UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Breast Cancer Tumor Detection and Segmentation from Ultrasound and Mammography Images

by

Student: Xudong Song
Supervisors: Wenjing Jia, Gengfa Fang

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2026

Certificate of Original Authorship

I, **Xudong Song**, declare that this thesis is submitted in fulfilment of the requirements for the award of **Doctor of Philosophy**, in the **School of Electrical and Data Engineering**, **Faculty of Engineering and Information Technology**, at the **University of Technology Sydney**.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

Signature: Production Note: Signature removed prior to publication.

Date: April 15, 2026

Acknowledgments

I would like to express my deepest gratitude to my PhD supervisors, whose guidance and support have been instrumental throughout my doctoral journey.

First, I am profoundly grateful to Dr Gengfa Fang, who served as my primary supervisor during the initial phase of my PhD. His invaluable assistance during my doctoral application and scholarship pursuit laid the foundation for this academic journey. Under his leadership, our research team successfully completed numerous projects, through which I developed strong collaborative skills and practical research capabilities. His most profound impact on me has been through his inspiring character—his unwavering optimism, relentless determination, and passionate dedication to research. These qualities have not only motivated me throughout challenging times but have fundamentally shaped my approach to both research and life.

I extend my sincere appreciation to Professor Xiangjian He, who initially served as my co-supervisor and had previously mentored me as a visiting scholar in his laboratory before my PhD enrollment. Under his expert guidance, I published my first academic paper, marking a significant milestone in my research career. Professor He's profound academic insights and remarkable ability to distill complex problems to their essence have been invaluable to my intellectual development. The collaborative supervision of Professor He and Gengfa Fang resulted in my first publication after PhD enrollment, setting a strong trajectory for my subsequent research.

I am deeply indebted to Professor Wenjing Jia, who became my primary supervisor following Dr Gengfa Fang's career transition and guided me through to the completion of this doctoral journey. Her meticulous approach to research and genuine care for her students have created an environment where rigorous scholarship and personal growth coexist harmoniously. During the most challenging periods of my PhD, when obstacles seemed insurmountable, Professor Jia's patient guidance and

steadfast encouragement became the beacon that kept me moving forward. What truly distinguishes her mentorship is the perfect balance between academic rigor and human compassion—she has taught me not only how to be a better researcher but also how to persevere with grace and integrity. I am profoundly grateful for her dedication to seeing me through to the finish line.

I would like to acknowledge my exceptional colleagues and collaborators who have enriched my PhD experience. Special thanks go to Dr Xiaochen Fan and Qianwen Ye, both outstanding researchers who have provided invaluable guidance in academic writing and research methodology. Our collaborative efforts have resulted in multiple joint publications, and their insights have significantly enhanced the quality of my work. I am also grateful to my research team members—Oliver, Le Cai, Dr Zhanzhong Gu, and Xiaoyang Lu—with whom I have had the privilege of working on various projects, including the Age Care initiative and the ABUS breast tumor detection project. Their collaborative spirit, technical expertise, and camaraderie have made even the most challenging experimental work enjoyable and productive.

I gratefully acknowledge the financial support provided by the International Research Scholarship (IRS) and the Faculty of Engineering and Information Technology Scholarship, which have enabled me to pursue this research without financial constraints.

Finally, I express my heartfelt gratitude to my family, whose unconditional love, understanding, and unwavering support have been my anchor throughout this journey. Their belief in my abilities and their sacrifices have made this achievement possible.

Xudong Song

Sydney, Australia, 2026

List of Publications

Journal Papers

J-1. **X. Song**, X. Lu, G. Fang, X. He, X. Fan, L. Cai, W. Jia, Z. Wang, "ABUSDet: A Novel 2.5D Deep Learning Model for Automated Breast Ultrasound Tumor Detection," *Applied Intelligence*, vol. 53, no. 21, pp. 26255–26269, 2023.

J-2. **X. Song**, W. Jia, X. Fan, G. Fang, X. He, Z. Wang, "TriNet: Multi-Stage Refinement Networks for Mammographic Breast Tumor Segmentation with Uncertainty Modeling," *Computer Methods and Programs in Biomedicine*, under submission.

