

Our research directly addresses the cognitive mismatch between AI and radiologists by explicitly aligning computational attention with human expertise, potentially leading to systems that perform well statistically but also "think" in ways that make sense to clinical users. These five research questions form a coherent progression from specific technical challenges in medical image analysis to broader questions about human-AI integration. Together, they represent a comprehensive approach to addressing the limitations of current methods while moving toward systems that more effectively complement and augment human clinical expertise.

1.4 Thesis Contributions

This dissertation makes several original contributions to medical image analysis, advancing both technical capabilities and clinical relevance of computational approaches. The key contributions are as follows:

1. Implementation of novel spatial correlation techniques for matching 2D projections with corresponding specimen radiographs.

2. A formal framework for integrating traditional active contour models within deep learning architectures, including mathematical formulations for differentiable ACM layers that enable end-to-end training.

3. A conceptual framework for aligning computational attention with human expert attention, including formal definitions of attention consistency metrics.

4. Creation of adaptive parameter selection mechanisms for ACMs that adjust to specific image characteristics and anatomical contexts.

5. Comparative analysis of boundary-aware transformers against standard transformer architectures, demonstrating significant improvements in segmentation accuracy at anatomical borders.

These contributions collectively advance the state of the art in medical image analysis while addressing critical barriers to clinical translation, particularly in boundary precision, interpretability, and alignment with clinical expertise.

1.5 Thesis Organization

The remainder of this dissertation is organized into the following chapters:

Chapter 2 reviews the relevant literature across multiple domains, including medical image analysis, active contour models, deep learning approaches to segmentation, attention

mechanisms, and human-AI integration. This chapter establishes the foundation for the research and identifies gaps in existing approaches.

Chapter 3 addresses the first research question, focusing on improving spatial matching between mammography and specimen radiography images. The chapter presents novel registration algorithms and evaluates their impact on lesion localization and surgical guidance.

Chapter 4 explores the second research question, developing synergistic frameworks that combine the strengths of traditional Active Contour Models with deep neural networks. The chapter details the architectural design, implementation, and evaluation of these hybrid approaches.

Chapter 5 addresses the third research question, presenting methods for making traditionally fixed parameters in ACMs learnable within neural network architectures. This chapter introduces differentiable formulations of ACM parameters and demonstrates their adaptability across different imaging contexts.

Chapter 6 focuses on the fourth research question, enhancing transformer-based architectures with explicit boundary awareness for improved segmentation accuracy. The chapter presents novel attention mechanisms and architectural components specifically designed for precise boundary delineation.

Chapter 7 addresses the fifth research question, examining how human attention patterns and clinical expertise can be effectively integrated into medical AI systems. This chapter presents methods for collecting and analyzing eye-tracking data and incorporating this information into model training and inference.

Chapter 8 synthesizes our findings across the research questions, discusses their implications for the field of medical image analysis, identifies remaining challenges, and outlines promising directions for future research.

Appendix A presents core terms and mathematical concepts.

Each technical chapter (Chapter 3–Chapter 7) follows a consistent structure: introduction to the specific problem, methodology, experimental design, results, discussion, and conclusions. This organization facilitates the presentation of the research as a coherent whole while allowing

each chapter to stand as a comprehensive treatment of its specific research question.

CHAPTER 2

Related Work

Neural networks, particularly CNNs, have revolutionized the field of image processing and computer vision due to their ability to learn spatial hierarchies of features from raw image data automatically. CNNs are particularly effective in medical image segmentation because of their capacity to capture complex patterns in highly dimensional data, such as CT scans, MRI, and X-ray images.

2.1 CNN-Based Medical Image Segmentation

U-Net, one of the most widely used architectures for medical image segmentation, is a CNN-based architecture specifically designed for biomedical image segmentation tasks. Ronneberger et al. (2015) introduced U-Net as a fully convolutional network that is particularly effective for segmenting small and irregular structures in medical images.

Several variations of U-Net have been proposed in the literature to improve performance in specific medical imaging tasks, such as 3D U-Net for volumetric data developed by Çiçek et al. (2016), which extends the original U-Net for 3D image segmentation, making it well-suited for MRI and CT scans. The attention U-Net of Oktay et al. (2018) introduces attention mechanisms to the U-Net architecture, allowing the model to focus on the most relevant regions of an image while suppressing irrelevant information, improving the segmentation quality in complex cases.

The U-Net design was extended to 3D by V-Net (Milletari et al., 2016) and 3D U-Net (Çiçek et al., 2016), addressing the volumetric nature of much medical imaging data. Further innovations included UNet++ (Zhou et al., 2018), which proposed additional skip connections,

and H-DenseUNet (Li et al., 2018), which combined 2D and 3D networks to reduce memory usage while maintaining 3D contextual information.

2.2 Vision Transformer-Based Segmentation

Several public challenges, such as the Medical Segmentation Decathlon (Antonelli et al., 2022), continue to be closely contested in attempts to provide reliable segmentation performance. The advent of Transformers has proved pivotal in taking meaningful steps towards achieving this goal. Initially developed for natural language processing (Vaswani et al., 2017), transformers have been adapted to vision tasks through the Vision Transformer (ViT). The seminal work on ViTs (Dosovitskiy et al., 2020) demonstrated the idea of leveraging the concepts of attention from the natural language processing domain and applying them to computer vision tasks. This model also demonstrated how pre-training the networks on large public datasets and finetuning them to specific benchmarks such as ImageNet-ReaL (Beyer et al., 2020) can help improve performance over purely convolutional architectures like ResNet. Several other researchers have demonstrated the power of self-supervised representation learning to train transformer networks (Dai et al., 2021).

The ViT architecture divides images into fixed-size patches, processes them as token sequences, and applies self-attention mechanisms to capture global relationships. Early works in this domain (Xie et al., 2021; Valanarasu et al., 2021; Xu et al., 2023) utilized transformer blocks in the network as bottleneck feature encoders. A key issue with ViTs was the quadratic complexity of the self-attention layers, which resulted in expensive computation times. The Swin Transformer (Liu et al., 2021) proposed a way to overcome this by using a hierarchical ViT architecture leveraging non-overlapping windows to compute self-attention, thereby reducing the computation to linear complexity. This approach made transformer-based segmentation computationally feasible for high-resolution 3D medical images.

UNETR (Hatamizadeh et al., 2022) pioneered the integration of transformers into medical image segmentation by using a ViT as an encoder within a U-Net-like structure. These ViT and SwinViT models were further leveraged to create the UNETR and SwinUNETR

(Tang et al., 2022) networks, directly embedding the transformer blocks as the encoder of the network in a U-shaped architecture, thereby allowing for dense inference from multi-scale features.

Other transformer-based approaches, including nnFormer (Zhou et al., 2021) and VT-UNet (Peiris et al., 2022), offer unique architectural modifications to enhance performance on medical segmentation tasks.

2.3 Large Vision Models

Large Vision Models (LVMs) have shown robust performance in segmentation and classification tasks of natural images using datasets of natural images such as COCO and ImageNet. However, fewer images are available for training a model in the medical domain, and even fewer are annotated. While large datasets such as MIMIC-CXR and the National Lung Screening Trial have been made publicly available, they lack annotations, such as lung and nodule segmentations. Moreover, differences in how images are acquired, even when imaging the same anatomical region, can affect the performance of segmentation methods. For example, differences in camera and lighting may cause algorithms to undersegment vessels on retinal images.

The Segment Anything Model (SAM) (Kirillov et al., 2023) solves different downstream tasks by getting prompts from the user. The prompt guides the model to look for image features in a frame or at a point. Researchers have shown that LVMs not purposely trained on medical images perform poorly on medical images. Some examples are shown in (Mazurowski et al., 2023). In this work, the model's performance improved after training the model on medical images. However, Shi et al. (2023) demonstrated that even after training SAM on medical images, it still fails to perform well on tree-like images.

2.4 Boundary-Aware Segmentation

Boundary detection has long been recognized as a critical component of accurate image segmentation. Early work by Xie and Tu (2015) demonstrated the effectiveness of edge detection networks, while holistically-nested architectures highlighted the importance of multi-scale boundary supervision. In the medical domain, boundary-aware approaches have shown particular promise due to the importance of precise delineation of anatomical structures.

Hatamizadeh et al. (2019b) introduced a boundary-aware CNN for medical image segmentation that incorporated an auxiliary boundary detection stream, demonstrating improved liver and tumor segmentation performance. Similarly, MSDS-UNet (Yang et al., 2021) employed multi-scale deep supervision to enhance boundary detection in lung tumor segmentation.

More recent approaches have extended boundary awareness to transformer architectures. Zhu et al. (2017) applied deep supervision to prostate segmentation. However, the integration of boundary awareness into transformer-based volumetric segmentation networks remains underexplored, particularly in the context of the Medical Segmentation Decathlon (MSD) challenge datasets, which span diverse anatomical structures and imaging modalities.

Our work builds upon these foundations by introducing a boundary-aware approach for the SwinUNETR architecture, specifically designed to detect edges and enhance boundary delineation across the MSD challenge tasks.

2.4.1 Edge Detection

A key difficulty in medical image segmentation involves accurately predicting the boundary shapes of organs and tumors. The tumors, in particular, can be difficult to predict due to their small and highly variable shapes. Research (Acuna et al., 2019; Xie and Tu, 2015) identified "boundary pixels" — i.e., pixels that belong to boundaries of the label classes — and proposed architectural modifications to enforce the network to predict a maximum response along the normal direction at an edge. Leveraging this concept, (Hatamizadeh et al., 2019b) used a Boundary Stream in a 2-D UNet, utilizing attention maps and a weighted

binary cross-entropy "edge loss" to improve segmentation performance. Our work extends this concept to the 3D Swin-UNETR with some additional modifications. Other works such as by (Hu et al., 2022) proposed the use of a separate boundary decoder to improve performance, while Kervadec et al. (2019) proposed a novel Boundary loss that makes use of distance maps on the space of contours.

2.4.2 Active Contour Models

Active Contour Models (ACMs) (Kass et al., 1988) evolve contours by minimizing energy functionals that attract the deformable contour to intensity edges in the image or that maximize differences in image intensity or texture between the target object (for example, a lesion) and its background (surrounding tissue) (Chan and Vese, 2001).

Researchers have recently sought to combine ACMs and deep learning approaches. Wang and Vemuri (2004) combined CNNs and parametric ACMs for the segmentation of buildings in aerial images; however, their method requires manual contour initialization, fails to delineate the boundary of complex shapes precisely, and segments only single objects, all of which limit its applicability to lesion segmentation due to the irregular shapes of lesion boundaries and the potential for needing to segment multiple lesions in a single image. Hatamizadeh et al. (2019a) implemented the Deep active lesion segmentation (DALS) framework that benefits from an improved level-set ACM formulation with a per-pixel-parameterized energy functional and a novel multiscale encoder-decoder convolutional neural networks (CNNs) that learns an initialization probability map along with parameter maps for the ACM. However, the issue with the DALS network is the fixed constant in the energy function, which needs to be tuned when applied to different datasets. Hatamizadeh et al. (2020) presented Trainable Deep Active Contours (TDACs), an automatic image segmentation framework that unites CNNs and ACMs. The Eulerian energy function of the ACM component includes per-pixel parameter maps predicted by the backbone CNN, which also initializes the ACM. Although the entire TDAC architecture is end-to-end automatically differentiable and backpropagation trainable without user intervention, the hyperparameters are still heuristically chosen. Given

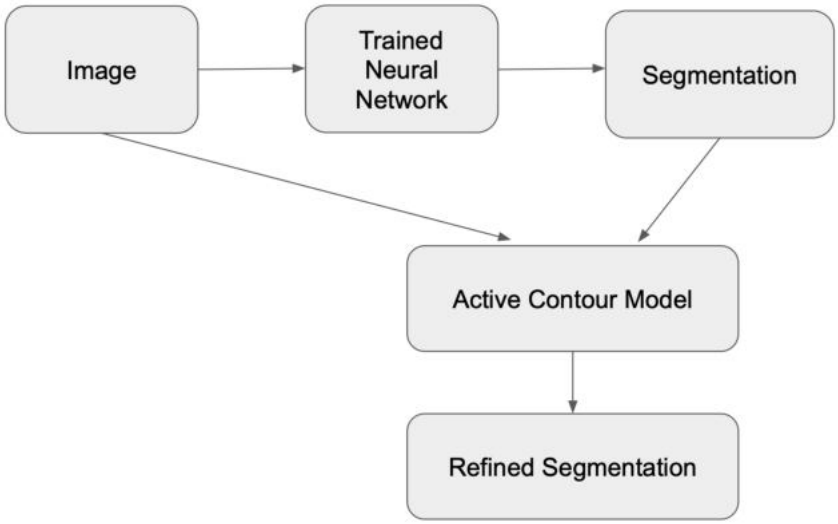


Figure 2.1: ACM appended after CNN

these limitations, we pursued a simpler approach to combining ACMs and SAM in order to to improve the medical image segmentation performance of the pretrained LVM.

Certain popular hybrid frameworks utilize deep learning frameworks to obtain initial segmentation masks for images and later refine these segmentation masks using active contour models outside the learning framework. An outline of this procedure is depicted in Figure 2.1. Hatamizadeh et al. (2019a) utilizes a multiscale encoder-decoder CNN architecture, whereas Nakhaei et al. (2024) uses the SAM architecture for the 'trained neural network' component in the image.

Vepa et al. (2022) proposed a weakly supervised framework for cerebral vessel segmentation in DSA images. They use an active contour model to generate weak annotations, refined through human-in-the-loop strategies. These labels are fed into a CNN for further refinement. CLAHE is applied as a pre-processing step to enhance vessel visibility. The method reduces reliance on manual annotations and achieves state-of-the-art performance, surpassing human annotators.

2.5 Domain Adaptation

The generalizability of machine learning algorithms, such as image segmentation, has been a longstanding problem. Domain adaptation seeks to generalize a model trained on a source dataset to achieve similar performance when applied to a related target dataset (Guan and Liu, 2021). One variant of domain adaptation is test-time domain adaptation, which adapts the trained model using a subset of unlabeled cases from the target dataset (Fleuret et al., 2021). We attempt to achieve test-time domain adaptation by adding a separate component based on ACMs that requires minimal parameterization using test data to improve the LVM's performance on the target dataset.

2.6 Integrating Human Visual Patterns

Recently, a focus has been on integrating human visual patterns with computer-aided analysis (Ibragimov and Mello-Thoms, 2024). In medical imaging, understanding the radiologist's visual search strategies through eye-tracking has become an important means of studying perceptual errors and guiding algorithm design for diagnosis (Karargyris et al., 2021).

Many studies have examined extracting radiologists' gaze to analyze how radiologists scan medical images (Bigolin Lanfredi et al., 2022; Demner-Fushman et al., 2015). Studies show that metrics, such as fixation duration, saccade length, and scan-path patterns, correlate with diagnostic accuracy and clinical expertise (Ibragimov and Mello-Thoms, 2024). Based on these findings, datasets such as REFLACX (Bigolin Lanfredi et al., 2022) have explored large-scale, timestamped gaze annotations for chest X-ray reading. Based on subsequent analyses of REFLACX, researchers have shown that expert radiologists typically have more targeted fixations over anatomically relevant regions compared to newly-trained radiologists, pointing to the fact that visual search efficiency (e.g., fewer unnecessary fixations, faster time to fixate on true abnormalities) is key to radiological expertise (Hsieh et al., 2024).

Studies have explored applying gaze data to interactive tasks. Some have used eye-tracking to provide annotations of abnormalities, reducing the labor involved in producing

manual bounding boxes and pixel-wise masks (Stember et al., 2019; Hsieh et al., 2024). The radiologists' gaze pattern acts as a weak supervisory signal, which preserves essential diagnostic cues while accelerating dataset annotation. In works such as (Khosravan et al., 2019), collaborative computer-aided diagnosis (C-CAD) frameworks allow the human reader and the system to iteratively refine predictions. The system utilizes gaze fixations to detect false positives or focus on suspicious regions, improving accuracy over standalone CAD pipelines.

Many deep learning-based methods have explored integrating gaze fixation with image features. Some studies use CNNs or transformer-based backbones to process images, while separately modeling gaze as a sequence of spatial coordinates (Ji et al., 2023).

Some studies explore teacher-student or attention alignment paradigms. Kumar and Marttinen (2024) introduced an enhanced CLIP variant that incorporates expert gaze heatmaps as auxiliary supervision for contrastive image–text learning, effectively aligning model attention maps with the radiologist's fixations.

Different models and architectures have attempted to use gaze in their training strategy, but all have one main objective: utilizing radiologists' visual expertise to improve both accuracy (e.g., lesion detection) and explainability (e.g., producing model saliency maps that align with human attention).

Current methods have the following limitations: First, they incorporate the gaze fixations in a late-fusion manner in their network and rely on manual pre-processing of the data (Ji et al., 2023; Ibragimov and Mello-Thoms, 2024). Second, few methods capture clinical reports or anatomical priors alongside gaze data, in turn missing the real-time contextual knowledge radiologists have access to while scanning an image. Patient history, textual notes, and external factors may influence a radiologist's attention. Lastly, some approaches use gaze data without the temporal information, or in a demanding way requiring pixel-level bounding boxes, limiting their applicability in the clinical environment.

CHAPTER 3

Spatial Matching of 2D Mammography Images and Specimen Radiographs

3.1 Introduction

Breast cancer is the most common invasive cancer and the second leading cause of death in women (Sun et al., 2017). Routine screening mammography has the potential to detect breast cancer in its earliest stage, before it becoming a potentially lethal invasive breast cancer. Ductal carcinoma in situ (DCIS) is breast cancer confined to the milk duct and is the earliest stage of breast cancer. DCIS may present with mammographically visible microcalcifications. However, making a diagnosis of DCIS can be challenging since most microcalcifications seen on a mammogram are benign or non-cancerous. So, for every woman who undergoes screening mammography followed by breast biopsy and receives a new diagnosis of breast cancer, approximately two additional women will undergo a benign breast biopsy. Although modern percutaneous core needle biopsy is safe, unnecessary biopsies are anxiety-provoking and are considered a risk associated with breast cancer screening. Microcalcifications with suspicious morphology (e.g., amorphous) are particularly challenging, given the difficulty characterizing them and the associated diagnostic uncertainty.

Figure 3.1 depicts the sequence of events when indeterminate screen-detected mammographic calcifications are identified. The patient is recalled for further evaluation with diagnostic mammography and magnification of the calcifications. Magnified diagnostic views allow the radiologist to determine if the mammographic calcifications are suspicious, probably benign, or benign. Suspicious calcifications warrant biopsy. Benign calcifications should probably be followed with repeat short-interval imaging in 6 months to assess for change or

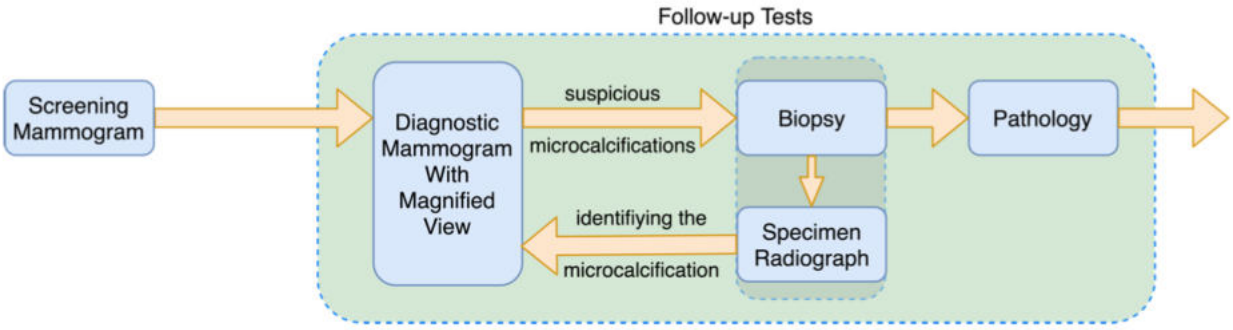


Figure 3.1: The general process for diagnosing suspicious microcalcifications. This work focuses on the magnification view obtained during the diagnostic mammogram and specimen radiographs acquired post-biopsy.

stability; those with benign calcifications return to routine mammographic screening. Women who are recalled from screening mammography for additional evaluation experience anxiety associated with the need to return for additional imaging and, if biopsy is indicated, anxiety about the procedure and results. Unnecessary benign biopsies are known as false positives. It is desirable to minimize these unnecessary biopsies while maintaining a high mammographic sensitivity (i.e., not missing any cancers).

Towards addressing the challenge of managing suspicious microcalcifications, our overall goal is to develop a computer-aided diagnosis algorithm that quantitatively analyzes the morphology and distribution of microcalcifications on diagnostic mammograms to distinguish between cases that should undergo biopsy versus short-term follow-up imaging. Given that microcalcifications are a byproduct of biological processes that may indicate the presence of invasive cancer, we wish to relate the mammographic appearance of microcalcifications to the tissue and cellular structure that is observable under the microscope. However, precisely pinpointing the region in the mammogram where a biopsy specimen was taken is challenging. One potential approach is to utilize the specimen radiographs of biopsy cores taken clinically to ensure that the targeted region of microcalcifications was collected during the biopsy. In this work, we investigate the feasibility of using the shape and distribution of microcalcifications in specimen radiographs to identify a region in the mammography image that appears to show similar microcalcifications. We posit that the ability to link a specific region in the mammography image to the tissue imaged in the specimen radiograph will enable more

precise radiology-pathology correlation.

The novelty of the work is summarized as follows:

1. Proximity functions represent individual MCs, providing a way to accommodate uncertainty in boundaries as part of classifier training.

2. A dense regression model and a novel blob segmentation algorithm are applied to generate MCs' accurate segmentation while achieving fewer false positives than other state-of-the-art algorithms.

3.2 Methods

Data were collected retrospectively following an institutional review board-approved protocol from patients seen at a single academic medical center. The dataset consisted of diagnostic mammograms from 80 patients. Each patient had two mammogram views: a magnified craniocaudal (CC) view and a magnified 90-degree mediolateral or lateromedial (ML/LM) view. The magnified CC-view images were used for training, as suggested by clinicians, because microcalcifications are more visible on CC versus other views. All other views (51 CC and ML/LM views) were used for testing. All images were acquired using Hologic Selenia full-field digital mammography equipment at 0.070 mm per pixel resolution and 12-bit grayscale. For each case, specimen radiographs of the biopsied tissue were also obtained by placing the tissue cores into a plastic tray and imaging the specimen via X-ray. A breast fellowship-trained, board-certified radiologist (BL) and a breast fellow (NC) reviewed all of the full views, magnified views, and the specimen X-rays and annotated individual microcalcifications. The open-source medical image viewer Horos was utilized to create the annotations, marking the spatial location of visible microcalcifications that were the biopsy target with a single point. The smallest bounding box containing the points annotated on the mammograms was used to denote.