Contents

Certificate ii

Acknowledgments iii

List of Publications v

List of Figures x

List of Tables xiv

List of Abbreviations xvi

Abstract xviii

1 Introduction 1

1.1 Research Problems 5

1.2 Research Methodology 7

1.3 Main Contributions 9

1.4 Thesis Organization 11

2 Related Work 13

2.1 Overview 13

2.2 Imaging Modalities for Breast Tumor Detection and Segmentation . . 14

2.3 Evolution of Breast Tumor Detection and Segmentation Techniques . . 22

2.3.1 Traditional Machine Learning Era (2000–2012) 23

2.3.2 CNN Dominance and the Transformer Transition (2014–2021) 26

2.3.3 Emerging Paradigms (2020–Present) 31

2.4 Clinical Requirements for ABUS Detection and Mammography Segmentation 34

2.5 Research Gaps and Motivation 35

3 ABUSDet: A Novel 2.5D Deep Learning Model for Automated Breast Ultrasound Tumor Detection 38

3.1 Introduction 38

3.2 Related Work 40

3.3 Methodology 43

3.3.1 ABUSDet Architecture 43

3.3.2 Pre-processing 45

3.3.3 2.5D Tumor Detection Model Design 47

3.3.4 Aggregation Module and Post-processing Strategy 48

3.3.5 Transfer Learning 50

3.4 Deliberate Training 51

3.5 Experiments and Results 54

3.5.1 Implementation Details 55

3.5.2 Results and Analysis 56

3.6 Summary 67

4 TriNet: Multi-Stage Refinement Networks for Mammographic Breast Tumor Segmentation with Uncertainty Modeling 69

4.1 Introduction 69

4.2 Related Work 74

4.2.1 Medical Image Segmentation Architectures 74

4.2.2 Multi-Stage Segmentation Frameworks 76
4.2.3 Attention Mechanisms in Medical Image Segmentation 78
4.2.4 Integrating Prior Knowledge and Uncertainty Modeling 81
4.2.5 Boundary-Aware and Edge-Enhanced Segmentation 82
4.2.6 Summary and Motivation for TriNet 84
4.3 Methodology 85
4.3.1 Overall Framework 86
4.3.2 Decoder Module 88
4.3.3 LRCA: Lightweight Row-Column Attention 89
4.3.4 Incorporating Prior Knowledge with Gray–Scott Noise 92
4.3.5 Loss Functions 93
4.4 Experiments 98
4.4.1 Datasets and Pre-processing 98
4.4.2 Evaluation Metrics 102
4.4.3 Experimental Environment and Training Setup 105
4.4.4 Experimental Results and Discussion 107
4.4.5 Ablation Study 112
4.5 Conclusion 120

5 Conclusion and Future Work 123

5.1 Conclusion 123
5.1.1 Addressing the Research Problems 123
5.1.2 Principal Contributions and Outcomes 124
5.1.3 Clinical Significance and Deployment Considerations 127
5.2 Future Work 128

5.2.1 Advancing Architectural Frameworks for Enhanced Performance 128

5.2.2 Leveraging Unlabeled Data through Advanced Learning Paradigms 129

5.2.3 Clinical Integration and Real-World Deployment 129

5.2.4 Synthesis 130

Bibliography **131**

List of Figures

1.1 Global age-standardized incidence and mortality rates of breast cancer based on GLOBOCAN 2022 estimates [1]. 2

1.2 Global new cases and deaths by cancer type based on GLOBOCAN 2022 estimates [1]. 3

1.2 Global new cases and deaths by cancer type based on GLOBOCAN 2022 estimates [1] (continued). 4

2.1 Mammography examples from three public datasets (left to right: INbreast [2], CBIS-DDSM [3], DMID [4]). Yellow contours mark the tumor regions. 15