The reference biopsied region to which the results of the template matching approach were compared. The process for finding the region on the mammogram associated with the

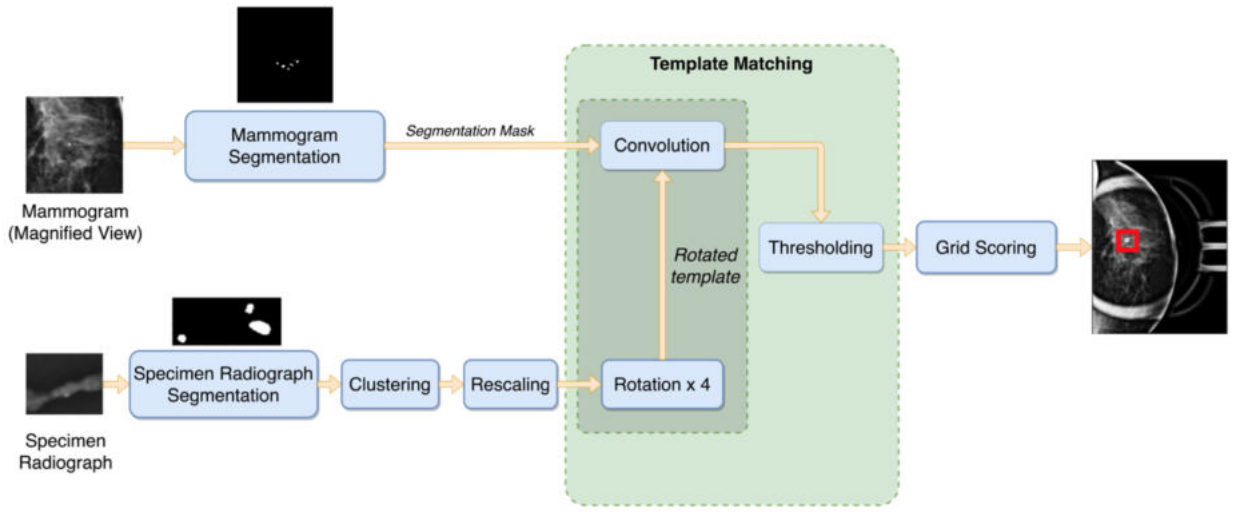


Figure 3.2: Overall approach for spatially matching microcalcifications imaged in the diagnostic mammogram and specimen radiograph.

calcifications in the specimens is illustrated in Figure 3.2. Briefly, magnification views of the diagnostic mammogram and specimen radiograph are automatically segmented using the same approach. The diagnostic mammogram is then divided into non-overlapping patches. A clustering algorithm is applied to the segmented specimen radiographs to group related calcifications. These groups are used as templates and matched against each patch within the mammogram. The output of the template matching process is a score that is used to determine the likely patch from which the microcalcification group in the specimen radiograph came. The following sections describe each step in detail.

3.2.1 Microcalcification Segmentation

Calcifications captured in magnification views were detected and segmented by an algorithm developed by our team (Marasinou et al., 2021). A sample result is shown in Figure 3.3. The method consists of two stages: (1) bright candidate objects are delineated using difference-of-Gaussians with Hessian analysis, and (2) a convolutional regression model is applied to choose the candidate objects corresponding to calcifications. The calcifications on the specimen radiograph are segmented using the same method; an example result is shown in Figure 3.3. Two image masks are generated: one of the microcalcifications from the magnified view and

the other of the microcalcifications from the specimen radiograph.

3.2.2 Clustering Microcalcification

To provide a more robust way to match specimen radiographs to mammograms, we investigated ways to identify groups of microcalcifications observed on the specimen radiographs that could be used as landmarks. To generate these groups, segmented microcalcifications on the specimen radiograph were clustered using an unsupervised clustering approach called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996). The objects were mapped to their centroid locations (points). Utilizing DBSCAN, points were grouped based on their neighborhood density and a minimum number of points within the group. Points that did not meet the minimum threshold for a group were labeled as outliers. A bounding box that encompasses the largest identified cluster of calcifications was cropped from the image and used as the template. Given that specimen radiographs may be magnified during acquisition, a field in the DICOM header called Estimated Radiographic Magnification Factor was used as a scaling factor. Relative orientations of the calcifications may vary between the mammogram and the specimen radiograph. As such, three additional templates were generated from each template by rotating the original template by 90, 180, and 270 degrees. All four templates were compared to the mammographic calcification mask. The mammogram was divided into a grid of 300×300 non-overlapping patches. Patches that overlapped with the bounding box annotations by the human readers, where biopsies were taken, were considered positive patches. All other patches were considered negative patches.

3.2.3 Template Matching

A template matching algorithm was applied to identify a patch in the magnification view with similar microcalcifications. The template matching algorithm (Di Stefano et al., 2003) requires a template image a (here, the group of microcalcifications segmented from the specimen radiograph) and a target image b (i.e., a patch from the diagnostic mammogram) as inputs. The size of the template image was smaller than the size of the target image

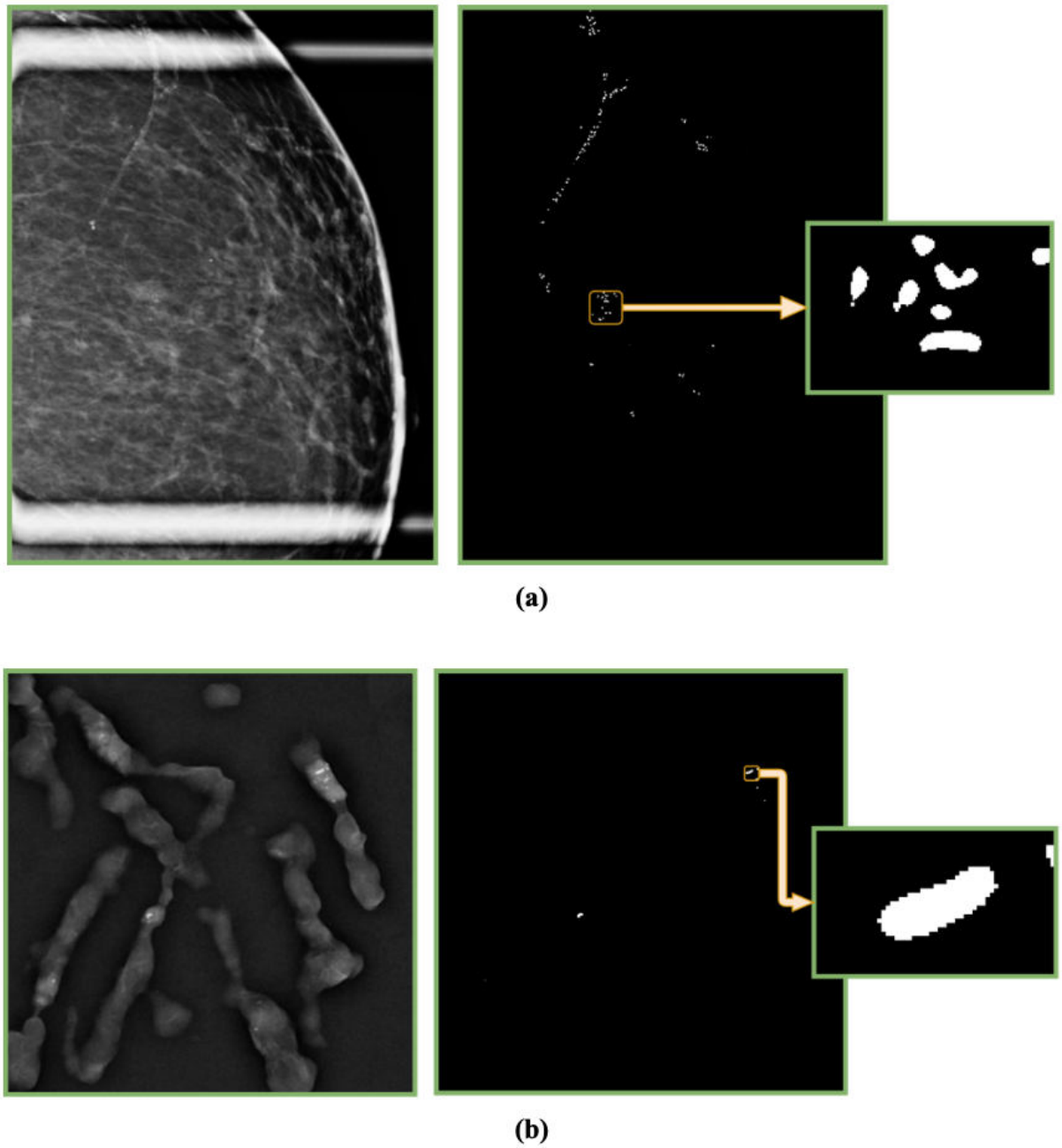


Figure 3.3: Examples of segmented microcalcifications (a) in the magnification view of the diagnostic mammogram and (b) the specimen radiograph.

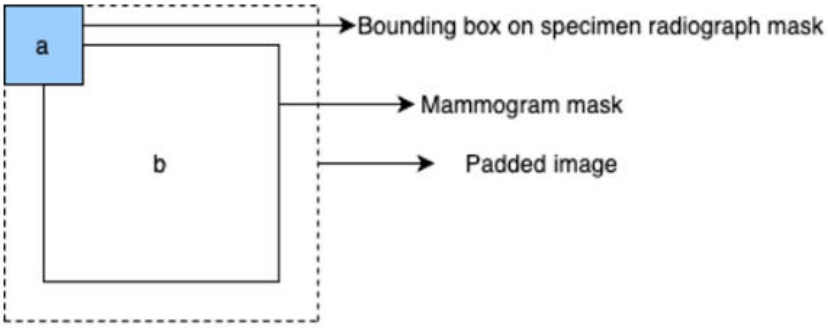


Figure 3.4: A schematic showing how images were padded prior to template matching

since the template was the smallest rectangle containing a group of calcifications, while the target image was the mammogram. Thus, for the computation to be performed on all (x, y) locations, the target image was padded with zero values in all directions by half the template size in each direction, as shown in Figure 3.4. The algorithm convolves template a with image b by aligning the center of the template with each location (x, y) of the image and computing the inner product between the overlapping pixels.

$$f(x,y) \;=\; \sum_{x',y' \in \Omega_a} a(x',y')\, b(x+x',\, y+y'), \tag{3.1}$$

We utilized a cross-correlation metric (3.1), the inner product of each region and each template image, as a similarity score of that patch to the template. The resulting value is interpreted as a similarity score at each location. Similarity scores are calculated for each of the four templates. Since the dimensions of each template are different (due to varying sizes of the cluster of calcifications extracted from the specimen radiograph), the scores are not comparable across different cases. We could not determine a universal threshold for choosing the best matching regions. Instead, the distribution of the scores generated for each location (x, y) of the image was plotted, and locations with a score higher than the 99th percentile of the score distribution were identified as the match.

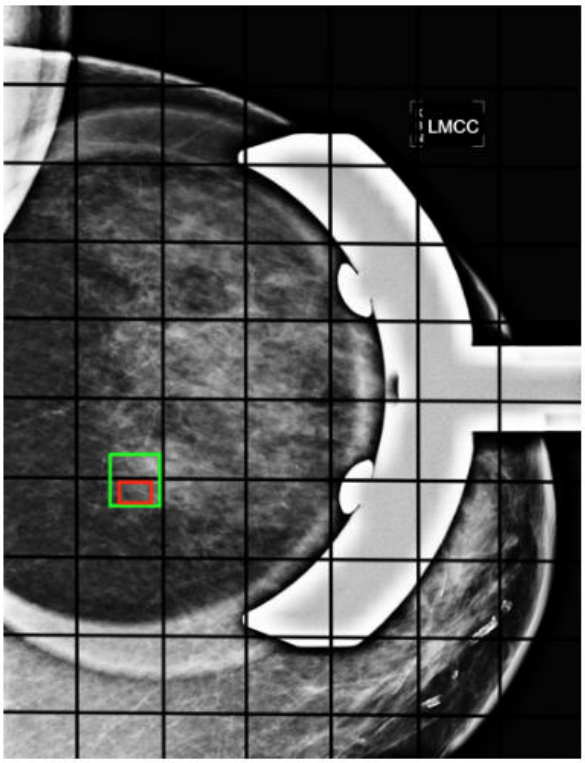


Figure 3.5: A schematic of the grid over the mammogram. The predicted patch is the one containing the red box, and the ground truth patches are the two patches with an overlap with the green box.

3.2.4 Region Scoring

The approach for scoring matched regions is illustrated in Figure 3.5. The top-scoring patch (shown in red) that overlapped with the reference bounding box (shown in green) was counted as a true positive. Patches that were not identified as matches or identified as biopsied regions by the human raters were considered true negatives (all other patches delineated by the grid). Metrics such as accuracy, precision, recall, specificity, and negative predictive value were calculated based on these definitions.

3.3 Experiments and Results

We evaluated our approach using a test set of 80 cases described in Section 3.2. We also visually inspected the results as part of failure analysis. The magnified ML/LM-view images

View	Accuracy	Precision	Recall	Specificity	NPV
Magnified ML	0.99	0.66	0.61	0.99	0.98
Full CC	0.99	0.67	0.58	1.00	0.99
Full ML/LM	0.99	0.69	0.63	1.00	0.99

Table 3.1: Performance of our template matching-based approach on the test cases.

and the full CC and ML views were used to test our approach. Table 3.1 summarizes the results. Our algorithm achieved consistent performance across different views. Given that the number of negative patches greatly outnumbered the number of positive patches, the accuracy and negative predictive value (NPV) were high, as expected. Precision and recall were reported at 0.66-0.69 and 0.58-0.63, respectively. The full ML/LM view achieved the highest precision and recall compared to other views.

3.4 Discussion

We visually inspected the results across all test cases to understand where the algorithm succeeded and failed. Figure 3.6 illustrates four cases: two where the algorithm correctly identified the region and two where the algorithm failed. Our algorithm was likely to fail in several scenarios: 1) the appearance of the microcalcifications was the same in biopsied and non-biopsied regions; 2) the segmentation did not accurately capture the microcalcification shape; and 3) the biopsied regions extended across different patches.

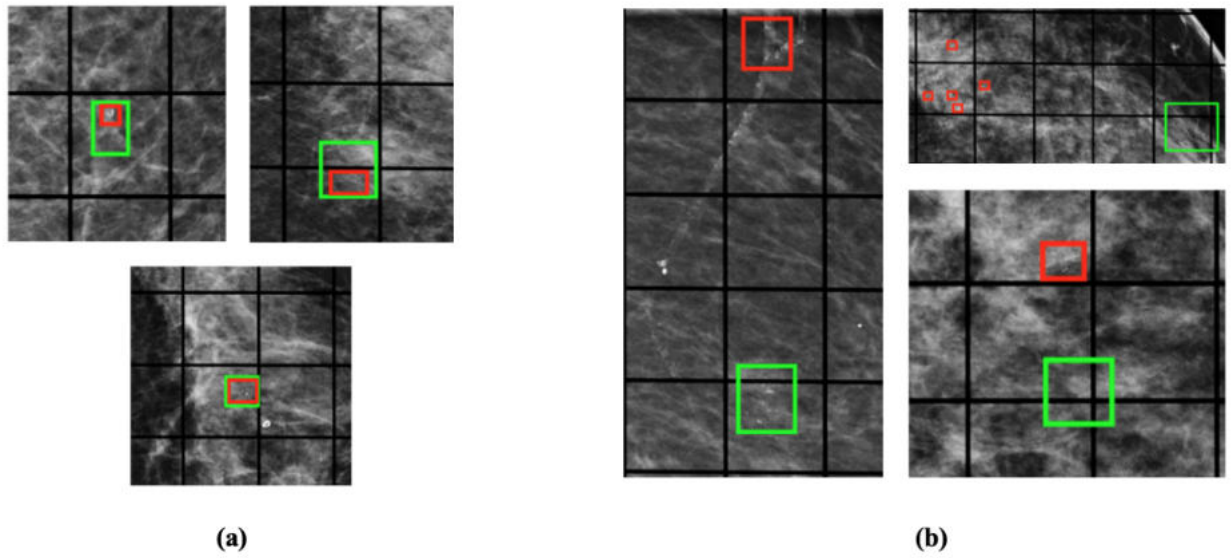


Figure 3.6: Example results from our template matching-based approach. (a) True positive matches made by the algorithm and (b) failure cases.

CHAPTER 4

Active Contour Model Refinement of SAM Boundaries in Medical Images

4.1 Introduction

Accurately capturing a lesion's boundaries is important in cancer diagnosis. However, the contrast between the boundary and surrounding tissue is frequently low, and the similar texture of the lesion and its surrounding area complicates the task of segmenting lesions compared to delineating objects in natural images. Recent advances in medical image segmentation have yielded generalized algorithms that can delineate a variety of regions (Mazurowski et al., 2023). Large vision models (LVM) such as the Segment Anything Model (SAM) have been shown to perform well across different tasks (Kirillov et al., 2023). However, these models also fail to segment the boundaries in noisy medical images or images with poor contrast, erroneously include non-relevant regions, or undersegment regions of interest. They have shown moderate to poor performance in segmenting the images, different from their training sets (Cheng et al., 2023; Shi et al., 2023; Mazurowski et al., 2023). While the performance of segmentation models can be improved with additional supervised training examples from the target dataset, manual annotation of data is labor-intensive, and for rare cancers, few examples may be available for training. As a result of these limitations, domain adaptation techniques that require large amounts of (labeled) data can be impractical.

Dataset Name	Kind	Number of Images	Sample Size	Image Size
CHASE_DB1	Retinal Vessels	28	28	999×960
DRIVE	Retinal Vessels	40	20	584×565
STARE	Retinal Vessels	20	20	700×605
ISIC 2018	Skin Lesions	2594	760	2166×3188

Table 4.1: Datasets

4.2 Methods

In our proposed method, we demonstrate how ACMs can be combined with an existing general-purpose segmentation model trained on large datasets, transferring their weights to work on specific tasks for which only a small number of available labeled cases may exist. For each dataset, we use the pretrained model, SAM, as the initial segmentation output, then try to minimize level-set energy around the contoured boundary and have the images segmented without requiring additional supervised training for fine-tuning.

4.2.1 Datasets

We conduct comparison experiments on three different medical image benchmark datasets to evaluate our proposed model. We focus on these datasets for two reasons: 1) our method can be evaluated and compared to prior literature, and 2) these datasets are sensitive to small differences along the boundaries.

4.2.1.1 Skin Lesion Segmentation

We conduct skin lesion segmentation experiments using the 2018 International Skin Imaging Competition (ISIC) 2018 dataset (Tschandl et al., 2019; Codella et al., 2019). This dataset consists of 2,594 RGB images of skin lesions with an average image size of 2166×3188 pixels. For our experiments, 786 cases are randomly sampled from the entire dataset.

4.2.1.2 Retinal Vessels

Examining retina blood vessels using fundus photography provides important information about the vascular health of the eye, body, and brain. Numerous studies have demonstrated a significant connection between the condition of the retinal blood vessels and many different diseases (Mookiah et al., 2021), thus making the task of retinal vessel segmentation of high importance. The **CHASE DB1** (Fraz et al., 2012), **DRIVE** (Staal et al., 2004), and **STARE** (Hoover et al., 2000) datasets serve as benchmarks for our approach. The CHASE DB1 dataset comprises 28 retinal images sized at 999×960. The STARE consists of 20 retinal images with dimensions of 700×605. The DRIVE dataset consists of 40 images sized at 584×565, but only 20 have annotations. For this reason, we only use those 20 images to be able to report the evaluation results.

4.2.2 Overall Approach

We combine SAM with the level-set ACM to yield a generalizable, fully automatic medical image segmentation method to produce more accurate and detailed boundaries. The framework includes a trained model that feeds a level-set ACM with per-pixel parameter functions. We used Hatamizadeh et al. (2019a) level-set active contour model with parameter functions. Figure 4.1 depicts the overall pipeline. The segmented mask by the model is fed into the ACM layer, and then ACM iterations refine the contour.

4.2.2.1 Pre-Processing

Each image is pre-processed and smoothed using a Gaussian filter with a kernel size of 1 or 5, for retinal vessel images and skin images, respectively. A smaller smoothing kernel was chosen for the retinal vessels due to their more delicate structure.

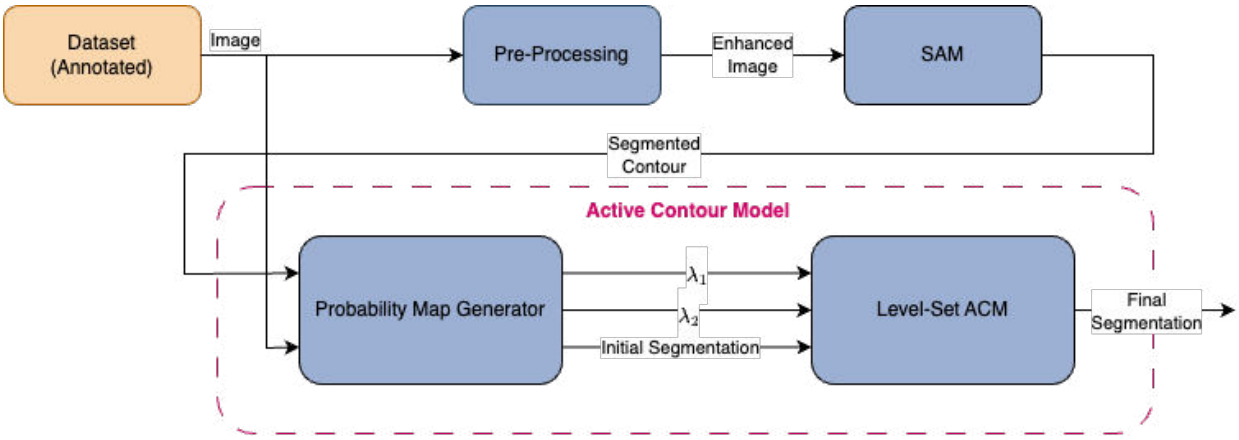


Figure 4.1: Segmentation pipeline.

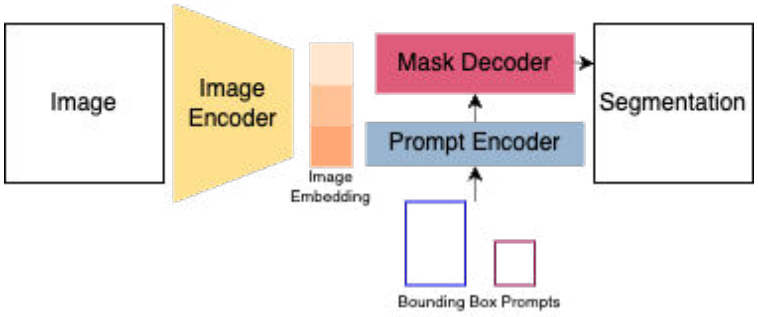


Figure 4.2: Proposed pipeline.

4.2.2.2 Segment Anything Model

SAM was introduced by Kirillov et al. (2023) as a large language model trained on 11 million images and 1 billion masks. SAM takes points or bounding boxes as a guide for focusing on each target in the image as a prompt. Figure 4.2 provides a high-level diagram of the model. In our approach, the prompt given to the model was a bounding box that contained all of the objects in the image.

4.2.2.3 Active Contour Models

The ACM level-set employed in our framework represents an extension of Chan and Vese's original method (Chan and Vese, 2001). The energy functional linked to a closed, time-varying contour C, denoted in $\Omega \in \mathbb{R}^2$ by the zero level set of the signed distance map $\phi(x, y, t)$, is

Method	IOU	Dice Score	Accuracy	Recall	Precision
SAM	0.647	0.379	0.894	0.653	0.878
SAM+ACM	0.756	0.424	0.921	0.851	0.962

Table 4.2: Results on the ISIC 2018 Dataset

formulated as follows: $E(\phi) = \int_\Omega \delta_\epsilon(\phi(x,y,t))(\mu|\nabla\phi(x,y,t)| + \int_\Omega W_s F(\phi(x,y,t))dudv)dxdy$

Here, μ penalizes the length of C, and the energy density is defined as: $F(\phi) = \lambda_1(u,v)(I(u,v) - m_1(x,y))^2 H_\Omega^I(\phi) + \lambda_2(u,v)(I(u,v) - m_2(x,y))^2 H_\Omega^E(\phi)$

To derive the curvature flow for the localized version of the uniform modeling energy, we followed the methodology outlined in (Lankton and Tannenbaum, 2008): $\frac{\partial\phi}{\partial t}(t) = \delta\phi(x)\int_{\Omega_y} B(x,y)\sigma\phi(y)\cdot(I(y) - u_x)^2 - (I(y) - v_x)^2)dy + \lambda\sigma\phi(x)\text{div}\left(\frac{\nabla\phi(x)}{|\nabla\phi(x)|}\right)$

In this context, x and y serve as independent spatial variables, each representing an individual point within $\Omega(x,y) = \begin{cases} 1, ||x-y|| < r \\ 0, \text{otherwise} \end{cases}$. λ_1, and λ_2 are created in the **Probability Map Generator** component by examining the segmentation outputs from the previous step. Then λ_1, λ_2, and the segmented mask, now as the initial contour, from the last step, are passed onto the **Level-Set ACM** component. In the **Level-Set ACM**, the contour is initialized as the segmented output mask. Through each iteration, the contour is updated to minimize the difference between the energy inside and outside of the contour.

4.3 Experiments and Results

4.3.1 Skin Lesion Segmentation

The initial contour was produced using SAM and the ACM was applied for 50 iterations. As shown in Table 4.2, an increase in the Intersection-over-Union (IoU) is observed by 0.11 (95% CI: 0.744–0.767). However, The improvement in segmentation is much more noticeable in cases where the SAM fails to segment the lesion in the presence of artifacts, such as hair, or where the contrast between the object of interest and background is low. Figure 4.3 provides examples of the results of SAM+ACM on the ISIC 2018 dataset.

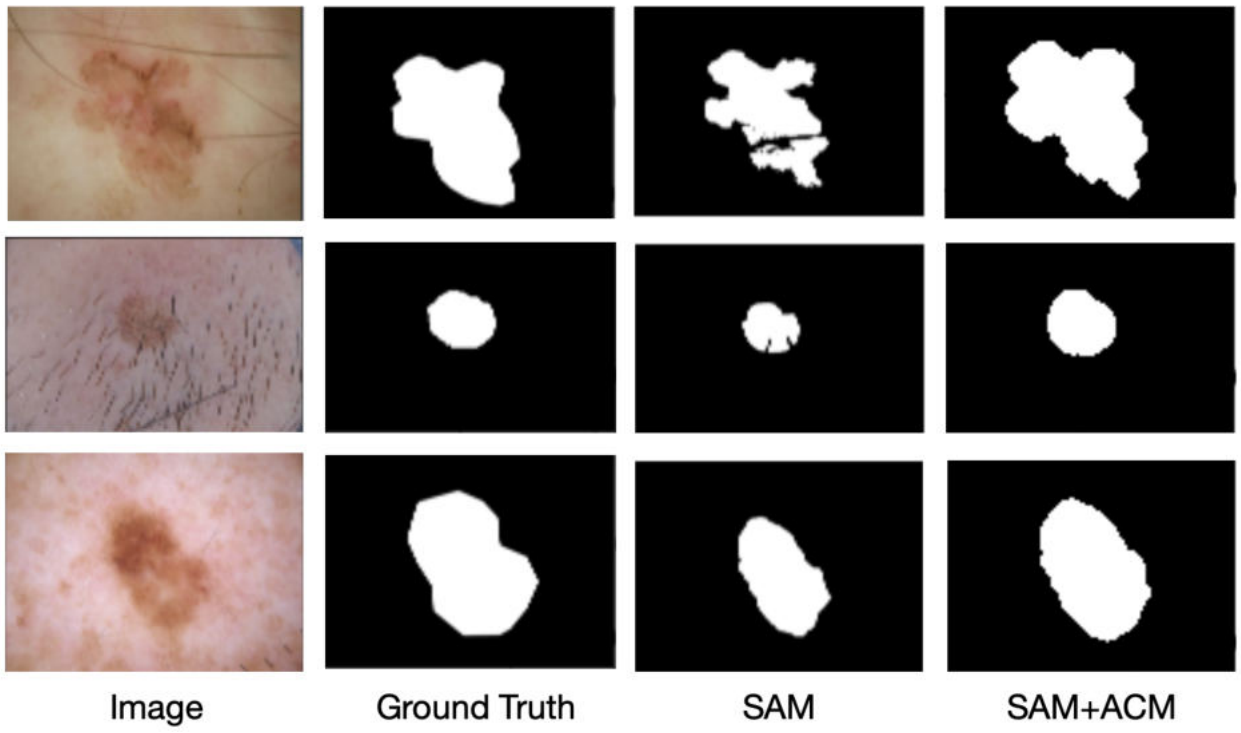


Figure 4.3: Examples of Improvement of Segmentation in ISIC 2018 Dataset

4.3.2 Retinal Vessels

Our approach was applied to three different retinal vessel datasets. First, using SAM, the initial contour was produced. We allowed the active contour models 50 update steps, enough for the contour to stabilize (additional steps yielded negligible boundary changes). As shown in Table 4.3, SAM+ACM achieved the best performance on the CHASE_DB1 dataset, increasing the IoU by 0.19 (95% CI: (0.056, 0.111). For the two other datasets, DRIVE

Dataset	Method	IOU	Dice Score	Accuracy	Recall	Precision
CHASE_DB1	SAM	0.084	0.074	0.780	0.282	0.103
CHASE_DB1	SAM+ACM	0.376	0.243	0.894	0.566	0.463
DRIVE	SAM	0.088	0.078	0.659	0.485	0.113
DRIVE	SAM+ACM	0.096	0.085	0.582	0.621	0.111
STARE	SAM	0.124	0.002	0.867	0.071	0.131
STARE	SAM+ACM	0.270	0.003	0.738	0.964	0.274

Table 4.3: Results on the Retinal Vessel Datasets

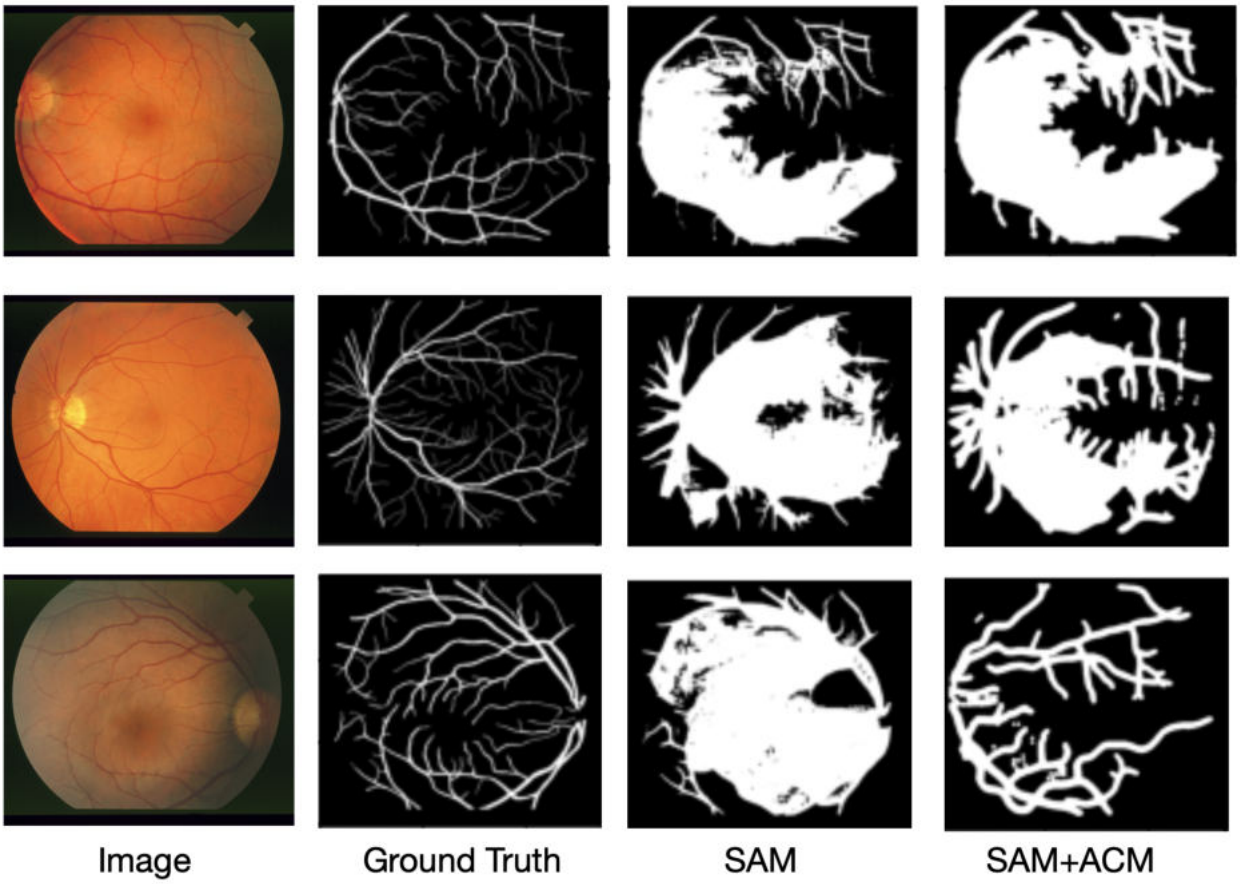


Figure 4.4: Examples of Improvement of Segmentation in STARE Dataset

and STARE, IoU was increased by 0.01 (95% CI: (0.069, 0.124)) and 0.15 (95% CI: (0.097, 0.151)), respectively. However, unlike the ISIC dataset, ACM fails to identify most vessels in cases where the SAM fails to segment any part of the vessel.

4.4 Discussion

Our objective in this chapter was to achieve domain adaptation with minimal retraining or parameter tuning, thereby enhancing the generalizability of SAM using ACMs. While we aimed to personalize object segmentation, there are certain limitations:

- We did not examine how different prompting strategies could have improved SAM's performance on medical images.

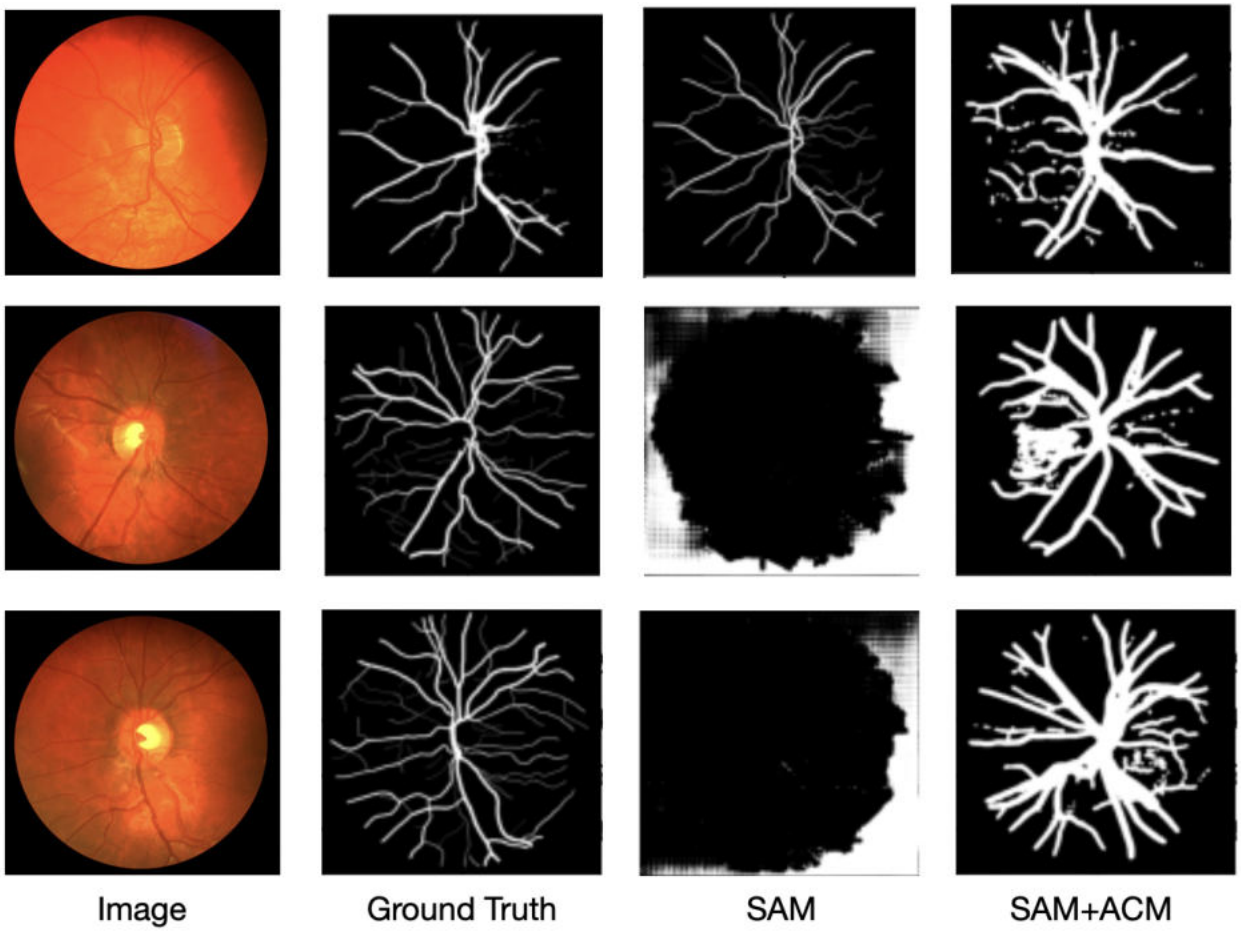


Figure 4.5: Examples of Improvement of Segmentation in CHASE_DB Dataset

- The ACM has hyperparameters that could be tuned to a specific dataset, which we did not investigate. For example, in the STARE and CHASE_DB datasets, failed segmentations suggest that allowing the ACM to go through more iterations may have improved results.

- Failures occurred in cases with objects that may obscure the skin lesion (e.g., hair, moles) and very narrow vessels in retinal vessels.

Figure 4.6, Figure 4.7, Figure 4.8, and Figure 4.9 provide examples of where the proposed method failed to segment the lesion due to the presence of hair, or when there is a low contrast between the object of interest and the background. Overall, the performance of the SAM+ACM approach on the DRIVE dataset was suboptimal. SAM failed to segment or identify vessels in all cases, and ACM proved unhelpful, as it relies on the initial contour

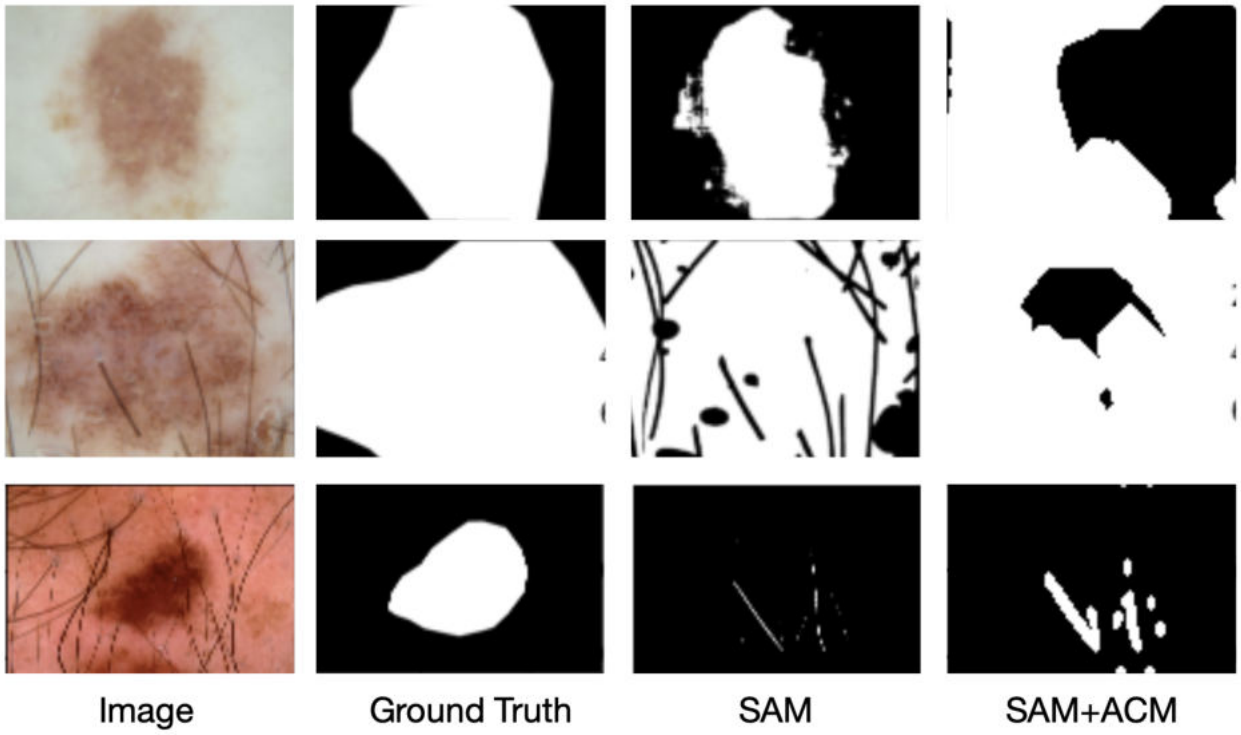


Figure 4.6: Examples of Cases Where Segmentation Failed in the ISIC 2018 Dataset

being specified around the object.

To summarize, obtaining the precise boundaries of challenging lesions when only a few labeled training cases are available is a difficult problem. Even with LVMs trained on large image datasets, fitting the models to the custom, smaller datasets is necessary. Based on our experiments, our approach offers three major improvements: 1) using the models trained on large datasets on unannotated datasets, 2) improved segmentation results on small datasets, and 3) providing personalized segmentation while minimizing the need for additional training and hyperparameter tuning using target data.