2.2 Digital breast tomosynthesis (DBT) examples from the BCS-DBT dataset [5]. Six reconstructed slices from a single volume are shown, sampled every 10 slices in the through-plane direction to highlight appearance changes with depth. Yellow boxes indicate the region of interest containing the tumor. 16

2.3 HHUS examples from the BrEaST dataset [6]. Left: rectangular FOV from a linear-array probe; right: sector FOV from a curvilinear/phased-array probe. Yellow contours mark the tumor regions. 18

2.4 Automated breast ultrasound (ABUS) example (de-identified). The same tumor is shown in the three standard planes from one volume: coronal (top-left), sagittal (right), and axial (bottom). Yellow boxes highlight the suspicious region of interest. 18

2.5 Breast MRI examples from the Duke-Breast-Cancer-MRI dataset [7]. Three orthogonal views from one exam are shown: *coronal* (top-left), *axial* (bottom), and *sagittal* (right). Yellow boxes highlight the suspicious region of interest. 20

3.1 Samples of tumors with different morphological patterns (Coronal View). 39

3.2 A range of slices of data blocks showing one tumor. 40

3.3 The system architecture of the proposed ABUSDet. 42

3.4 Illustration of the key modules in the proposed ABUSDet breast tumor detection model. 45

3.5 The schematic diagram of the proposed deliberate training scheme for training the deep learning model. 52

3.6 Histogram of tumor size distribution in the ABUS dataset. 55

3.7 Using data augmentation and emptyGT data 56

3.8 Comparison of different deliberate losses 57

3.9 Comparison of different channel numbers of the Coronal View. 58

3.10 Comparison of different Channel numbers of the Aerial View and Sagittal View. 59

3.11 Sensitivity obtained with different strides. 60

3.12 Comparison of different aggregation parameters “eps” (a) and different CS (Channel Squeeze) methods in (b). 61

3.13 Selected examples of tumor detection results on 2.5D data block of the Test set using the proposed ABUSDet model. The yellow boxes are the Ground Truth tumors, and the red boxes are the predicted boxes of tumor candidates. 62

3.14 Compare different Faster R-CNN models. 66

3.15 Comparison among 2D, 3D, and 2.5D methods 66

4.1 Examples of mammographic images showing low tumor-tissue contrast, ambiguous boundaries, and substantial size variability from the INbreast dataset [2]. Left: a small, low-contrast tumor with indistinct margins. Right: two tumors of different sizes within the same mammogram. Blue contours indicate the ground-truth tumor boundaries. 70

4.2 Overall architecture of TriNet. The three stages progressively refine the mask: S1 (feature extraction), S2 (initial decoding), and S3 (prior-guided refinement). The 3D cuboids represent the output feature maps of the encoder and decoder; orange boxes denote decoder modules, and light-blue boxes denote ConvBlocks. 86

4.3 Core building blocks: (a) **Decoder** with optional prior injection in refinement stage; (b) **ConvBlock** (pointwise–depthwise–pointwise with ECA); (c) **LRCA** placed after the ConvBlock. 88

4.4 **LRCA** pipeline. Patchify & Tokenize produces a square token grid; row-wise and column-wise self-attentions operate on axial sequence views; results are fused and restored by Untokenize & Unpatchify. The implementation computes the two axial attentions in a single SDPA call by concatenating the row/column batches. 89

4.5 Effect of Gray–Scott noise on the prior mask. (a) Original prior mask $S_{n-1}L_0$. (b) Masks after Gray–Scott perturbation. The perturbation adds plausible boundary uncertainty while preserving overall tumor structure, which improves robustness during stage-n refinement (*e.g.*, S3). 93

4.6 Qualitative illustration of the proposed ABALoss on two mammograms. The top row shows a case with a large tumor, and the bottom row shows a case with a small tumor. Columns show the ground-truth mask, predicted mask, their automatically extracted edge maps, and the edge overlay maps. Note how the edge thickness adapts to the tumor scale, providing balanced supervision across lesion sizes. 96