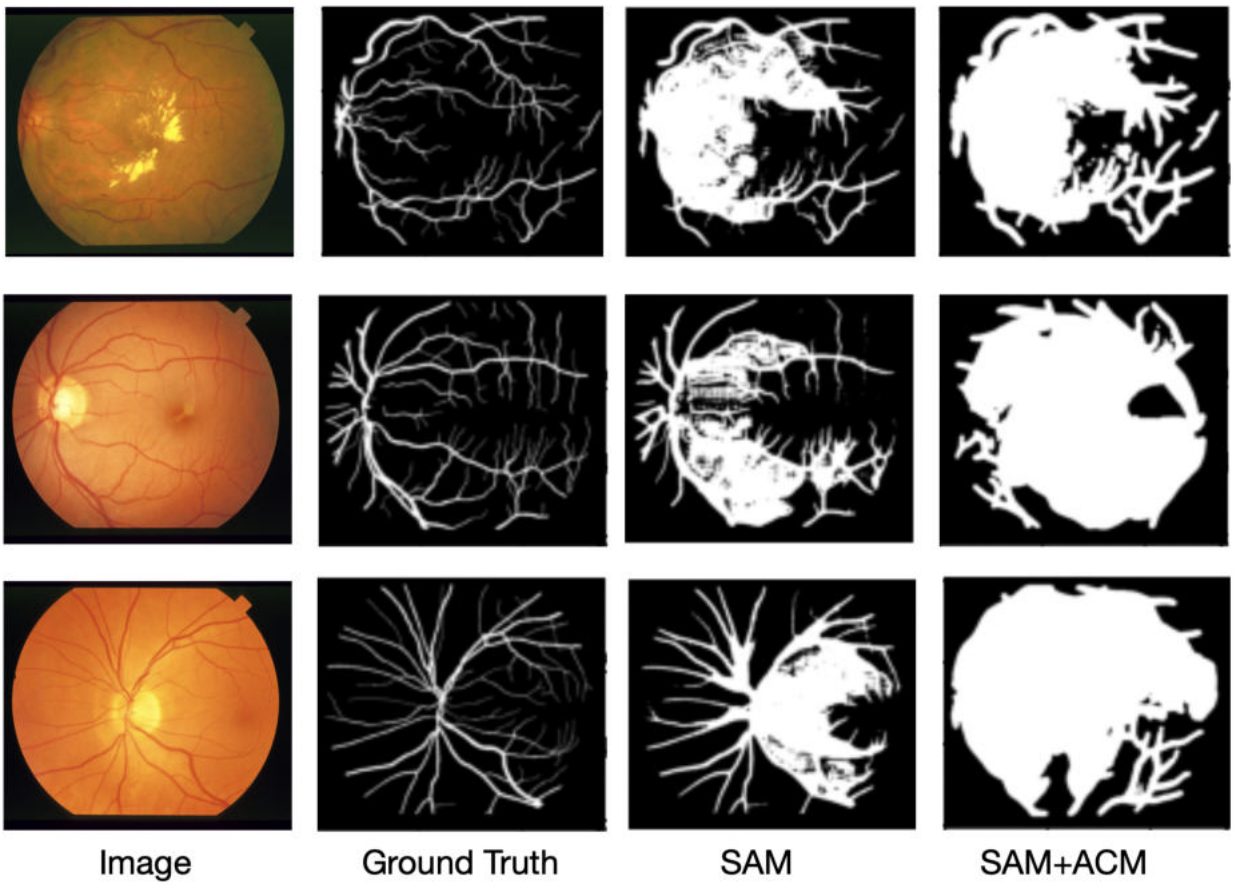


Figure 4.7: Examples of Cases Where Segmentation Failed in the STARE Dataset

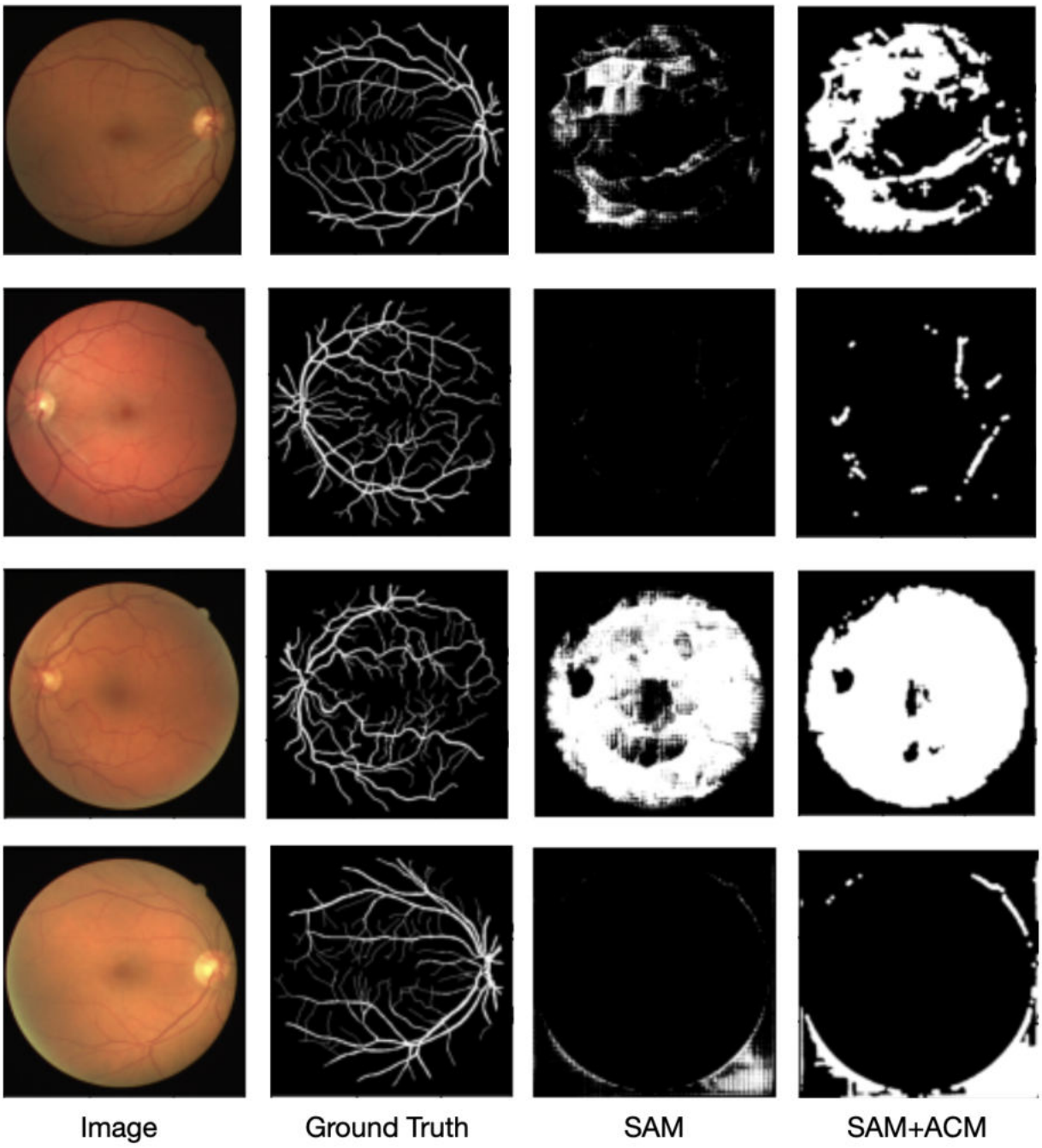


Figure 4.8: Examples of Cases Where Segmentation Failed in DRIVE Dataset

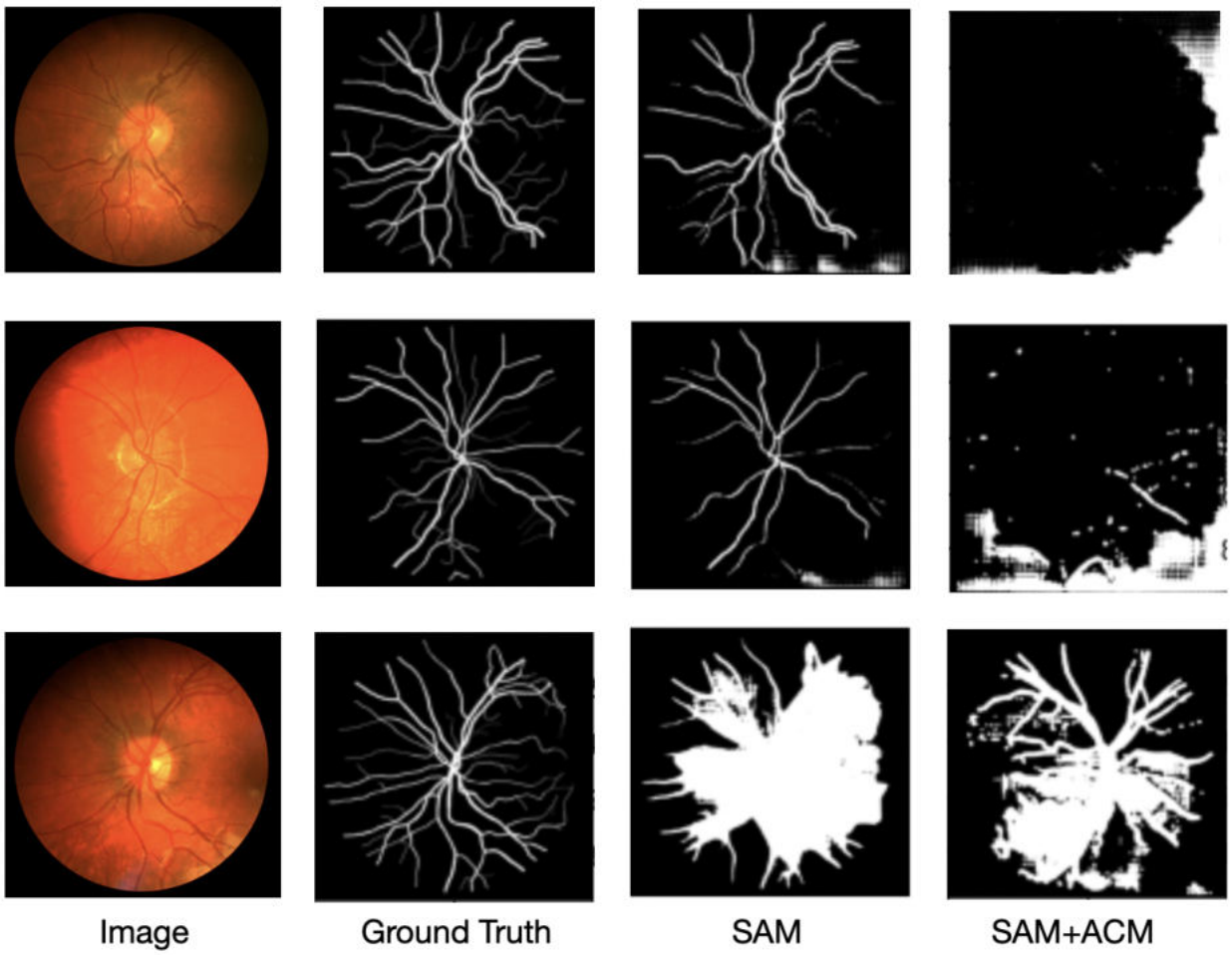


Figure 4.9: Examples of Cases Where Segmentation Failed in the CHASE_DB Dataset

CHAPTER 5

Active Contour Models With Attention for Medical Image Segmentation

edical image segmentation is essential for accurate diagnosis and treatment planning, particularly in identifying anomalies in CT and other imaging modalities. Our objective is to develop a deep learning-based segmentation framework that combines Active Contour Models (ACMs) and attention mechanisms, thus advancing automated medical image segmentation for improved clinical decision-making. Our approach enhances boundary precision and segmentation accuracy in medical images by integrating ACMs with Convolutional Neural Networks (CNNs) and edge detection techniques. Moreover, we investigate the novel idea of fine-tuning ACM hyperparameters by learning them within the CNN and backpropagating the loss through the ACM. Our methodology addresses challenges such as weak boundaries, noise, and variability in anatomical structures, contributing to more robust and interpretable medical image analysis. This chapter develops a prototype hybrid model and thoroughly evaluates it on multiple medical images.

5.1 Introduction

Medical image segmentation is a critical task in medical diagnostics, enabling the precise delineation of anatomical structures and abnormalities in imaging modalities such as CT, MR, and ultrasound (US). However, segmentation remains challenging due to the complexity of medical images, which often exhibit irregular shapes, subtle boundaries, and significant noise. Additionally, effective segmentation often requires models trained on domain-specific datasets, making generalization difficult. Various approaches exist for medical image segmentation,

each offering distinct strengths and limitations.

Deep learning-based segmentation models have gained prominence due to their ability to learn complex patterns from medical images. Notable architectures include U-Net and SegNet, which are widely used in medical image segmentation due to their encoder-decoder structures that help retain spatial information. When integrated with CNNs, attention mechanisms can further enhance the model's ability to focus on fine details, leading to more precise segmentation. The choice of loss function plays a crucial role in optimizing performance, with common functions such as dice loss (DCS), intersection over union (IoU), and binary cross-entropy (BCE) tailored to handle class imbalance and segmentation accuracy.

Edge-based segmentation relies on detecting strong gradients to define object boundaries, but it may struggle with weak edges and noisy images. Recent vision-language models, such as the Segment Anything Model (SAM), leverage pre-trained transformers for zero-shot segmentation, though they may require fine-tuning for domain-specific tasks. Hybrid techniques combine traditional segmentation approaches with deep learning, offering improved accuracy, especially when dealing with weak boundaries or noisy data.

Active Contour Models (ACMs) offer an alternative approach to medical image segmentation by iteratively refining contours to fit object boundaries. ACMs provide several advantages, including adaptive boundary detection that dynamically adjusts contour points to capture complex and irregular shapes. They also allow for integrating prior knowledge by fine-tuning energy terms based on domain-specific constraints, enhancing segmentation robustness. Unlike purely pixel-intensity-based methods, ACMs incorporate internal energy constraints, making them more resilient to image noise. Variants, such as region-based ACMs, further enhance flexibility. ACMs have proven versatile across various imaging modalities, including CT, MRI, and US, and are valuable for organ boundary detection and tumor segmentation tasks.

Our work introduces a novel approach to CNN-based segmentation for medical image datasets by integrating ACMs to enhance boundary precision. A key innovation is the hybrid ACM model, which leverages pretrained edge attention mechanisms and incorporates ACM

logic directly into the CNN architecture. This allows the model to predict hyperparameters via an ACM hyperparameter generator as part of the training process, effectively setting these parameters dynamically rather than relying on manual tuning. By combining the strengths of deep learning and incorporating the DALS Level Set ACM (Hatamizadeh et al., 2019a), our approach aims to improve segmentation accuracy, robustness, and generalization across diverse medical images, marking a significant advancement in the field.

5.2 Methods

In this project, we focused on two different methods. One involves an edge-focused attention module, and the other continues our previous work on integrating ACMs and foundation model.

5.2.1 Edge Segmentation

The edge segmentation module is designed to extract and refine edge features from an input tensor, which is crucial in enhancing segmentation accuracy, particularly around object boundaries. The process consists of four main steps: edge detection, edge thresholding, filling interior regions, and final processing.

The edge detection step applies the Roberts Operator (Roberts, 1963), a simple edge filter that computes horizontal and vertical gradients using a 2×2 convolutional kernel. It calculates the gradient magnitude to highlight regions of significant intensity changes corresponding to edges. Each input channel is processed separately, combining the results into a final edge map.

Next, edge thresholding dynamically determines an adaptive threshold based on edge intensities' mean and standard deviation. This step ensures that only the most prominent edges are retained, producing a binary edge map.

The third step fills interior regions by employing morphological operations to refine the detected edges. Dilation expands edge regions to close small gaps, while erosion shrinks

Figure 5.1: Example result from the edge segmentation module. The Roberts operator applied to image bjorke_86.png (left). Edge-based segmentation using the Roberts operator with thresholding and enhancements to obtain strong binary results (center). The target segmentation (right).

edges to retain only meaningful structures. These operations help fill interior regions within strong edges, ensuring enclosed areas are appropriately marked and improving segmentation consistency.

Finally, final processing enhances the detected edges for better contrast and interpolates the edge map to match the original input dimensions. The result is a refined edge-aware feature map later utilized in the Edge Attention mechanism.

This approach is crucial for segmentation models, as it effectively captures boundary details, leading to sharper object delineation and improved accuracy. However, while it captures major edges, it struggles to find specific abnormalities in the target, as the Roberts operator focuses on general edges rather than finer details. Despite this, it provides a solid foundation for training a neural network to leverage edge-based segmentation. An example result from this approach is shown in Figure 5.1.

5.2.2 Convolutional Neural Network

5.2.2.1 Architecture

The network follows an encoder-bottleneck-decoder structure. The encoder extracts hierarchical features, incorporating attention mechanisms to suppress irrelevant information and refine

feature selection. The bottleneck enhances feature representation. The decoder reconstructs the segmentation mask using transpose convolutions and attention layers, refining spatial details before outputting the final mask via a Sigmoid activation. Edge-based post-processing can further improve boundary accuracy.

5.2.2.2 Data Preprocessing

The data preprocessing pipeline for the edge segmentation model standardizes the dataset by processing grayscale images and their corresponding segmentation masks. Images are resized to 1024×1024 pixels, normalized to the $[0, 1]$ range, and converted into PyTorch tensors. A custom EdgeSegmentationDataset class is implemented to load and preprocess image-mask pairs, ensuring only matching files are considered. The pipeline supports optional transformations like data augmentation (random flips, rotations) and sample size limitations. This setup ensures consistency in input data, enhancing the model's stability and performance during training.

5.2.2.3 Attention Mechanisms

We compare two approaches: Base Attention and Edge Attention, highlighting their differences in feature refinement and edge awareness.

Our model supports Base and Edge Attention mechanisms. Base Attention refines feature importance, while Edge Attention enhances segmentation by incorporating edge detection using the Roberts Operator.

Base Attention learns an attention map by processing the input feature map through a series of 1×1 convolutions, followed by a sigmoid activation to normalize attention values. The model scales the input features using these learned weights, refining feature importance. The process involves four key steps: feature transformation through a 1×1 convolution, refinement using another 1×1 convolution, normalization via sigmoid activation, and feature modulation, where the input features are multiplied by the attention map to selectively enhance relevant regions. However, Base Attention does not explicitly consider edge information, which can

limit segmentation accuracy near boundaries. Since it applies attention uniformly across the feature space, it lacks specialized treatment of high-gradient regions.

Edge Attention extends Base Attention by explicitly incorporating edge information into the attention computation. It applies a separate edge-detection convolution, ensuring that attention focuses on boundary details to improve segmentation performance. This process starts with feature extraction using a 3×3 convolution, followed by edge awareness, where edges are extracted from the input. Then, a convolutional layer transforms edge features into attention weights normalized via sigmoid activation. Finally, these computed attention weights are multiplied by the extracted features, enhancing segmentation precision around boundaries.

5.2.3 Active Contour Models

ACMs rely on several key components that collectively govern the contour evolution process. The intensity image serves as the primary input, representing the image on which the contour evolves. The initial contour defines the starting boundary, critical in the model's convergence and final segmentation outcome. The algorithm and energy functional dictate how the contour evolves, balancing internal forces for smoothness and external forces derived from the image. Additionally, hyperparameters control various aspects of the model's behavior, such as regularization strength and step size, significantly impacting the performance and stability of the segmentation.

Continuing our previous publication, we designed a hybrid retinal-vessel segmentation pipeline that combines a lightweight CNN trained end-to-end on DRIVE, STARE, and CHASE, SAM zero-shot masks generated per image via the Segment Anything Model's ACM refinement, with per-image hyperparameters predicted by fusing CNN and SAM features.

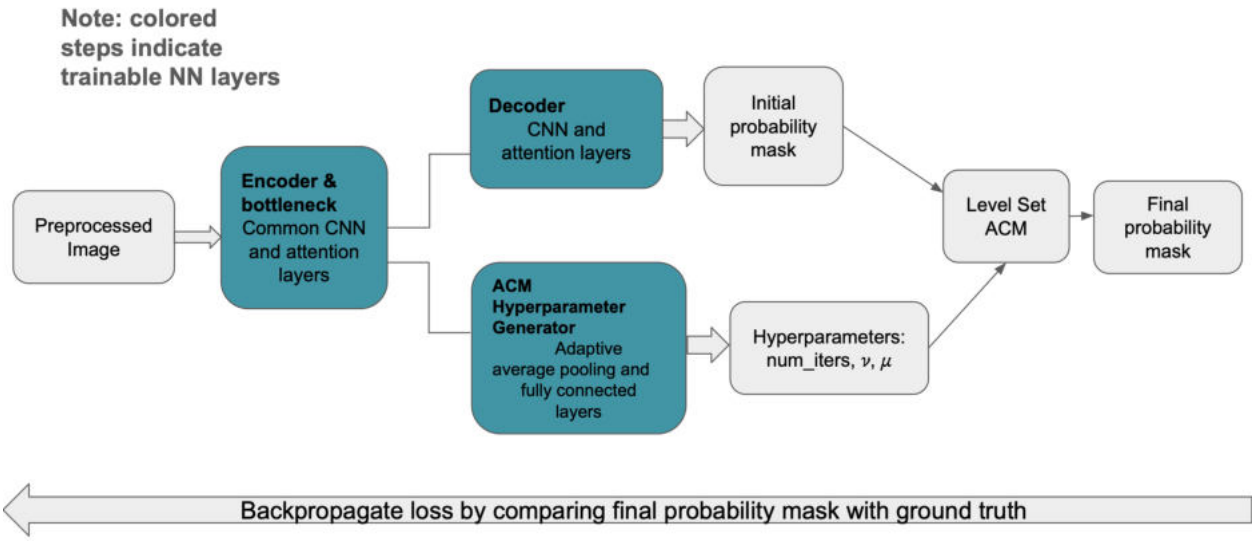


Figure 5.2: Hybrid CNN + ACM system diagram

5.2.4 Overall Architecture

5.2.4.1 System Diagram

Figure 5.2 illustrates our hybrid ACM system. Enabling the ACM functionality in our architecture will include the 'ACM Hyperparameter Generator,' its outputs, and 'Level Set ACM', which will otherwise be disabled.

ACM Hyperparameter Generator The ACM Hyperparameter Generator uses adaptive average pooling as a flattening mechanism, followed by a fully connected layer with a ReLU activation and another fully connected layer with a Sigmoid activation. It produces three outputs: num_iters, ν, and μ, which are scaled to the correct ranges: 0–100 for num_iters, 0-10 for ν, and 0-1 for μ. The main novelty is to leverage the properties learned through the previous layers (encoder and bottleneck) and further use fully connected layers to learn image properties relevant to determining an optimal set of ACM hyperparameters that will result in effective ACM contour evolution.

Incorporating ACM into the Neural Network Incorporating the DALS Level Set ACM into the neural network required converting the original TensorFlow implementation to

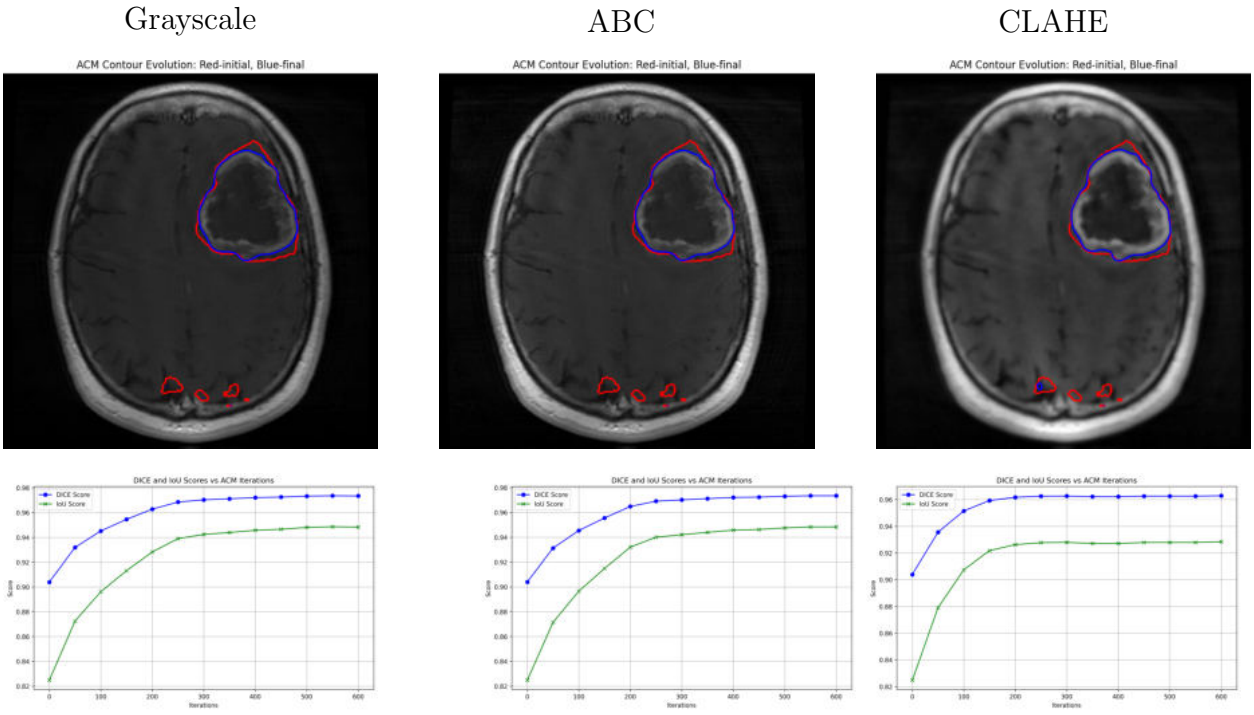


Figure 5.3: Experiment 1: DALS LSA demonstration on a brain image using a pretrained-CNN generated initial contour, $\nu = 5$, $\mu = 0.2$, number of iterations = 600.

PyTorch, as our CNN baseline was in PyTorch. For the brain demonstration in Figure 5.3, the torch version of DALS LSA takes 1.770149 seconds, while the TensorFlow version takes 92.741031 seconds. This is a speedup by a factor of 52.4. Hence, in addition to being compatible with existing torch architecture, a torch version of DALS significantly reduces the training time.

Additionally, the DALS LSA uses non-differentiable scipy functions to calculate the initial signed distance map from the probability mask. Figure 5.4 illustrates how this issue can cause ineffective learning. The initial signed distance map is updated through the ACM iterations to get the final probability mask. Leaving this out of the gradient computational graph will hinder backpropagation along this important path and effective learning. To address this, we aimed to replicate this logic using differentiable operations, ensuring compatibility with backpropagation. Making the DALS LSA fully differentiable should theoretically allow for more accurate gradient updates, improving the overall learning process.

Figure 5.5 illustrates the differences in ACM contour evolution and demonstrates the

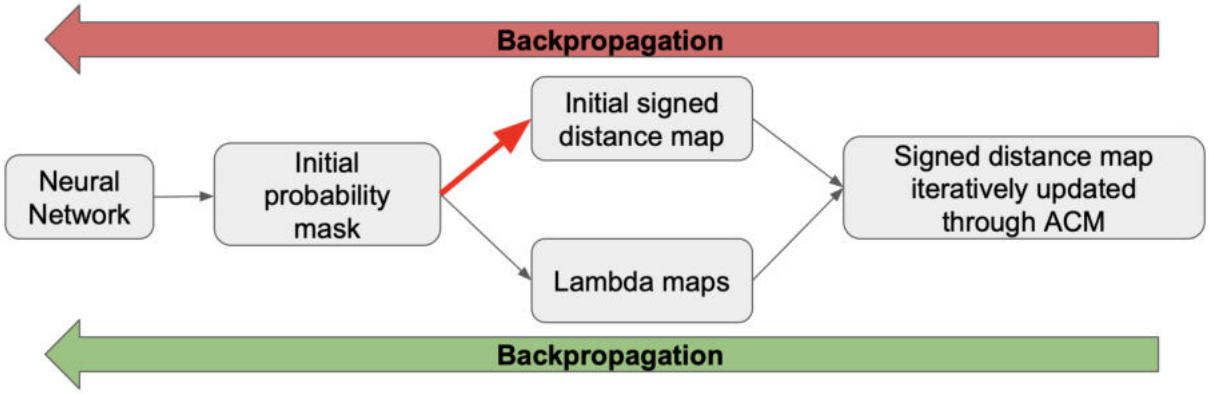


Figure 5.4: Differentiability and Backpropagation

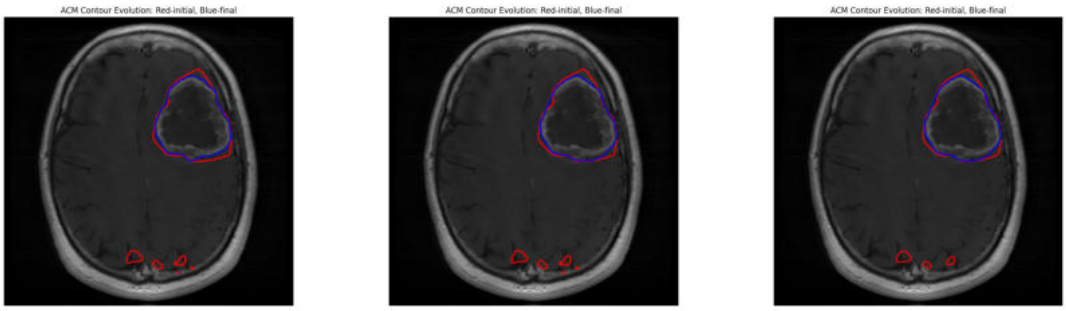


Figure 5.5: Demonstration of Tensorflow, Torch, and fully differentiable Torch versions of the DALS Level Set ACM.

success of our approach through the DALS brain demonstration results, showing that the converted DALS LSA logic remains intact.

5.2.4.2 Training

The training procedure for the edge segmentation model follows a structured approach for efficient learning and generalization. The model is initialized, and its configuration is saved for reproducibility. Training is conducted in mini-batches, where predictions are made, and the loss is computed against ground truth masks. A learning rate scheduler adjusts the learning rate based on training progress, and gradient clipping prevents instability by controlling large gradients. Early stopping halts training if no improvement occurs after several epochs. Loss is logged and visualized to monitor progress, ensuring stable, efficient training and avoiding

overfitting.

Note about training hybrid ACM model: Since Active Contour Models (ACMs) perform poorly when initialized with suboptimal or random probability masks, the training process for the hybrid ACM architecture incorporates a weight initialization strategy. Specifically, before training begins, the weights of a reasonably trained edge attention model are loaded into the non-ACM-related layers of the neural network. This serves as a pretraining step, providing a more stable starting point for training the hybrid ACM model.

5.2.4.3 Testing

The testing process evaluates the trained model's performance on a test dataset. It begins by organizing the output folder based on the model's timestamp and epoch. The model is set to evaluation mode to ensure consistent behavior. Loss tracking and metric computation are initialized, and the model processes each batch in the test set to generate predictions. Loss is calculated, and predictions are converted to probabilities for metric evaluation. Visual results, including input-output-target comparisons, are saved. After processing all batches, performance metrics like accuracy and IoU are computed and stored. The results are saved for further analysis, providing insights into the model's effectiveness.

5.2.4.4 Evaluation

At test time, we compute four segmentation modes per image:

$$\{\texttt{cnn\_base},\ \texttt{cnn\_acm},\ \texttt{sam\_zero},\ \texttt{sam\_acm}\} \tag{5.1}$$

and report Jaccard (IoU) and F_1 scores, averaged over DRIVE, STARE, and CHASE. We also log the learned average iters, ν, μ to interpret the contour refinement behavior.

Metric	Base Attention Model	Edge Attention Model	Hybrid ACM Model
Average Test Loss	0.0257	0.0229	0.0389
AUROC	0.9882	0.9921	0.8758
AUC	0.9882	0.9921	0.8758
Precision	0.7624	0.7570	0.8175
Recall (Sensitivity)	0.5360	0.6477	0.4023
F1 Score	0.6294	0.6981	0.5393
IoU	0.3412	0.4285	0.3820
Dice Score	0.4570	0.5519	0.4427

Table 5.1: Performance metrics for the Base Attention Model (50 epochs trained), Edge Attention Model (50 epochs trained), and Hybrid ACM Model (10 epochs trained with eight training images) Evaluation.

5.3 Experiments and Results

5.3.1 Considerations

Our model is trained with an initial learning rate of 0.001, which is reduced by a factor of 0.5 if no improvement is greater than 1e-4 after three epochs. The dataset used is the COVID-19 CT scan lesion segmentation dataset (COVID-19, 2020).

The paper discusses the use of different training configurations for various attention modules. The base attention module utilizes a model trained for 50 epochs on the entire COVID-19 CT scan dataset with an 80/20 train-test split ratio. The edge attention module also uses a model trained for 50 epochs on the full dataset with the same train-test split. Due to time complexity constraints, the hybrid ACM module leverages the pretrained edge attention model (50 epochs) and is fine-tuned on a smaller subset of 8 training images and two testing images from the complete dataset for a total of 10 epochs.

5.3.2 Overall Results

In testing, the results are better and are produced faster for the Edge Attention model than the Base Attention model. It can be seen from Table 5.1 that the edge attention model has a

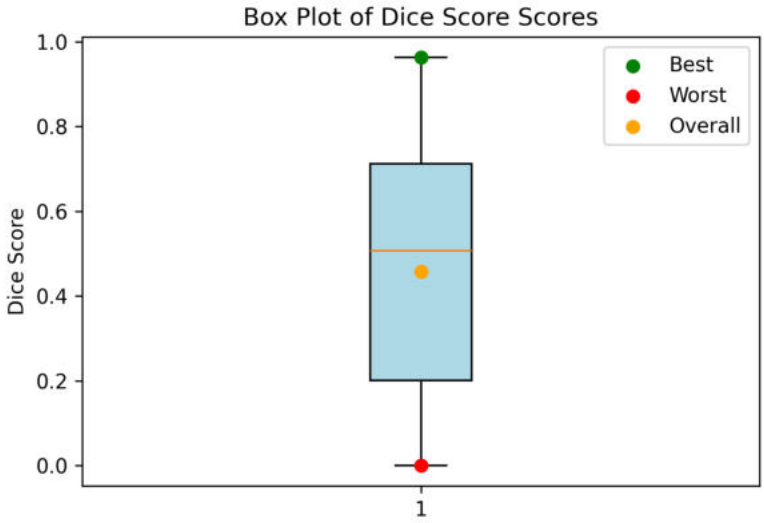


Figure 5.6: Dice Score Box Plot comparing results in the Base Attention Model

lower average test loss and higher values for other scores like IoU and DICE than the base attention model. While the Hybrid ACM model has slightly greater loss than the Base and Edge Attention CNNs, its ability to have hyperparameter tuning for the ACM provides novel results and a foundation for future adaptation and training.

5.3.3 CNN With Base Attention

For a clearer comparison of the overall performance for the base attention CNN module, please refer to Figure 5.6. While the average Dice Score of 0.4570 is respectable for a model with only a basic attention module, there is a significant disparity between the best and worst-case scenarios.

The image that stands out as the best performer is shown in Figure 5.7. It achieved the highest precision among all the test images from the base attention model, with a Dice Score of 0.9628, which is an excellent result. While this is a strong performance, there is still potential for further improvement. Overall, these results highlight the model's strong capability and suggest opportunities for additional refinement.

The image that demonstrates the worst-case scenario is shown in Figure 5.8. This image exhibited the lowest precision among all the test images from the base attention model,

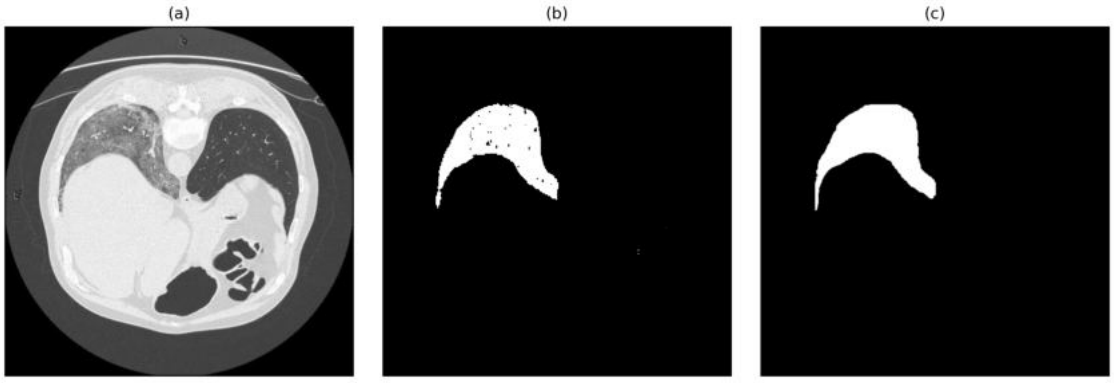


Figure 5.7: Best Base Attention Model result with a precision score of 0.9630 and DICE score of 0.9628 for the Jun_radiopaedia_7_85703_0_case20_39.png image. (a) Input image in the dataset. (b) Predicted segmentation output from the trained model. (c) Target segmentation.

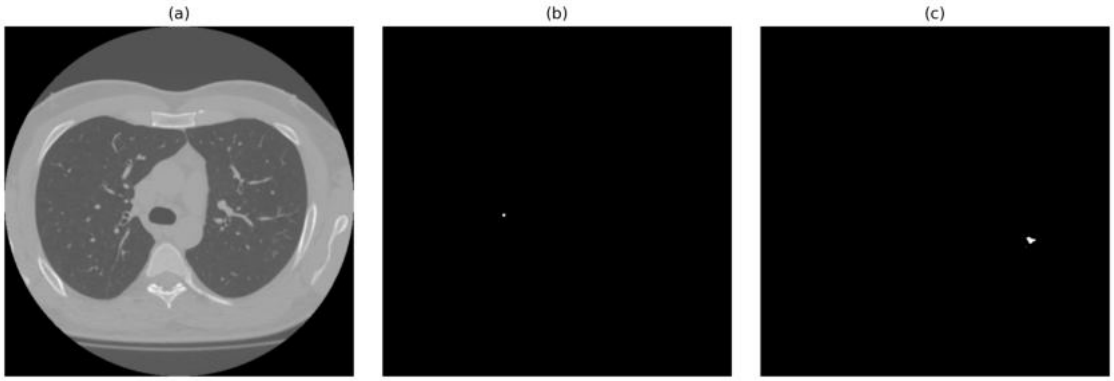


Figure 5.8: Worst Base Attention Model result with a Dice Score of 0.0 for the Morozov_study_0295_28.png image. (a) Input image in the dataset. (b) Predicted segmentation output from the trained model. (c) Target segmentation.

achieving a Dice Score of 0, indicating complete prediction failure in this instance. As illustrated in Figure 5.8, the final target image lacks sufficient detail, likely confusing the model and impairing its ability to identify and highlight the relevant features correctly.

5.3.4 CNN with Edge Attention

For a clearer comparison of the overall performance for the edge attention CNN module, please refer to Figure 5.9. This has an average Dice Score of 0.5519, which is greater than the average Dice score of the base attention model.

An image to highlight is Figure 5.10. This image had the best precision out of all the

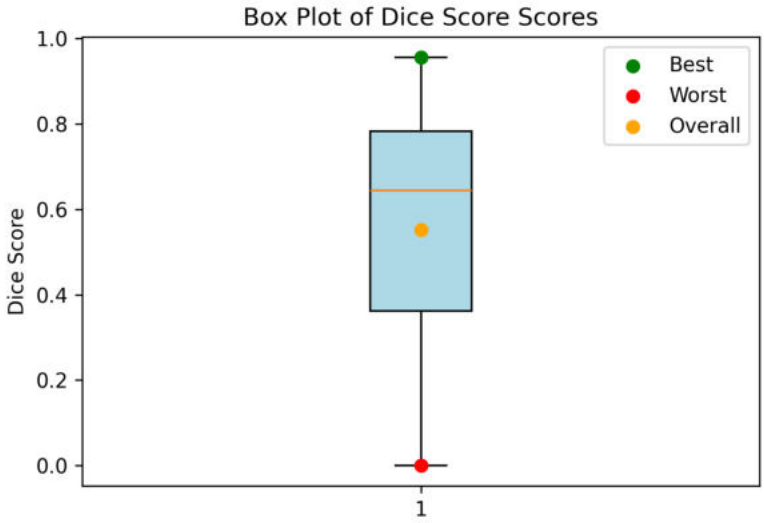


Figure 5.9: Dice Score Box Plot comparing results in the Edge Attention Model

images tested from the edge attention model, with a Dice Score of 0.9561. While this max result is not better than the Base Attention model, the average test DICE score for edge attention is still better, as discussed before. Moreover, comparing the dice score box plots in Figure 5.6 and Figure 5.9 you can see that in edge CNN most of the test data dice scores fall in higher range compared to the base CNN case.

Another image to highlight is Figure 5.11. This image had the worst precision out of all the images tested from the edge attention model, with a Dice Score of 0.0. Similar to Figure 5.8, Figure 5.11 struggled to find the features to emphasize in segmentation, which is clear in its target result, which is a very small segmentation.

5.3.5 Pretrained CNN with ACM

The main innovation of this approach is the ability to determine optimal ACM hyperparameters using fully connected layers, which significantly enhances the effectiveness of ACM contour evolution. Despite being trained on a small set of images atop the pretrained Edge Attention model, the results show a notable improvement in the performance of the ACM hyperparameter generator and the Hybrid ACM model. As shown in Table 5.1, the Hybrid ACM model achieves the best precision and comparable scores for all other metrics, even

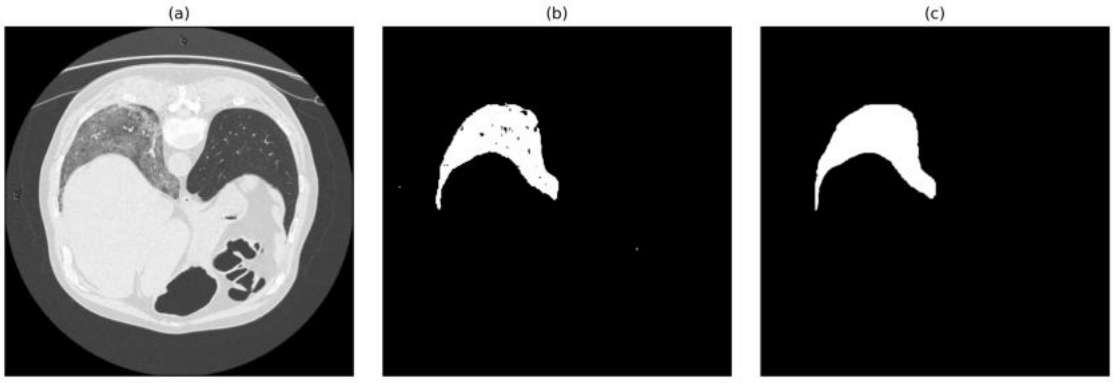


Figure 5.10: Best Edge Attention Model result with a Dice Score of 0.9561 for the Jun_radiopaedia_7_85703_0_case20_39.png image. (a) Input image in the dataset. (b) Predicted segmentation output from the trained model. (c) Target segmentation.

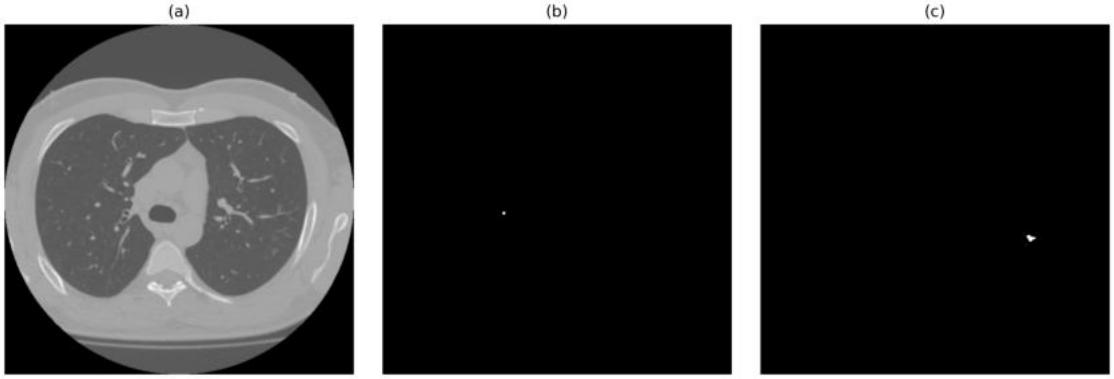


Figure 5.11: Worst Edge Attention Model result with a Dice Score of 0.0 for the Morozov_study_0295_28.png image. (a) Input image in the dataset. (b) Predicted segmentation output from the trained model. (c) Target segmentation.

though it was only trained for 10 epochs.

An image to highlight is Figure 5.12. This image had the best precision out of all the images tested from the Hybrid ACM model. The Dice Score is 0.8569. Despite this DICE score being lower than the best base and edge attention results, this training occurred on fewer images and with only 10 epochs.

Another test result is Figure 5.13, and it had a non-zero Dice Score of 0.0284.

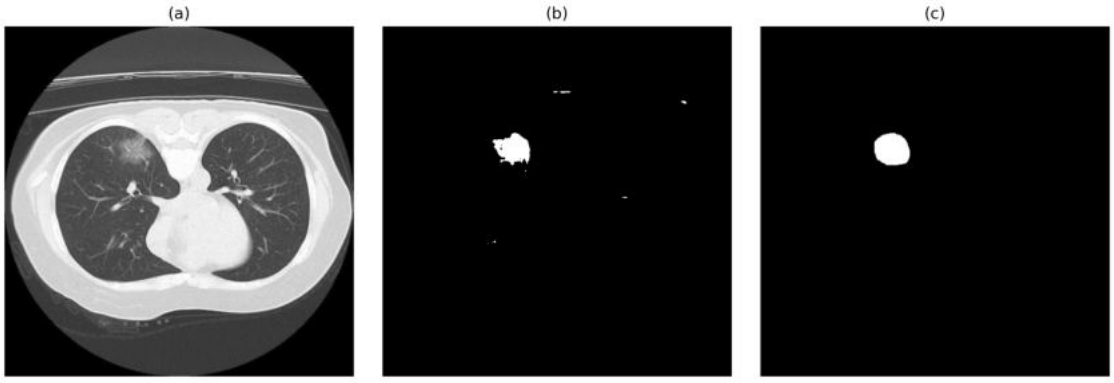


Figure 5.12: Hybrid ACM Model result with a Dice Score of 0.8569 for the Jun_radiopaedia_29_86491_1_case16_20.png image. (a) Input image in the dataset. (b) Predicted segmentation output from the trained model. (c) Target segmentation.

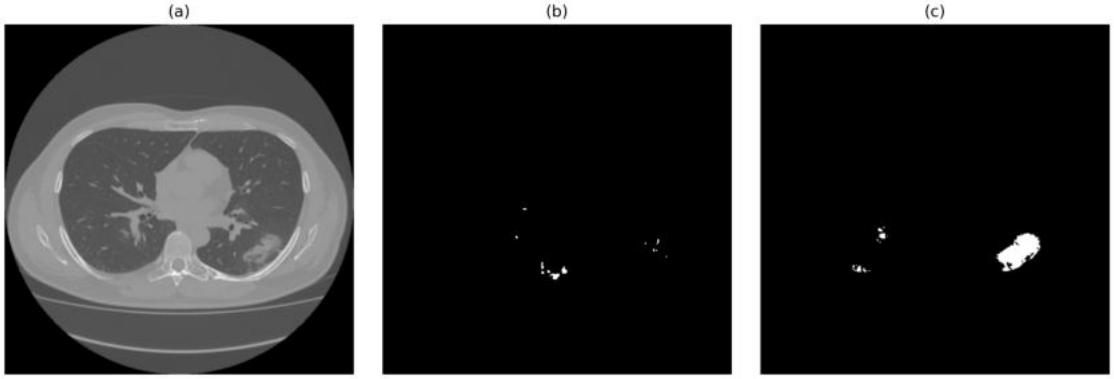


Figure 5.13: Hybrid ACM Model result with a Dice Score of 0.0284 for the Morozov_study_0258_22.png image. (a) Input image in the dataset. (b) Predicted segmentation output from the trained model. (c) Target segmentation.

5.4 Discussion

The overall results from the Hybrid ACM model are novel in their approach to utilizing the ACM hyperparameter generator and provide a profound innovation in their possibility to learn future contours with a combined neural network and level-set ACM. When trained and tested on small datasets, the Active Contour Model (ACM) with hyperparameter tuning demonstrates good performance. However, its application to larger datasets presents significant computational challenges due to the high time and memory consumption associated with gradient tracking over hundreds of ACM iterations. Since each iteration contributes to refining the segmentation mask, the backpropagation process becomes increasingly expensive,

leading to inefficiencies in training. This limitation makes large-scale training impractical without substantial computational resources. Despite this, the model's strong results on small datasets highlight its potential effectiveness, suggesting that it is well-suited for applications requiring high-quality segmentation on limited data samples. Future optimizations could focus on reducing computational overhead while preserving the model's ability to learn optimal ACM hyperparameters.

Future work should focus on several key areas for improvement. This includes exploring additional image transformations and pre-processing techniques to enhance model performance. Further training will be conducted on various variations to increase the model's robustness. Additionally, model improvements such as incorporating batch normalization and optimizing the ACM to enhance speed and memory efficiency will be explored. A sliding window approach will also be investigated to improve the handling of large images during training. Finally, further training will be conducted with the optimized ACM to refine its performance.

CHAPTER 6

Boundary-Aware SwinUNETR for Medical Image Segmentation

Medical image segmentation presents several challenges, including high variance in the appearance and shape of target regions, high dimensionality of volumetric data, and label imbalance. Recent transformer-based architectures like SwinUNETR have shown promising results but struggle with precise boundary delineation, particularly for small structures. We present Boundary-Aware SwinUNETR, a novel approach that incorporates deep supervision mechanism into the SwinUNETR architecture to enhance boundary detection capabilities. Our method leverages intermediate features from multiple resolution levels to generate boundary attention maps, which are combined with the main segmentation pathway. The resulting segmentation performance on the Medical Segmentation Decathlon Pancreas dataset shows an average dice improvement of 1%.