4.7 Visual comparison for the ablation study. Rows correspond to different combinations of components evaluated in Table 4.6. The acronym **DAPGL** (inserted above the symbols) indicates the presence (✓) or absence (✗) of **D**eep supervision, **A**BALoss, **P**rior Layer, **G**ray-Scott, and **L**RCA, sequentially. The full TriNet configuration (bottom row) demonstrates the most coherent boundary delineation and fewest false predictions. 113

4.8 Comparison of breast tumor segmentation results on the INbreast dataset [2]. Rows correspond to different models (including Original and Ground Truth), and columns correspond to different images. **TriNet** demonstrates strong boundary delineation. 117

List of Tables

3.1 The hyperparameters of the proposed architecture. 43

3.2 Execution time per volume of different strides in the ABUSDet model. 58

3.3 Stratified performance analysis of ABUSDet with respect to different tumor sizes. 63

3.4 Sensitivity and FPs/volumes of different models. 65

4.1 Comprehensive training and model configuration parameters used across all experiments. 102

4.2 Quantitative comparisons of segmentation performance of different models on the INbreast dataset [2] implemented in this research. Results are reported as mean ± std over five folds. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; ACC: Accuracy; SEN: Sensitivity; SPEC: Specificity; Params: Parameter count; FLOPs: Floating Point Operations; Time: Inference time per image; Best results are shown in boldface. 108

4.3 Quantitative comparisons of segmentation performance of different models on the CBIS-DDSM dataset [3]. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; ACC: Accuracy; SEN: Sensitivity; SPEC: Specificity; Params: Parameter count; FLOPs: Floating Point Operations; Time: Inference time per image; Best results are shown in boldface. 108

4.4 Quantitative comparisons of segmentation performance of different models on the DMID dataset [4]. Results are reported as mean ± std over five folds. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; ACC: Accuracy; SEN: Sensitivity; SPEC: Specificity; Params: Parameter count; FLOPs: Floating Point Operations; Time: Inference time per image; Best results are shown in boldface. 109

4.5 Quantitative comparisons of SOTA results as reported in literature on the INbreast dataset [2]. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; Best results are shown in boldface. 110

4.6 Ablation study on the INbreast dataset [2]. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; ACC: Accuracy; SEN: Sensitivity; SPEC: Specificity. 112

List of Abbreviations

ABUS: Automated Breast Ultrasound

ACC: Accuracy

AI: Artificial Intelligence

AUC: Area Under the Curve

BCE: Binary Cross-Entropy

CBIS-DDSM: Curated Breast Imaging Subset of DDSM

CNN: Convolutional Neural Network

CS: Channel Squeeze

DBT: Digital Breast Tomosynthesis

DDPM: Denoising Diffusion Probabilistic Models

DICOM: Digital Imaging and Communications in Medicine

DMID: Digital Mammography Dataset for Breast Cancer Diagnosis Research

DSC: Dice Similarity Coefficient

ECA: Efficient Channel Attention

FFDM: Full-Field Digital Mammography

FN: False Negative

FP: False Positive

FROC: Free-Response Receiver Operating Characteristic

GBCI: Global Breast Cancer Initiative

HHUS: Handheld Ultrasound

IOU: Intersection over Union

LBP: Local Binary Patterns

LRCA: Lightweight Row-Column Attention

MRI: Magnetic Resonance Imaging

RFE: Recursive Feature Elimination

SAM: Segment Anything Model

SDPA: Scaled Dot-Product Attention

SE: Squeeze-and-Excitation

SEN: Sensitivity

SPEC: Specificity

SRFM: Selective Receptive Field Module

SSM: State Space Model

TP: True Positive

WHO: World Health Organization

ABSTRACT

Breast Cancer Tumor Detection and Segmentation from Ultrasound and Mammography Images

by

Student: Xudong Song

Supervisors: Wenjing Jia, Gengfa Fang

Breast cancer is among the most commonly diagnosed cancers in women worldwide. While early screening reduces deaths, clinical practice requires both accurate detection for identifying suspicious regions and precise segmentation for treatment planning. Current research disproportionately focuses on 2D mammography due to its widespread availability and superior sensitivity to microcalcifications, despite 3D Automated Breast Ultrasound Systems (ABUS) offering superior performance in dense breast tissue without ionizing radiation. These complementary modalities each have limitations: 3D ABUS detection methods struggle with computational complexity, while 2D processing ignores inter-slice continuity, causing high false positives; mammographic segmentation struggles with extreme scale variation and generates over-smoothed boundaries, particularly problematic for spiculated margins.

This thesis presents two specialized models addressing these challenges. **ABUSDet** introduces a 2.5D detection framework that balances volumetric context with computational efficiency. It partitions ABUS volumes into overlapping multi-slice blocks, stacking 28 adjacent slices and compressing them via learnable channel reduction, enabling efficient detection while preserving 3D spatial relationships. A deliberate training strategy prioritizes under-trained batches by temporarily skipping well-trained ones to improve hard-example learning. ABUSDet achieves **77.94%** sensitivity at 3.0 false positives per volume, outperforming existing methods, with

11.0 seconds inference time.

TriNet achieves precise segmentation through progressive multi-stage refinement. It employs a single MobileViT encoder that feeds sequential decoder stages, which iteratively refine tumor boundaries. Each decoder integrates a newly proposed Lightweight Row–Column Attention (LRCA) module, reducing computational complexity while capturing global spatial dependencies. Gray–Scott reaction–diffusion perturbations are applied to the previous stage's prediction to create uncertainty-aware spatial priors for subsequent refinement. Stage-specific supervision: S2 prioritizing recall via optimizing DiceBCE and MissPenalty losses, S3 emphasizing boundary fidelity via optimizing Adaptive Boundary-Alignment Loss, enables coarse-to-fine refinement. Evaluated across three benchmark datasets under a whole-image segmentation setting, TriNet achieves state-of-the-art performance on INbreast with **88.40%** Dice and **88.20%** sensitivity using only **2.83M** parameters, while maintaining robustness on CBIS-DDSM (**55.94%** Dice) and DMID (**75.68%** Dice).

Together, ABUSDet and TriNet form a clinically aligned pipeline—ABUSDet for volumetric detection and TriNet for precise boundary delineation—advancing breast cancer screening while respecting computational constraints.

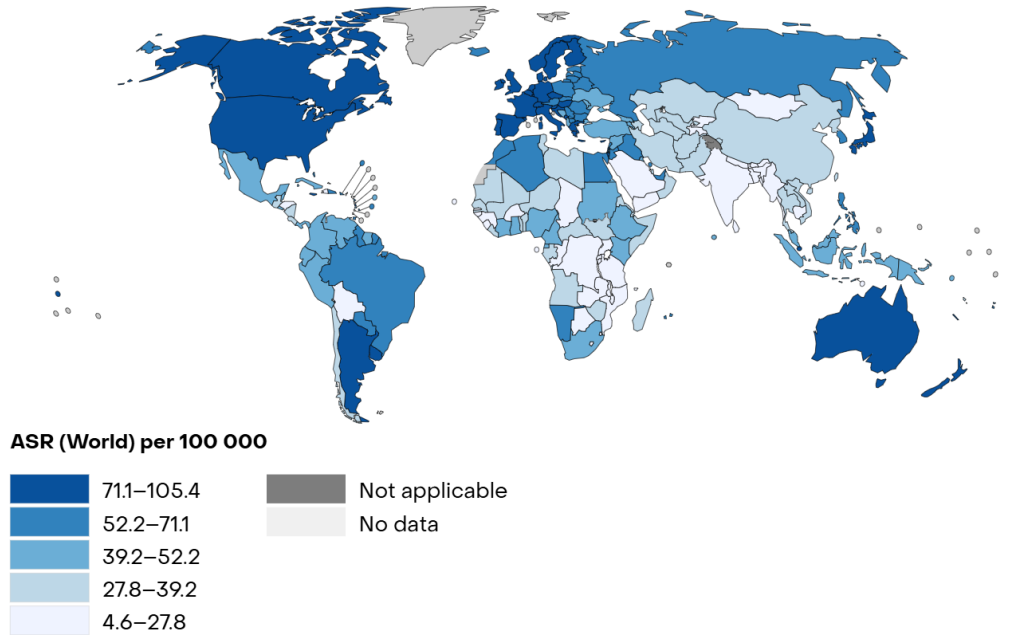
Keywords: Tumor Detection, Tumor Segmentation, ABUSDet, TriNet, Breast Cancer, ABUS, Mammography.