6.1 Introduction

The field of computer vision has seen remarkable development over the last decade, with a large focus on domains such as classification, segmentation, and image captioning. In particular, the domain of segmentation (including Semantic segmentation) has seen continuous improvements, from the advent of architectures such as the U-Net (Ronneberger et al., 2015) to the development of Vision Transformers (Dosovitskiy et al., 2020). Transformer networks, in particular, have provided a new approach to segmentation problems, leveraging self-attention (Vaswani et al., 2017) blocks to model long-range global information better. In addition, these transformer networks allow for pre-training on related segmentation tasks in the domain,

making improving performance on newer downstream tasks easier.

In spite of all these advancements, medical image segmentation continues to stand out as a challenging task. Some key challenges associated with segmentation in the medical image domain include: a) a large imbalance in labels, particularly involving organ and tumor segmentation, and b) stark differences between regular images and various imaging modalities, including CT and MR. Additionally, representative datasets are often limited in the medical imaging domain, thus making it difficult to develop models with reliable performance.

3D medical image segmentation is challenging due to the high variance in appearance and shape of target regions, high dimensionality of volumetric data, and label imbalance. These challenges become more limiting in multi-organ and tumor segmentation tasks where boundaries between different structures can be ambiguous or hard to segment. While Convolutional Neural Networks (CNNs) have traditionally dominated this field (Ronneberger et al., 2015; Çiçek et al., 2016; Milletari et al., 2016),

Some recent work, however, has shown encouraging improvements in performance on these tasks, including complex transformer models such as SwinUNETR (Tang et al., 2022; Hatamizadeh et al., 2022). This architecture, which leverages pre-training an encoder on tasks such as contrastive learning, masked volume inpainting, and 3D rotation prediction, shows SOTA performance on several segmentation tasks, such as the Decathlon Challenge (Antonelli et al., 2022) tasks.

To address these challenges, we propose Boundary-Aware SwinUNETR, a boundary-aware architecture incorporating a dedicated boundary attention stream and multi-level deep supervision into the SwinUNETR framework.

We seek to extend the SwinUNETR architecture by providing additional context to focus on the boundaries of the regions of interest (organ/tumor). In particular, inspired by efforts such as the one by Hatamizadeh et al. (2019b), the present chapter focuses on adding a boundary stream feeding off the outputs of the encoder. We also use an associated boundary stream "Edge loss" component, which helps increase performance.

Our key contributions are:

- A novel boundary attention module that generates edge-focused attention maps from intermediate encoder features and integrates them into the main segmentation pathway.
- A comprehensive loss function that combines Dice loss for region-based segmentation with a custom edge loss for boundary delineation, weighted appropriately to balance their contributions.

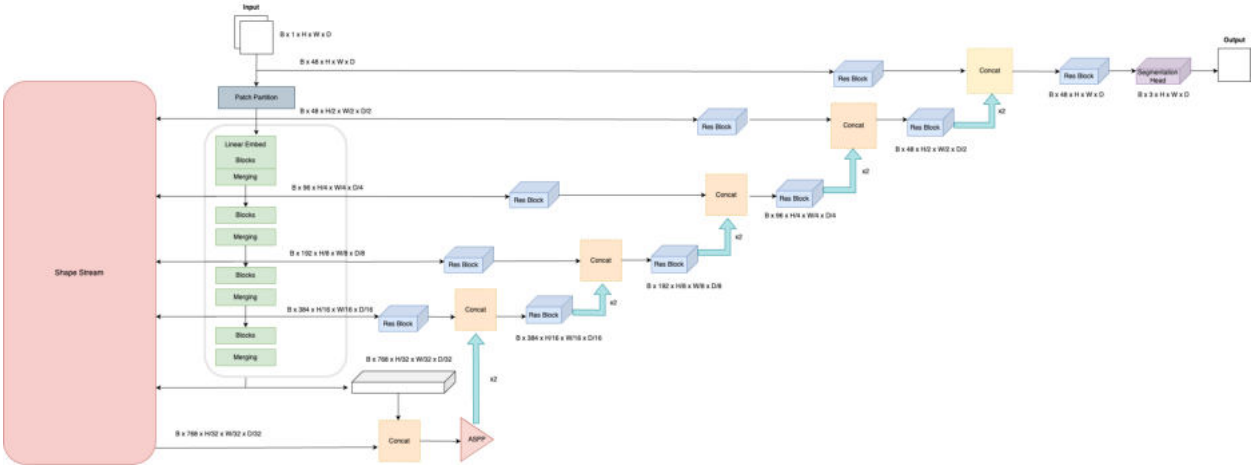
6.2 Methods

6.2.1 Architecture

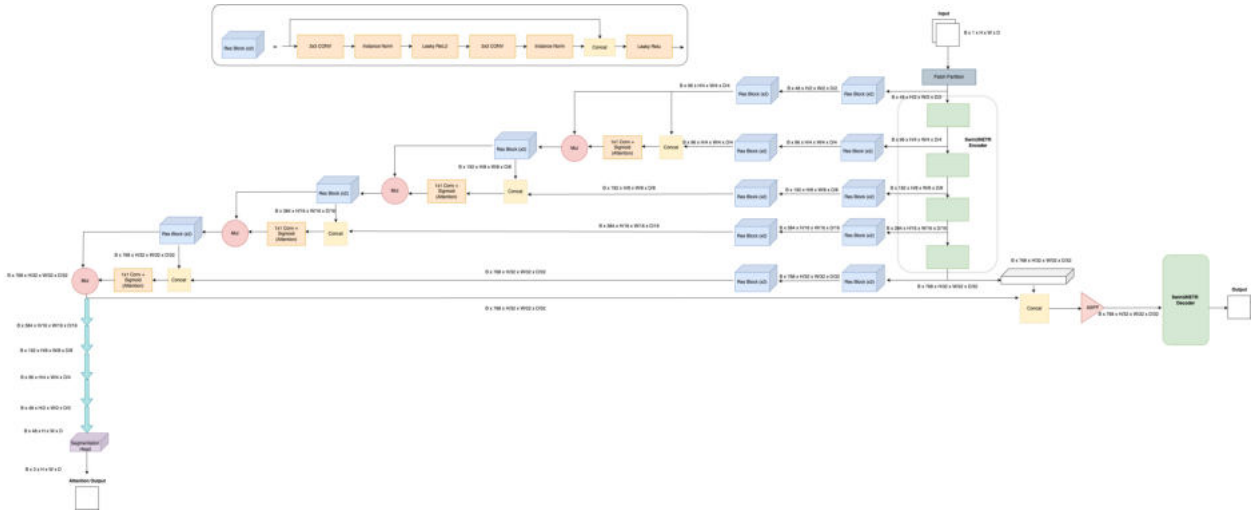
Figure 6.1 depicts the overall architecture of the Boundary Aware SwinUNETR model. There are two main components, the SwinUNETR encoder-decoder model, which is described in Section 6.2.1.1, and the Boundary Stream, which is described in Section 6.2.1.2.

6.2.1.1 SwinUNETR

The SwinUNETR model (Tang et al., 2022) leverages a Swin Transformer as the encoder and is connected to a CNN-based decoder at different resolutions in a U-shaped network. The encoder makes use of a shifted window-based self-attention approach. An input of size $(H \times W \times D \times C)$ is partitioned into a sequence of $H/h \times W/w \times D/d$ partitions, where $(h \times w \times d)$ is the patch resolution, which are then projected into an F-dimensional space, where F is the feature size. These partitions are then split into non-overlapping windows at each layer l, and self-attention is computed within each of these windows (Liu et al., 2021a). If the window is of dimension $(R \times R \times R)$, then the windows are shifted in layer $l+1$ by $R/2, R/2, R/2$ voxels. The self-attention equations (Tang et al., 2022) at layer l and $l+1$



(a) SwinUNETR encoder-decoder



(b) Boundary Stream

Figure 6.1: Overall architecture for the Boundary Aware SwinUNETR model.

can be written as follows:

$$\hat{a}^{l} = \text{W-MSA}(\text{IN}(\hat{a}^{l-1})) + \hat{a}^{l-1}, \tag{6.1}$$

$$\hat{a}^{l} = \text{MLP}(\text{IN}(\hat{a}^{l})) + \hat{a}^{l}, \tag{6.2}$$

$$\hat{a}^{l+1} = \text{SW-MSA}(\text{IN}(\hat{a}^{l})) + \hat{a}^{l}, \tag{6.3}$$

$$\hat{a}^{l+1} = \text{MLP}(\text{IN}(\hat{a}^{l+1})) + \hat{a}^{l+1}, \tag{6.4}$$

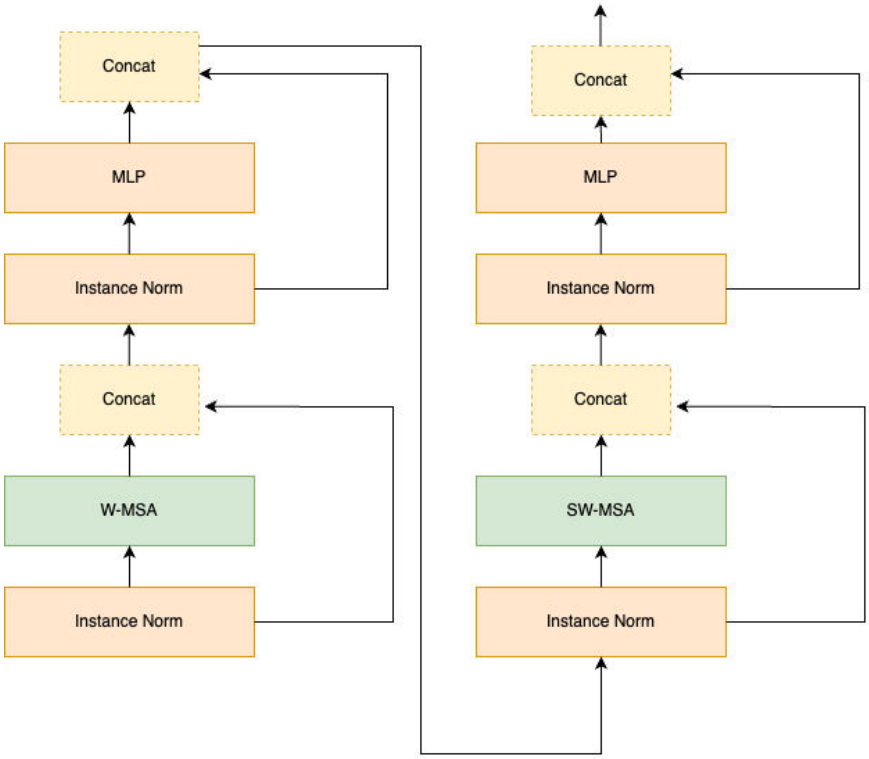


Figure 6.2: Internal structure of a Swin Transformer block (Tang et al., 2022).

where W-MSA refers to regular multi-head self-attention, SW-MSA refers to window partitioning multi-head self-attention, MLP refers to a multi-layer perceptron, and IN refers to an Instance Normalization (Ulyanov et al., 2016) layer.

Figure 6.2 demonstrates the internal structure of an encoder Swin-Transformer block. The encoder in this model is composed of 4 stages of these two-block units, with the hierarchical representation at various spatial resolutions serving the purpose of multi-scale feature extraction for downstream tasks such as segmentation. The decoder shown in Figure 6.1a creates a U-shaped structure with the encoder outputs at various resolutions, and it is comprised primarily of CNN blocks. At each level, the features from the previous level of decoding are up-sampled via transpose convolutions and concatenated with the encoder outputs at the current level. The final node helps compute the segmentation output by converting the feature vector into the required n-channel output.

6.2.1.2 Shape Stream and Attention

The boundary stream, shown in Figure 6.1b, takes inspiration from the Attention layer of the model by Hatamizadeh et al. (2019c). The encoder intermediate feature vectors serve as inputs to this stream. At the first level, the first hidden-state feature vector is down-sampled by a factor of 2 and concatenated with the second hidden-state feature vector. This concatenated feature vector is then fed into a 1×1 convolution followed by a sigmoid layer, and the output attention map is multiplied by the first hidden-state feature vector. This output, denoted as $attn_0$, is the top layer attention output.

This attention output is down-sampled using a 3×3 convolution, and is then concatenated with the encoder hidden-state feature vector at the next level. Similar computations as above are repeated. In general, the computation of the layer l attention output can be written as

$$\alpha_l = \sigma\big(C_{1\times1}(t_{l-1} \| h_l)\big), \quad \text{attn}_l = t_{l-1} \odot \alpha_l, \quad t_l = C_{3\times3}(\text{attn}_l), \tag{6.5}$$

where t_l is the output obtained by passing the attention output from layer l into a down-sampling transition block.

The final output from the shape stream is utilised in two different ways. First, it is concatenated with the bottleneck output of the encoder, which is then passed into an ASPP (Chen et al., 2016) layer before being passed into the decoder. In parallel, the boundary stream output is also upsampled using transpose convolutions and passed through a segmentation head, similar to the output at the end of the decoder. This final attention logit is compared against the 3D Spatial Gradient of the label to obtain the boundary stream loss, the details of which are discussed next.

6.2.2 Loss Function

The overall loss function is

$$\text{Loss} = \lambda_1 L_{\text{Dice}}(y_{\text{pred}}, y_{\text{true}}) + \lambda_2 L_{\text{Dice}}(s_{\text{pred}}, s_{\text{true}}) + \lambda_3 L_{\text{CE}}(s_{\text{pred}}, s_{\text{true}}) + L_{\text{DS}}, \tag{6.6}$$

where L_{Dice} denotes the Dice Loss, L_{CE} denotes the categorical cross entropy loss, and

$$L_{\text{DS}} = \sum_{l=0}^{3} w_l \big(\lambda_2 \, L_{\text{Dice}}(\text{attn}_l, s_{l,\text{true}}) + \lambda_3 \, L_{\text{CE}}(\text{attn}_l, s_{l,\text{true}})\big) \tag{6.7}$$

is the deep supervision loss. Additionally, y_{pred} refers to the decoder output, y_{true} refers to the label, s_{pred} refers to the up-sampled output from the boundary stream (i.e., attention logit), and s_{true} refers to the 3D Spatial Gradient of the label, which serves as the edge map for the label.

The edge loss, denoted by L_{CE}, is a weighted categorical cross-entropy loss (as opposed to a weighted binary cross-entropy loss proposed by Hatamizadeh et al. (2019b), which can be written as

$$L_{\text{CE}} = -\sum_{i=1}^{n} \beta_i \log\left(\frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}\right) y_i, \tag{6.8}$$

where n is the number of output classes, and where the weights are obtained as

$$\beta_i = \frac{\text{Total number of pixels in boundary map } s_{\text{true}} - \text{Number of boundary pixels of class } i}{\text{Total number of pixels in boundary map } s_{\text{true}}}. \tag{6.9}$$

The deep supervision loss L_{DS} helps make the model more robust to various target sizes by ensuring the boundary matches at various resolutions, as first proposed by Wang et al. (2015). Internally, it comprises the same dice loss and weighted cross-entropy loss used to evaluate s_{pred}. These losses are weighted by a scale weight w_l, which starts at 0.5 and reduces by half as the layer number l increases, thus ensuring that higher spatial resolutions contribute more to the loss. The attention output at layer l of the shape stream is attn_l, and $s_{l,\text{true}}$ is s_{true} interpolated to match the dimensions of attn_l.

6.3 Experiments and Results

6.3.1 Datasets

The primary dataset was the Pancreas dataset released as part of the Medical Segmentation Decathlon (MSD) Challenge. We disregarded the test set for this experiment since it does not have publicly released labels. The train set, which served as the experiment's dataset, comprises 281 images. This was split into 5-folds across which experiments were run. The images are portal-venous contrast CT images, with a large imbalance between the organ's size and the tumor. Results were also generated using the train set of the Liver dataset, which was also a part of the MSD challenge to compare performance across datasets. This dataset comprises 131 portal-venous contrast CT images, again split across 5-folds.

6.3.2 Implementation Details

The experiments were run on two main hardware resources. The first cluster uses a single Tesla V100 GPU with 32 GB memory and 24 CPU cores. The second cluster uses an RTX A6000 GPU with 50 GB memory and 8 CPU cores. For each experiment, the number of workers is set to the maximum available workers on the machine. A learning rate of 2×10^{-4} was used for all experiments, along with the AdamW optimizer (Loshchilov et al., 2017). Each model was fine-tuned for 600 epochs with a batch size of 1, and used the pre-trained encoder provided. For the loss function, the lambda weights were set to ($\lambda_1 = 1.0$), ($\lambda_2 = 0.5$), and ($\lambda_3 = 0.1$). The scaling intensity parameters and augmentation probabilities for experiments with each dataset were chosen as prescribed by Tang et al. (2022). However, some changes were made on the spacing and ROI sizes to allow faster computation on the available resources, as noted below:

- For experiments involving the pancreas dataset, the training samples are cropped to a size of $64 \times 64 \times 64$ and interpolated to an isotropic voxel spacing of 1.5.
- For experiments involving the liver dataset, the training samples are cropped to a size of $96 \times 96 \times 96$ and interpolated to an isotropic voxel spacing of 1.5.

Model	Pancreas Dice	Tumor Dice	Average Dice
SwinUNETR	0.769441248	0.452041274	0.610741308
Boundary-Aware SwinUNETR	0.783133084	0.459104462	0.621118788

Table 6.1: 5-fold cross-validation scores on the Pancreas dataset.

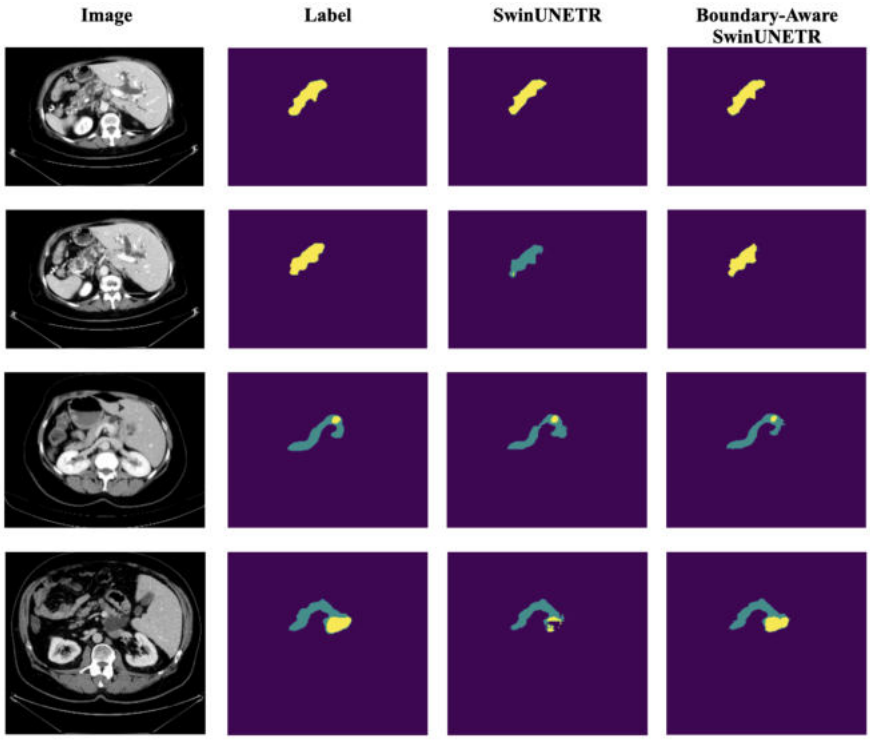


Figure 6.3: Qualitative segmentation results for SwinUNETR and Boundary-Aware SwinUNETR. Yellow pixels indicate the pancreas, and Blue pixels indicate the tumor.

6.3.3 Results

In this section, we report the results of the various experiments conducted. Table 6.1 reports the Dice scores of the Boundary-Aware SwinUNETR architecture compared to SwinUNETR on the MSD Pancreas dataset. Table 6.1 shows the average Dice scores from five-fold cross-validation. As observed in the table, the Boundary-Aware SwinUNETR model outperforms SwinUNETR on both the Pancreas Dice score (by a factor of 1.4%) and the Tumor Dice score (by a factor of 0.7%). Figure 6.3 shows qualitative results for a comparison between the two models.

Loss Function	Pancreas Dice	Tumor Dice	Average Dice
Without Edge Loss and Deep Supervision	0.758926	0.42427203	0.59159905
Without Deep Supervision	0.77432615	0.4422641	0.60829514
With Edge Loss and Deep Supervision	0.77774006	0.44873494	0.6132377

Table 6.2: Effect of Edge Loss and Deep Supervision.

Model	Liver Dice	Tumor Dice	Average Dice
SwinUNETR	0.946518274	0.62109556	0.80081126
Boundary-Aware SwinUNETR	0.94881628	0.625404694	0.80404893

Table 6.3: 5-fold cross-validation scores on the Liver dataset.

Further experiments assessed the impact of the loss described in Section 6.2.2. Table 6.2 reports the resulting Dice scores absent different loss components. Note that the results in this section are obtained using only a single model instead of averaging cross-validation results. As the results in Table 6.2 show, both the Edge Loss and Deep Supervision Loss are crucial to the improved performance.

The final set of results assesses the performance on a secondary dataset, namely the MSD Liver dataset. Again, the results reported are averaged across a five-fold cross-validation. Table 6.3 reports the results of this set of experiments. In the case of the Liver dataset, the model performs marginally better than the standard SwinUNETR model, both on Liver Dice (0.2%) and Tumor Dice (0.4%).

6.4 Discussion

We presented an improvement on the SwinUNETR transformer model for image segmentation. The project demonstrated the utility of an additional boundary stream that focuses on the shape of the segmentation target. Our results on the Pancreas dataset show that this architecture has promise in improving medical image segmentation.

However, a key issue with this model remains the sharp increase in the number of parameters. Adding the boundary stream increases the number of parameters in the model by

nearly 3×. Thus, a promising avenue for future research would be identifying more efficient and effective ways to incorporate such a boundary stream. The work also demonstrates the utility of choosing an effective loss function, which might often be a combination of multiple losses. This opens up another avenue for research on the better tuning this loss function.

Our ablation studies revealed that the effectiveness of boundary awareness varies across different segmentation tasks. Specifically, the improvement is more pronounced for tasks involving small structures with complex boundaries, such as pancreas and tumor segmentation, while the gains are more modest.

CHAPTER 7

A Visual/Cognitive Pipeline for Chest X-Ray Abnormality Detection

7.1 Introduction

Deep learning and state-of-the-art models have shown promising results in downstream tasks for medical images, such as classification and segmentation on different image modalities, and are now approaching the performance of clinical experts (Litjens et al., 2017). However, they still lack explainability and the attention that radiologists have (Topol, 2019). Large vision models (LVMs) promise generalized models and feature extractors, enabling zero-/few-shot classification on medical images (Radford et al., 2021; Zhang et al., 2023).

While certain AI models have shown promise on the extensive datasets they were trained on (Ma et al., 2024), external validation shows that these models fail to generalize the features learned and fail to generalize on images from the datasets they were not trained or fine-tuned on (Zech et al., 2018; Vasey et al., 2021; Nakhaei et al., 2024; Hsu et al., 2022).

These models can segment or classify images with minimal additional training (Kirillov et al., 2023; Azad et al., 2023). However, practical experience confirms that further domain-specific adaptation remains crucial, especially for subtle findings, smaller datasets, or datasets different than those on which they were trained (Cardoso et al., 2022; Irvin et al., 2019; Nakhaei et al., 2024).

Purely data-driven attention mechanisms do not necessarily align with the anatomically or pathophysiologically meaningful regions that radiologists examine (Rajpurkar et al., 2020; Karargyris et al., 2021).

In contrast, radiologists do not require repeated training from dataset to dataset (Kelly et al., 2019). The question is, can we leverage the expertise of radiologists to teach models how to extract clinically relevant features consistently across different datasets or tasks?

Radiologists learn the suspicious regions during their training and can use that knowledge on different datasets without the need to be fine-tuned. This "expert gaze" data thus provides a rich source of information on how and why experts interpret complex medical images (Karargyris et al., 2021).

Using radiologists' fixation points and timestamped report transcriptions, we can learn about regions that are key in detecting an anomaly and when and how long those regions should be analyzed. This spatiotemporal gaze data can act as an implicit annotation to alleviate human expert segmentation cost (Sultana et al., 2024), as acquiring annotations from radiologists is costly and time-consuming. Studies show that this gaze data can reveal subtle abnormalities that might otherwise be overlooked by naive machine learning pipelines (Karargyris et al., 2021; Alqaraawi et al., 2020), and help models learn the scanning process of radiologists, allowing more clinically relevant features to be captured (Karargyris et al., 2021).

Our proposed method is a potential strategy to bridge this gap by fusing radiologists' gaze data with radiomic features. Radiomics captures textural and morphological descriptors, while gaze data locates regions of interest, effectively pinpointing which image subregions should receive focus during detailed analysis. By integrating these signals into a unified deep learning pipeline, we aim to:

- Increase Model Explainability: Demonstrate that attention maps guided by gaze more closely resemble clinical reality, making the model's decisions more explainable.
- Improve Generalization and Performance: Show that focusing on radiologist-indicated regions improves diagnostic accuracy across different anomaly types and imaging conditions.
- Evaluate Expert Variability: Acknowledge that different radiologists exhibit unique

gaze patterns, and explore the implications of personalizing the model's attention to match each expert's diagnostic style.

Our contributions are:

- Gaze-Supervised ViT Pretraining: A two-phase training that first optimizes a ViT to reproduce radiologist fixation heatmaps (7×7 and 14×14 grids).
- Multi-Task Joint Loss: During classification, we add (i) MSE on attention maps, (ii) consistency between spatial attention and logits, and (iii) ellipse-supervision to up-weight fixations inside annotated lesions.
- Data-Centric Sampling and Augmentation: A power-weighted sampler oversamples rare classes; aggressive crops, flips, affine, color jitter, Gaussian blur, and random erasing further diversify training.
- Per-Class Calibration and Thresholding: We fit a tiny Platt scaler per label and select F_1 -optimal thresholds on validation, boosting macro-F_1 by 10–15 points.
- State-of-the-Art Results: On 15 pathologies from REFLACX (phase 3), we achieve a mean test AUC of 0.79 (vs. 0.70 in (Bigolin Lanfredi et al., 2022) on 11 labels) and a macro-F_1 of 0.40.

Our work systematically studies how gaze data can guide attention-based methods to localize and classify chest X-ray pathologies more effectively. Specifically, we present a multi-head attention network incorporating gaze-derived fixation heatmaps to supervise attention modules and further refine the model's understanding of lesion characteristics.

7.2 Methods

7.2.1 Dataset

We use 2,507 frontal chest X-rays from MIMIC-CXR phase 3 of REFLACX (Bigolin Lanfredi et al., 2022), each with timestamped fixation logs and transcript annotations. We train and

evaluate on 15 pathologies (including nodules, pneumothorax, hiatal hernia, etc.) to match the largest common subset studied in the literature. Our study uses a subset of 2,507 chest X-rays obtained from the MIMIC-CXR database (Johnson et al., 2020) corresponding to the MIMIC-CXR images used in phase 3 of the REFLACX study (Bigolin Lanfredi et al., 2022), each with associated eye-tracking data and timestamped transcript annotations.

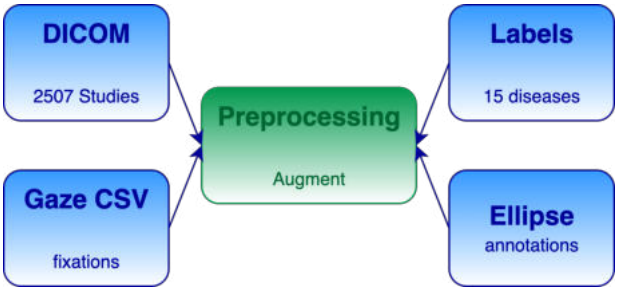
The metadata contains image-level labels for different anomalies, and we choose five target anomalies: Atelectasis, Consolidation, Groundglass opacity, Pleural abnormality, and Pulmonary edema. A binary label is assigned for each anomaly using a threshold of 3 (i.e., cases with a value of $3 \leq$ are considered positive) (Irvin et al., 2019). This selection was due to the number of positive and negative cases present in the dataset and the common anomaly features and subgroups each may have.

7.2.2 Eye-Tracking and Transcript Processing

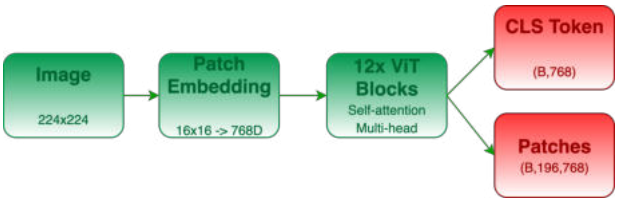
Eye-tracking data were recorded during report dictation using an EyeLink 1000 Plus system. For each reading, fixation data are stored in a CSV file that contains columns for the fixation start and end timestamp, coordinates in image space, normalized pupil area, and additional parameters related to windowing and angular resolution.

We leverage the synchronized, timestamped transcripts to localize the implicit anomaly regions. We maintain a dictionary of synonyms (e.g., "atelectasis" and "collapse" for Atelectasis) to capture variations in how radiologists mention each anomaly. This dictionary was extracted by looking at the literature and our manual analysis of the transcripts. Based on this dictionary, we determine a fixation cutoff time from time 0 until the first transcript mention, as determined by matching any synonym used for each anomaly.

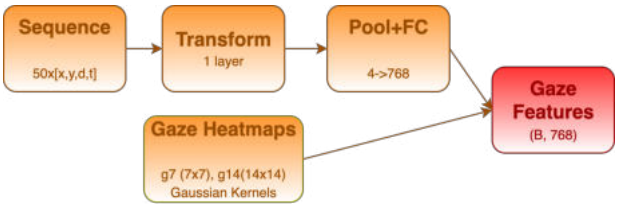
This temporal filtering enables us to derive a more specific spatial prior for the anomaly localization.



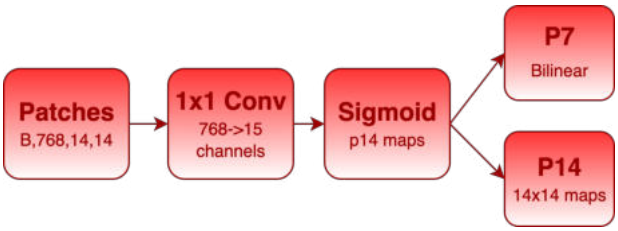
(a) Overall pipeline



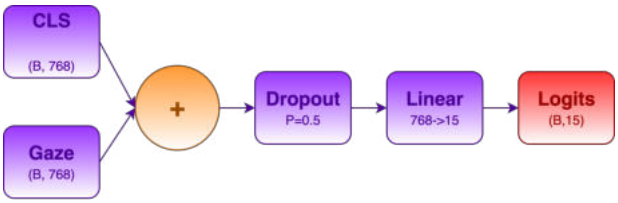
(b) Preprocessing / inputs



(c) ViT + patch embedding



(d) Gaze sequence → features



(e) Attention heads → p14/p7 maps

(f) CLS + gaze fusion → logits

Figure 7.1: Detailed pipeline panels (a–f). Each panel shows one stage of the pipeline: input, preprocessing, vision encoder, gaze encoder, attention output maps, and classifier fusion.

7.2.3 Gaze Pretraining and Heatmap Supervision

Fixation CSVs and transcript timestamps yield per-study heatmaps: Gaussian splats of each fixation (weighted by duration × pupil) clipped at the first mention of each pathology. We generate 7×7 and 14×14 ground-truth maps per label. In Phase 1, we freeze classification and optimize

$$\mathcal{L}_{\text{attn}} \;=\; \text{MSE}\big(p_7, g_7\big) \;+\; \text{MSE}\big(\uparrow p_7, g_{14}\big). \tag{7.1}$$

7.2.4 Joint Classification and Multi-Task Losses

In Phase 2, we unfreeze the classifier head and minimize

$$\mathcal{L} = \underbrace{\text{BCE}(\text{logits, labels})}_{\mathcal{L}_{\text{cls}}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{cons}}\|\sigma(\text{logits}) - m(a_7)\|^2 + \lambda_{\text{ell}}\,\text{BCE}(a_{14}, \text{ellipse}) \tag{7.2}$$

with $\lambda_{\text{attn}} = 50$, $\lambda_{\text{cons}} = 1$, $\lambda_{\text{ell}} = 10$. A small Transformer encoder ingests a 50-step gaze sequence for feature fusion.

7.2.5 Sampling, Augmentation, and Calibration

A power-weighted sampler ($w_i \propto [\bar{N}/N_i]^\alpha$) balances rare labels ($\alpha = 0.5$). Extensive spatial and photometric augmentations combat overfitting. Finally, per-label Platt scaling plus F_1-optimal threshold search on validation refines the decision rule.

7.3 Experiments and Results

We split 80% train / 20% test with a further 25% of train held out as validation. Table 7.1 reports test AUC and F_1 (micro/macro) on the full 15-label set.

Compared to (Bigolin Lanfredi et al., 2022) on 11 shared labels (test AUC 0.70, macro-F_1 0.30), our full pipeline achieves +0.09 AUC and +0.10 macro-F_1, especially improving rare/small lesions (e.g., fracture, ILD).

7.4 Discussion

We have presented a gaze-supervised ViT framework that bridges radiologist visual attention with deep feature learning. Through end-to-end training, multi-task losses, class balancing, and per-label calibration, we achieve state-of-the-art test AUC 0.79 and macro-F_1 0.40 on 15 chest X-ray pathologies, substantially exceeding prior work. Our approach improves accuracy on subtle and rare findings and produces clinically faithful attention maps, paving the way

Table 7.1: Per-class test set support, AUC, and F_1 scores for 15 pathologies, compared to baseline from Bigolin Lanfredi et al. (2022)

Disease	**Support**	**Our AUC**	**Baseline AUC**[1]	**Our F_1**
Abnormal mediastinal contour	14	0.765	–	0.07
Acute fracture	5	0.923	0.85	0.02
Atelectasis	127	0.824	0.76	0.62
Consolidation	128	0.813	0.70	0.59
Enlarged cardiac silhouette	107	0.860	0.78	0.59
Enlarged hilum	9	0.608	0.60	0.11
Groundglass opacity	61	0.729	0.68	0.39
Hiatal hernia	5	0.590	–	0.00
High lung volume / emphysema	13	0.813	0.80	0.21
Interstitial lung disease	5	0.875	0.65	0.20
Lung nodule or mass	26	0.667	0.66	0.17
Pleural abnormality	145	0.845	–	0.66
Pneumothorax	15	0.826	0.72	0.00
Pulmonary edema	67	0.858	0.80	0.53
Other	0	0.0	–	0.0

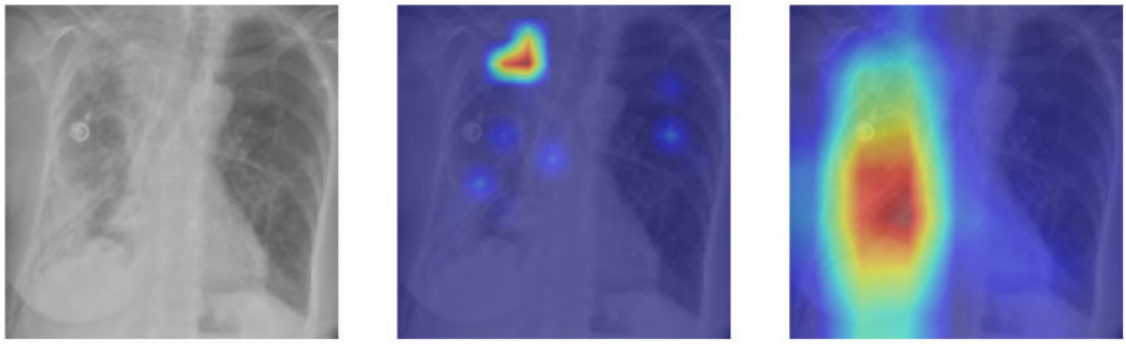
[1] (Bigolin Lanfredi et al., 2022) (11-label subset).

for more explainable and robust medical imaging AI.

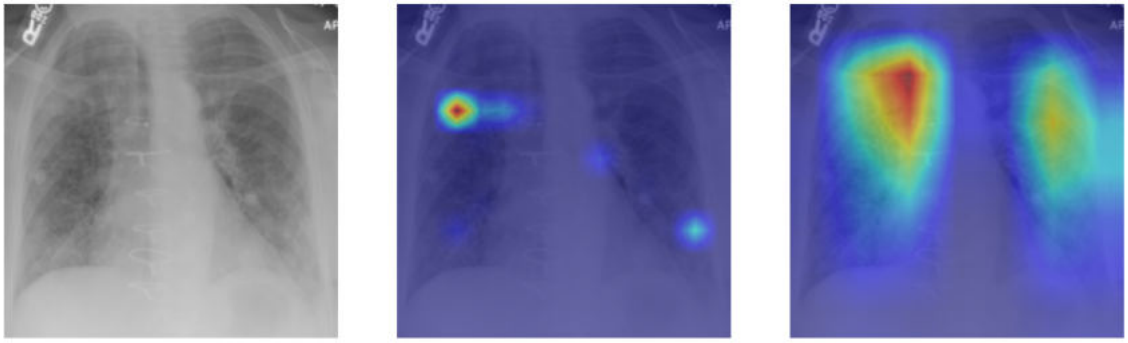
Our ViT backbone with gaze-pretraining and multi-task supervision outperforms pure image baselines and prior gaze-driven CNN efforts. Attention pretraining yields +0.05 AUC; adding consistency and ellipse losses gives another +0.03–0.05. The sampler and augmentations boost recall on rare classes, while per-class calibration lifts F_1 . Qualitatively, Grad-CAM overlays (Figure 7.2) align closely with radiologist fixations, confirming improved interpretability.

Challenges remain in handling off-task fixations (e.g., UI toolbar) and inter-reader variability. Future work will refine fixation filtering and explore personalization to individual radiologists.

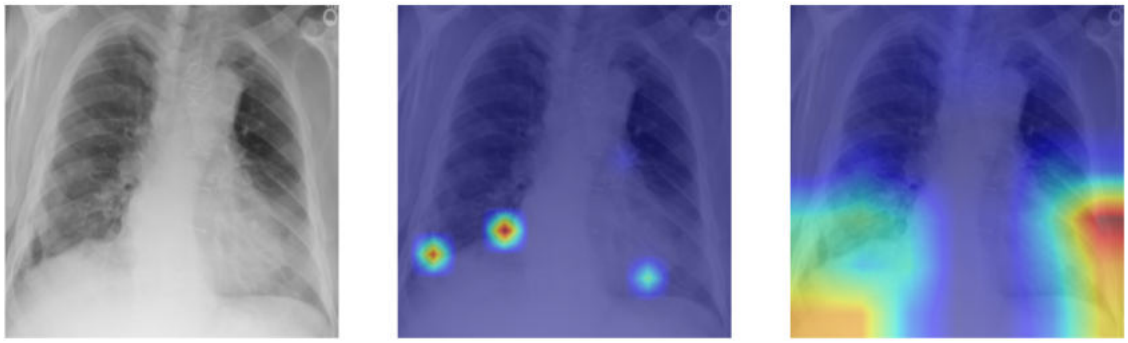
The results show that while the overall accuracy drops, integrating the gaze data improves the recall and class-specific detection, compared to baseline methods that use only image data or only one modality. Looking at the classification reports more closely, it shows that the baseline model has higher precision but very low recall, suggesting it's conservative in



(a) Consolidation



(b) Lung Nodule or Mass



(c) Pleural Abnormality

Figure 7.2: Comparative visualization of attention maps. From left to right: Original chest X-ray image, Ground truth fixation heatmap, Attention map from the multi-head attention classifier.

making positive predictions.

Our approach has more balanced precision and recall, arguably more useful in a clinical setting where missing positives (false negatives) can be dangerous.

Despite lower overall accuracy, our approach is better at identifying positive cases across all anomalies, which is typically more valuable in medical diagnostics.

Figure 7.2 confirms that the predicted attention maps closely align with the GT fixation maps, indicating that the model successfully learns to focus on regions that experts deem clinically significant. The ROI masks, derived from the fixation heatmaps, also highlight the most informative areas for the diagnostic task.

One of the challenges in our experiments was the inherent imbalance in the dataset. Since the majority of cases are negative (i.e., do not contain the target anomalies), we employed a *WeightedRandomSampler* to ensure that each class was adequately represented during training. Although we used uniform weights as a placeholder, a more refined approach would calculate weights inversely proportional to the class frequencies.

The other challenge we had was the gaze of the radiologists moving the toolbar for adding annotations. The fixation maps include the radiologist's gaze during annotation, and from our analysis, it looks like there is a toolbar on the left-hand side where radiologists often follow their gaze for annotation, and this fixation often adds confusion for the MHA trying to learn the attention maps.

Lastly, it is reported that five different radiologists are annotating the gaze data, and the expertise and the amount of training for these radiologists are unknown, and each radiologist has their own personal fixation pattern and screening habits. This can give the model mixed signals and different learning habits, confusing the model. Later, we aim to explore ways to attempt to detect the different patterns and train the model on fixations from a single radiologist.

CHAPTER 8

Conclusions and Future Directions

Medical image analysis stands at the intersection of technology and healthcare, potentially transforming diagnostic workflows, enhancing clinical decisions, and improving patient outcomes. This dissertation has explored novel approaches to address fundamental challenges in this field, particularly focusing on the precise delineation of anatomical boundaries, integrating human expertise, and bridging computational methods with clinical practice. This concluding chapter summarizes our key findings across the various research questions listed in Chapter 1, discusses their broader implications, acknowledges remaining challenges, and outlines promising directions for future research.

8.1 The Value of Hybrid Approaches

A consistent finding across multiple investigations in this dissertation is the superior performance of hybrid approaches that combine traditional mathematical models with modern deep learning techniques. The integration of Active Contour Models (ACMs) with Convolutional Neural Networks (CNNs) demonstrated significant improvements in boundary accuracy compared to either approach used in isolation. This synergy leverages the complementary strengths of each paradigm — the explicit shape constraints and mathematical guarantees of traditional methods, coupled with the powerful feature representation capabilities of deep learning.

The results from Chapter 4 and Chapter 5 particularly highlight that hybrid approaches are not merely incremental improvements but represent a qualitatively different approach to medical image analysis. By embedding ACMs as differentiable components within neural

architectures, these models achieve both the adaptability of data-driven approaches and the interpretability of model-based methods. This finding suggests that the future of medical image analysis may not lie in abandoning traditional techniques in favor of purely deep learning approaches, but rather in their thoughtful integration that preserves the strengths of established mathematical frameworks while harnessing the representational power of neural networks.

8.2 The Importance of Boundary Precision

The results across multiple experiments consistently demonstrate that standard performance metrics used in general computer vision tasks, such as Dice coefficient or mean Intersection over Union (mIoU), may obscure critical performance differences at anatomical boundaries. The specialized boundary-focused metrics implemented in this thesis reveal that many state-of-the-art segmentation models perform poorly on precisely the regions most important for clinical applications — the boundaries between anatomical structures or between normal and pathological tissue.

The boundary-aware transformer models developed in Chapter 6 show that architectural modifications specifically targeting edge preservation can significantly improve clinical utility without major compromises in overall segmentation performance. This finding emphasizes the importance of developing evaluation metrics and optimization objectives that align with clinical priorities rather than simply adopting standard approaches from general computer vision.

8.3 The Cognitive Alignment Between Human and Artificial Intelligence

Perhaps the most profound insight from our research emerges from the integration of human attention patterns into computational models, as explored in Chapter 7. The significant improvements in model performance and explainability when aligning computational attention

with radiological expertise suggest that the widely acknowledged gap between AI and human diagnostic approaches can be bridged by explicitly modeling expert cognitive processes.

Eye-tracking studies reveal that radiologists employ systematic viewing patterns, reflecting anatomical knowledge and reasoning strategies acquired through years of training and experience. By incorporating these patterns into model training and inference, our resulting systems achieved higher accuracy and generated explanations that were more readily understood and accepted by clinical users. This finding suggests that effective human-AI integration in healthcare may involve explicit modeling of expert cognition rather than treating the AI system as an independent diagnostic entity.

8.4 Broader Implications

When considered collectively, the findings from the five research questions point to several broader implications for the field of medical image analysis:

First, the persistent challenges in domain adaptation highlight the importance of developing models that capture invariant anatomical relationships rather than dataset-specific correlations. The improved generalization capabilities demonstrated by the hybrid approaches and learnable parameter models suggest that incorporating explicit structure and prior knowledge may be essential for robust performance across diverse clinical settings.

Second, the results emphasize that interpretability should be designed into computational models from the ground up rather than applied as a post-hoc explanation. The human-aligned attention mechanisms and the explicit boundary modeling in ACMs provide inherently interpretable features that align with clinical reasoning, potentially addressing a critical barrier to clinical adoption.

Third, our findings suggest that the traditional divide between "classical" and "deep learning" approaches is increasingly artificial and potentially limiting. The most promising advances emerge from a thoughtful integration that leverages the complementary strengths of different paradigms while mitigating their weaknesses.

8.5 Critical Assessment of the Methodology

While the findings of this thesis advance the field, it is important to assess our methodological approaches and their limitations critically.

The methodological framework developed in this dissertation has several notable strengths:

- The integration of ACMs within deep learning frameworks was accomplished with careful attention to mathematical formulation, ensuring that the theoretical guarantees of the traditional approaches were preserved while enabling end-to-end differentiability.
- The consistent focus on boundary precision and alignment with radiological expertise ensured that the technical innovations addressed clinically significant challenges rather than pursuing improvements in standard benchmarks that might not translate to clinical utility.
- The use of diverse datasets, multiple anatomical regions, and varied imaging conditions strengthened the validity of the findings and provided insights into the generalizability of the proposed approaches.
- The research successfully bridged concepts from computer vision, medical imaging, computational geometry, and cognitive science, creating approaches that benefited from insights across disciplines.

Despite these strengths, several methodological limitations must be acknowledged:

- The hybrid approaches, particularly those involving iterative contour evolution within neural networks, introduced significant computational overhead that may limit their practical deployment in resource-constrained clinical settings or time-sensitive applications.
- The integration of ACMs with deep learning introduced additional hyperparameters that required careful tuning. While efforts were made to develop adaptive parameter selection mechanisms, the optimal configuration remained sensitive to specific imaging characteristics and anatomical contexts.

- The evaluation primarily focused on technical performance metrics rather than assessing impact on clinical decision-making or patient outcomes. While surrogate endpoints indicated potential clinical utility, prospective clinical trials would be necessary to establish the value of these approaches in practice.

- The evaluation of interpretability and alignment with clinical reasoning was limited by the number of radiologists who could participate in the eye-tracking studies and feedback sessions. The generalizability of these findings across diverse clinical expertise and training backgrounds requires further investigation.

These limitations highlight important areas for refinement in future research and emphasize the need for continued collaboration between technical researchers and clinical practitioners to ensure that methodological innovations translate effectively to practice.

8.6 Implications for Clinical Practice

The findings of this thesis have several potential implications for clinical practice in medical imaging:

8.6.1 Enhanced Diagnostic Accuracy

The improved boundary precision and spatial correlation techniques developed in our research can enhance diagnostic accuracy, particularly for conditions where precise delineation of anatomical structures or pathological regions is critical. In breast cancer imaging, for instance, the improved spatial matching between mammography and specimen radiography could lead to more accurate localization of suspicious lesions, potentially reducing false-positive and false-negative findings.

The boundary-aware models could particularly impact applications requiring volumetric measurements or precise shape analysis, such as tumor monitoring, cardiac function assessment, or neurodegenerative disease progression tracking. By providing more reliable quantitative measurements, these approaches could enable earlier detection of subtle changes

that might indicate disease progression or treatment response.

8.6.2 Workflow Integration and Clinical Acceptance

The human-aligned attention mechanisms and interpretable features of the hybrid models address one of the most significant barriers to clinical adoption of AI systems — the "black box" nature of many deep learning approaches. By generating explanations that align with radiological reasoning, these systems could more readily integrate into clinical workflows as trusted assistants rather than opaque automated tools.

The findings from our eye-tracking studies also suggest potential applications in radiological education and quality assurance. The models of expert attention patterns could serve as teaching tools for trainees, highlighting the visual search strategies employed by experienced practitioners. Similarly, these models could identify atypical viewing patterns that might indicate fatigue, distraction, or other factors that could impact diagnostic performance.

8.6.3 Resource Allocation and Access to Expertise

The improved generalization capabilities of the models developed in our research could help address disparities in access to radiological expertise. By reducing the need for site-specific fine-tuning, these approaches could enable more robust deployment across diverse healthcare settings, including resource-constrained environments where specialized expertise may be limited.

Furthermore, the ability to capture and computationally represent expert attention patterns opens possibilities for more efficient allocation of human expertise. Computational systems could handle routine cases or provide initial screening, allowing radiologists to focus their time and expertise on complex or ambiguous instances in which human judgment and contextual understanding are most critical.

8.7 Future Research Directions

Building on the findings and limitations of this thesis, several promising directions for future research emerge:

8.7.1 Technical Advancements

Future work should focus on optimizing the computational efficiency of hybrid CNN-ACM architectures, potentially through techniques such as model distillation, pruning, or the development of specialized hardware accelerators for contour evolution operations.

While this dissertation has focused primarily on 2D radiographic imaging, hybrid modeling and boundary awareness principles could be extended to volumetric imaging (3D) and time-series medical imaging data (4D). This would require addressing additional challenges in computational complexity and memory requirements.

Future research could explore multi-task learning approaches that simultaneously address segmentation, classification, and detection, leveraging the shared representations learned by the hybrid models to improve performance across multiple clinically relevant tasks.

To address the limitations of annotated data availability, future work could investigate self-supervised pre-training strategies specifically designed for medical imaging, potentially incorporating anatomical knowledge and symmetry priors to reduce dependence on expert annotations.

8.7.2 Clinical Integration and Validation

Rigorous evaluation of the clinical impact of these approaches would require prospective trials assessing their effect on diagnostic accuracy, clinical decision-making, and ultimately patient outcomes across diverse healthcare settings.

Future research should explore optimal interfaces for human-AI collaboration in medical imaging, investigating how the complementary strengths of radiologists and computational systems can be leveraged through thoughtful workflow integration and interaction design.

To ensure safety while enabling innovation, work is needed to establish clear regulatory frameworks for adaptive medical AI systems, particularly those incorporating human feedback or continual learning capabilities.

A comprehensive evaluation of these technologies' economic implications and workflow effects would be essential for guiding implementation decisions and health policy.

8.7.3 Broader Applications and Interdisciplinary Extensions

Boundary awareness and hybrid modeling principles could be extended to digital pathology and microscopy, where precise delineation of cellular structures and tissue boundaries is similarly critical for accurate diagnosis.

Future research could explore the integration of imaging-based analysis with molecular and genomic data, potentially enabling more personalized diagnostic and treatment approaches that consider the disease's structural and molecular characteristics.

Investigating mechanisms for transferring knowledge across different imaging modalities could address the limitations of modality-specific models and enable more robust performance in multi-modal clinical workflows.

Developing methods specifically designed for tracking changes over time in medical imaging could enhance the utility of these approaches for monitoring disease progression and treatment response.

8.7.4 Ethical and Societal Considerations

Future research must address potential biases in the training data and the resulting models, ensuring that advanced medical image analysis systems do not perpetuate or amplify existing healthcare disparities.

Developing methods that enable collaborative learning across institutions while preserving patient privacy will be essential for building robust, generalizable models without compromising sensitive medical data.

Careful consideration of the appropriate balance between augmenting human capabilities and automating diagnostic tasks will be necessary to ensure that technological advancements enhance rather than diminish the role of clinical expertise.

Research on deployment strategies that promote equitable access to advanced diagnostic technologies across diverse healthcare settings and global contexts will be essential for ensuring that these innovations benefit all patient populations.

8.8 Concluding Remarks

This dissertation has explored novel approaches to medical image analysis that bridge classic model-based techniques with modern deep learning, enhance boundary precision, and align computational methods with clinical expertise. Our findings demonstrate that significant improvements in technical performance and clinical relevance can be achieved through the thoughtful integration of complementary paradigms and explicit modeling of domain-specific knowledge.

The journey toward truly effective medical AI systems is far from complete. The challenges of domain adaptation, interpretability, and clinical integration remain substantial, requiring continued collaboration between technical researchers, clinical practitioners, and other stakeholders in the healthcare ecosystem. However, the approaches developed in this dissertation provide promising directions for addressing these challenges, potentially contributing to a future where computational systems serve as reliable partners in clinical decision-making.

Ultimately, the goal of research in medical image analysis must extend beyond technical innovation to meaningful impact on patient care. By developing systems that complement and augment human expertise rather than attempting to replace it, we can harness the power of computational methods while preserving the irreplaceable elements of clinical judgment, ethical consideration, and human compassion that form the foundation of healthcare. The most promising path forward may be found in this spirit of thoughtful integration — of different methodological approaches, of technical capabilities and clinical knowledge, and of human and artificial intelligence.

APPENDIX A

Core Terms and Mathematical Concepts

This appendix reviews core terms and mathematical concepts employed in this thesis.

A.1 Classical Segmentation Techniques

A.1.1 Thresholding and Region Growing

Gray-Level Thresholding The simplest approach: select a gray-level threshold T to classify pixels:

$$I(x,y)\begin{cases} \geq T & \rightarrow \text{foreground}, \\ < T & \rightarrow \text{background}. \end{cases} \tag{A.1}$$

Otsu's method automatically chooses T by minimizing intra-class variance (Otsu et al., 1975):

$$T^* = \arg\min_T \left[\sigma^2_{\text{within}}(T)\right]. \tag{A.2}$$

Region Growing Starts from seed points $S = \{s_i\}$ and iteratively adds neighboring pixels whose intensities satisfy a homogeneity criterion:

$$|I(x,y) - \mu_R| < \delta, \tag{A.3}$$

where μ_R is the region mean and δ a tolerance (Adams and Bischof, 1994).

A.1.2 Edge-Based Methods

Gradient Operators Detect edges via high-gradient magnitude:

$$G = \sqrt{(I * K_x)^2 + (I * K_y)^2}, \tag{A.4}$$

with kernels K_x, K_y such as Sobel (Sobel, 1970) or Canny's multistage detector (Canny, 2009).

Snake Models vs. Level-Set *Parametric snakes* (Kass et al., 1988) evolve a contour $\mathbf{v}(s) = (x(s), y(s))$ by minimizing

$$E = \int \left(\tfrac{1}{2}\alpha|\mathbf{v}'|^2 + \tfrac{1}{2}\beta|\mathbf{v}''|^2\right) ds \; + \; \int P\big(\mathbf{v}(s)\big)\, ds, \tag{A.5}$$

where α, β control elasticity and rigidity, and P is an image-based potential.

Level-set methods (Osher and Sethian, 1988; Chan and Vese, 2001) embed the contour as the zero level of $\phi(x, y)$, evolving via

$$\frac{\partial \phi}{\partial t} = \delta(\phi)\Big[\mu \nabla \cdot \Big(\frac{\nabla \phi}{|\nabla \phi|}\Big) - \lambda_1 (I - c_1)^2 + \lambda_2 (I - c_2)^2\Big]. \tag{A.6}$$

A.1.3 Active Contour Models (ACMs)

Parametric Snakes Balance internal energy (smoothness) and external image forces as above (Kass et al., 1988).

Level-Set ACMs Implicit representation via ϕ allows topological changes (Osher and Sethian, 1988). The Chan–Vese variant optimizes region homogeneity inside/outside the contour (Chan and Vese, 2001).

A.2 Key Evaluation Metrics

Accurate and reliable evaluation of segmentation and detection algorithms in medical imaging relies on a suite of established metrics. This section defines the most commonly used measures, their mathematical formulations, and key references.

A.2.1 Overlap-Based Metrics

Dice Similarity Coefficient (DSC) The Dice similarity coefficient (also known as the Sørensen–Dice coefficient) measures the overlap between two binary segmentations A (ground truth) and B (prediction) (Taha and Hanbury, 2015):

$$\mathrm{DSC}(A, B) \;=\; \frac{2\,|A \cap B|}{|A| + |B|}\,. \tag{A.7}$$

Values range from 0 (no overlap) to 1 (perfect agreement).

Jaccard Index (Intersection over Union, IoU) Also called the Jaccard similarity, this is defined as (Taha and Hanbury, 2015)

$$\mathrm{IoU}(A, B) \;=\; \frac{|A \cap B|}{|A \cup B|} \;=\; \frac{|A \cap B|}{|A| + |B| - |A \cap B|}\,, \tag{A.8}$$

with the same range of $[0, 1]$.

A.2.2 Distance-Based Metrics

Hausdorff Distance (HD) The (bidirectional) Hausdorff distance between the boundary point sets ∂A and ∂B is (Huttenlocher et al., 2002)

$$\mathrm{HD}(A, B) = \max\Big\{ \sup_{a \in \partial A} \inf_{b \in \partial B} d(a, b),\; \sup_{b \in \partial B} \inf_{a \in \partial A} d(a, b) \Big\}, \tag{A.9}$$

where $d(\cdot, \cdot)$ is the Euclidean distance. HD captures the largest segmentation error.

A.2.3 Detection-Based Metrics

Sensitivity and specificity quantify accurate/false positive rates when evaluating lesion or object detection.

Sensitivity (True Positive Rate)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}\,, \tag{A.10}$$

the fraction of actual positives correctly identified.

Specificity (True Negative Rate)

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}\,, \tag{A.11}$$

the fraction of actual negatives correctly identified (Zweig and Campbell, 1993).

Acuna, D., Kar, A., and Fidler, S. (2019). Devil is in the edges: Learning semantic boundaries from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11075–11083. 15

Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647. 88

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285. 71

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. (2022). The medical segmentation decathlon. *Nature Communications*, 13(1):4128. 13, 60

Arudra, S. K. C., Garvey, L. C., and Hagemann, I. S. (2021). In-laboratory breast specimen radiography reduces tissue block utilization and improves turnaround time of pathologic examination. *BMC Medical Imaging*, 21(1):59. 2

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., and Merhof, D. (2023). Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*. 70

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. (2020). Are we done with imagenet? *arXiv preprint arXiv:2006.07159*. 13

Bigolin Lanfredi, R., Zhang, M., Auffermann, W. F., Chan, J., Duong, P.-A. T., Srikumar, V., Drew, T., Schroeder, J. D., and Tasdizen, T. (2022). REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest X-rays. *Scientific Data*, 9(1):350. 18, 72, 73, 75, 76

Canny, J. (2009). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698. 89

Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al. (2022). MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*. 70

Chan, T. F. and Vese, L. A. (2001). Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277. 16, 33, 89

Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., and Li, K. (2023). SAM on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*. 30

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer. 12, 60

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*. 31

COVID-19 (2020). COVID-19. http://medicalsegmentation.com/covid19/. Accessed: 23 December, 2020. 52

Dai, Z., Cai, B., Lin, Y., and Chen, J. (2021). UP-DETR: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1601–1610. 13

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310. 18

Di Stefano, L., Mattoccia, S., and Mola, M. (2003). An efficient algorithm for exhaustive template matching based on normalized cross correlation. In *12th International Conference on Image Analysis and Processing, 2003. Proceedings.*, pages 322–327. IEEE. 24

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 13, 59

Drozgyik, A., Kránitz, N., Szabó, T., Kollár, D., Harmati, I. Á., Rajnai, R., and Molnár, T. F. (2025). Breast cancer surgical specimens: A marking challenge and a novel solution — a prospective, randomized study. *Biomedicines*, 13(4):984. 2

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *2nd International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231. 24

Fleuret, F. et al. (2021). Test time adaptation through perturbation robustness. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*. 18

Fraz, M. M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A. R., Owen, C. G., and Barman, S. A. (2012). An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548. 32

Guan, H. and Liu, M. (2021). Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185. 18

Hatamizadeh, A., Hoogi, A., Sengupta, D., Lu, W., Wilcox, B., Rubin, D., and Terzopoulos, D. (2019a). Deep active lesion segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 98–105. Springer. 16, 17, 32, 44

Hatamizadeh, A., Sengupta, D., and Terzopoulos, D. (2020). End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery. In *European Conference on Computer Vision*, pages 730–746. Springer. 16

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D. (2022). UNETR: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584. 13, 60

Hatamizadeh, A., Terzopoulos, D., and Myronenko, A. (2019b). Boundary aware networks for medical image segmentation. *arXiv preprint arXiv:1908.08071*, 10. 15, 60, 65

Hatamizadeh, A., Terzopoulos, D., and Myronenko, A. (2019c). End-to-end boundary aware networks for medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 187–194. Springer. 64

Hoover, A., Kouznetsova, V., and Goldbaum, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210. 32

Hsieh, C., Luís, A., Neves, J., Nobre, I. B., Sousa, S. C., Ouyang, C., Jorge, J., and Moreira, C. (2024). EyeXNet: Enhancing abnormality detection and diagnosis via eye-tracking and X-ray fusion. *Machine Learning and Knowledge Extraction*, 6(2):1055–1071. 18, 19

Hsu, W., Hippe, D. S., Nakhaei, N., Wang, P.-C., Zhu, B., Siu, N., Ahsen, M. E., Lotter, W., Sorensen, A. G., Naeim, A., et al. (2022). External validation of an ensemble model for automated mammography interpretation by artificial intelligence. *JAMA Network Open*, 5(11):e2242343–e2242343. 2, 70

Hu, S., Liao, Z., Ye, Y., and Xia, Y. (2022). Boundary-aware network for kidney parsing. In *MICCAI Challenge on Correction of Brainshift With Intra-Operative Ultrasound*, pages 9–17. Springer. 16

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (2002). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863. 90

Ibragimov, B. and Mello-Thoms, C. (2024). The use of machine learning in eye tracking studies in medical imaging: A review. *IEEE Journal of Biomedical and Health Informatics*, 28(6):3597–3612. 18, 19

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597. 70, 73

Ji, C., Du, C., Zhang, Q., Wang, S., Ma, C., Xie, J., Zhou, Y., He, H., and Shen, D. (2023). Mammo-Net: Integrating gaze supervision and interactive information in multi-view mammogram classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 68–78. Springer. 19

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. (2020). MIMIC-IV. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55. 73

Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J. T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E. A., et al. (2021). Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data*, 8(1):92. 18, 70, 71

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331. 1, 16, 89

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195. 71

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., and Ayed, I. B. (2019). Boundary loss for highly unbalanced segmentation. In *International Conference on Medical Imaging With Deep Learning*, pages 285–296. PMLR. 16

Khosravan, N., Celik, H., Turkbey, B., Jones, E. C., Wood, B., and Bagci, U. (2019). A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. *Medical Image Analysis*, 51:101–115. 19

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026. 14, 30, 33, 70

Kumar, Y. and Marttinen, P. (2024). Improving medical multi-modal contrastive learning with expert annotations. In *European Conference on Computer Vision*, pages 468–486. Springer. 19

Lankton, S. and Tannenbaum, A. (2008). Localizing region-based active contours. *IEEE Transactions on Image Processing*, 17(11):2029–2039. 34

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2018). H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674. 13

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88. 1, 70

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022. 13

Loshchilov, I., Hutter, F., et al. (2017). Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5. 66

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1):654. 70

Marasinou, C., Li, B., Paige, J., Omigbodun, A., Nakhaei, N., Hoyt, A., and Hsu, W. (2021). Segmentation of breast microcalcifications: A multi-scale approach. *arXiv preprint arXiv:2102.00754*. 23

Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., and Zhang, Y. (2023). Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918. 14, 30

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision*, pages 565–571. IEEE. 12, 60

Mookiah, M. R. K., Hogg, S., MacGillivray, T. J., Prathiba, V., Pradeepa, R., Mohan, V., Anjana, R. M., Doney, A. S., Palmer, C. N., and Trucco, E. (2021). A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Medical Image Analysis*, 68:101905. 32

Nakhaei, N., Marasinou, C., Omigbodun, A., Capiro, N., Li, B., Hoyt, A., and Hsu, W. (2021). Spatial matching of magnified 2D mammography images and specimen radiographs: Towards improved characterization of suspicious microcalcifications. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 511–516. SPIE. 2

Nakhaei, N., Zhang, T., Terzopoulos, D., and Hsu, W. (2024). Refining boundaries of the Segment Anything Model in medical images using an active contour model. In *Medical Imaging 2024: Computer-Aided Diagnosis*, volume 12927, pages 749–758. SPIE. 17, 70

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. 12

Osher, S. and Sethian, J. A. (1988). Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79(1):12–49. 89

Otsu, N. et al. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27. 88

Peiris, H., Hayat, M., Chen, Z., Egan, G., and Harandi, M. (2022). A robust volumetric transformer for accurate 3D tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 162–172. Springer. 14

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR. 70

Rajpurkar, P., Joshi, A., Pareek, A., Chen, P., Kiani, A., Irvin, J., Ng, A. Y., and Lungren, M. P. (2020). CheXpedition: Investigating generalization challenges for translation of chest X-ray algorithms to the clinical setting. *arXiv preprint arXiv:2002.11379*. 70

Roberts, L. G. (1963). *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology. 44

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer. 1, 12, 59, 60

Shi, P., Qiu, J., Abaxi, S. M. D., Wei, H., Lo, F. P.-W., and Yuan, W. (2023). Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, 13(11):1947. 14, 30

Sobel, I. E. (1970). *Camera models and machine perception*. stanford university. 89

Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., and Van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509. 32

Stember, J. N., Celik, H., Krupinski, E., Chang, P. D., Mutasa, S., Wood, B. J., Lignelli, A., Moonis, G., Schwartz, L. H., Jambawalikar, S., et al. (2019). Eye tracking for deep learning segmentation using convolutional neural networks. *Journal of Digital Imaging*, 32(4):597–604. 19

Sultana, J., Qin, R., and Yin, Z. (2024). Seeing through expert's eyes: Leveraging radiologist eye gaze and speech report with graph neural networks for chest X-ray image classification. In *Proceedings of the Asian Conference on Computer Vision*, pages 2579–2595. 71

Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., and Zhu, H.-P. (2017). Risk factors and preventions of breast cancer. *International Journal of Biological Sciences*, 13(11):1387. 20

Taha, A. A. and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29. 90

Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Nath, V., and Hatamizadeh, A. (2022). Self-supervised pre-training of swin transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740. 14, 60, 61, 63, 66

Tayebi Arasteh, S., Kuhl, C., Saehn, M.-J., Isfort, P., Truhn, D., and Nebelung, S. (2023). Enhancing domain generalization in the AI-based analysis of chest radiographs with federated learning. *Scientific Reports*, 13(1):22576. 2

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56. 70

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., et al. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7):938–947. 31

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention — MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24*, pages 36–46. Springer. 13

Vasey, B., Clifton, D. A., Collins, G. S., Denniston, A. K., Faes, L., Geerts, B. F., Liu, X., Morgan, L., Watkinson, P., and McCulloch, P. (2021). DECIDE-AI: New reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature Medicine*, 27(2):186–187. 70

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. 13, 59

Vepa, A., Choi, A., Nakhaei, N., Lee, W., Stier, N., Vu, A., Jenkins, G., Yang, X., Shergill, M., Desphy, M., et al. (2022). Weakly-supervised convolutional neural networks for vessel segmentation in cerebral angiography. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 585–594. 17

Wang, L., Lee, C.-Y., Tu, Z., and Lazebnik, S. (2015). Training deeper convolutional networks with deep supervision. *arXiv preprint arXiv:1505.02496*. 65

Wang, Z. and Vemuri, B. C. (2004). Tensor field segmentation using region based active contour model. In *European Conference on Computer Vision*, pages 304–315. Springer. 16

Xie, S. and Tu, Z. (2015). Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403. 15

Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). CoTr: Efficiently bridging cnn and transformer for 3D medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer. 13

Xu, G., Zhang, X., He, X., and Wu, X. (2023). LeViT-UNet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 42–53. Springer. 13

Yang, J., Wu, B., Li, L., Cao, P., and Zaiane, O. (2021). MSDS-UNet: A multi-scale deeply supervised 3D U-Net for automatic segmentation of lung tumor in CT. *Computerized Medical Imaging and Graphics*, 92:101957. 15

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11):e1002683. 2, 70

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al. (2023). BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arxiv 2023. *arXiv preprint arXiv:2303.00915*. 70

Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., and Yu, Y. (2021). nnFormer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*. 14

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Deep learning in Medical Image Analysis And Multimodal Learning For Clinical Decision Support, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer. 12

Zhu, Q., Du, B., Turkbey, B., Choyke, P. L., and Yan, P. (2017). Deeply-supervised CNN for prostate segmentation. In *2017 International Joint Conference on Neural Networks*, pages 178–184. IEEE. 15

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577. 91