Chapter 1

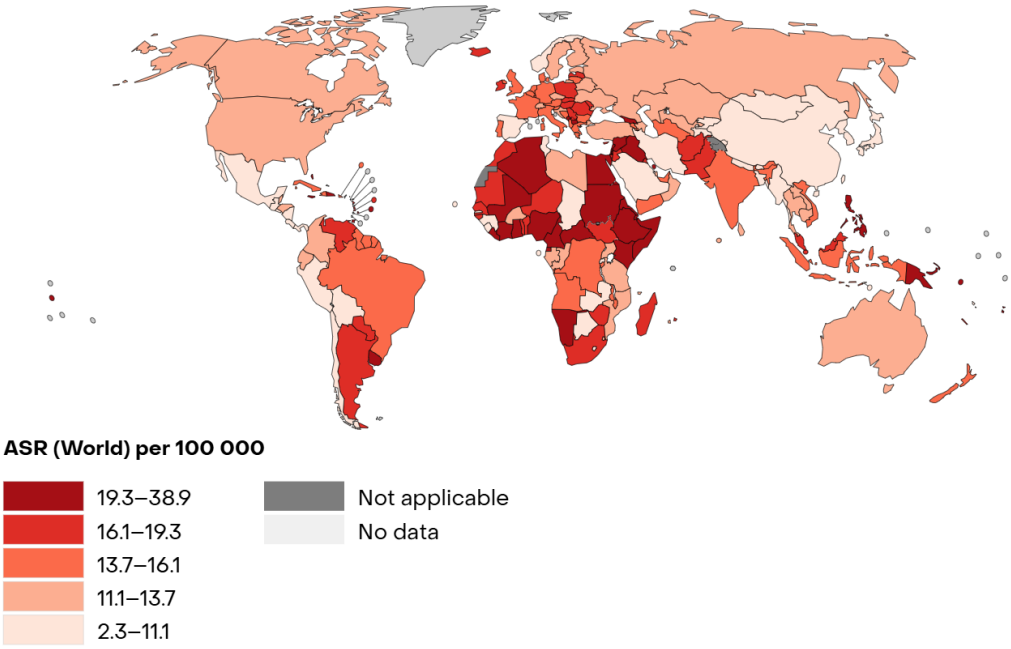
Introduction

Breast cancer is the most common cancer in women worldwide and a leading cause of cancer-related death. The latest GLOBOCAN 2022 estimates [1] report approximately 2,300,000 new cases and about 670,000 deaths annually. These patterns are not uniform: Fig. 1.1a shows the highest age-standardized incidence across high-income regions—Europe, North America, and Australia/New Zealand—whereas Fig. 1.1b shows disproportionately high age-standardized mortality across many lower-income settings, especially sub-Saharan Africa, indicating survival and access gaps. Figs. 1.2a and 1.2b corroborate the scale of the burden, with breast cancer ranking second for new cases and fourth for total cancer deaths worldwide. Because the overwhelming majority of cases occur in women, breast cancer is the leading cause of cancer death among women. Participation in organized screening programmes is associated with a 20–40% reduction in breast-cancer mortality, underscoring the importance of diagnostic technologies that are both accurate and broadly deployable [8].

In clinical workflows, population control of breast cancer naturally separates into two stages with distinct algorithmic targets. The screening stage aims not to miss tumors in large cohorts, where performance is measured as tumor-level sensitivity at constrained false positives per examination. The preoperative decision-making stage aims to quantify boundaries and size for staging, margin planning, and risk stratification. These goals are complementary: first, make a suspicious focus reliably visible and localizable, then draw its boundary precisely.



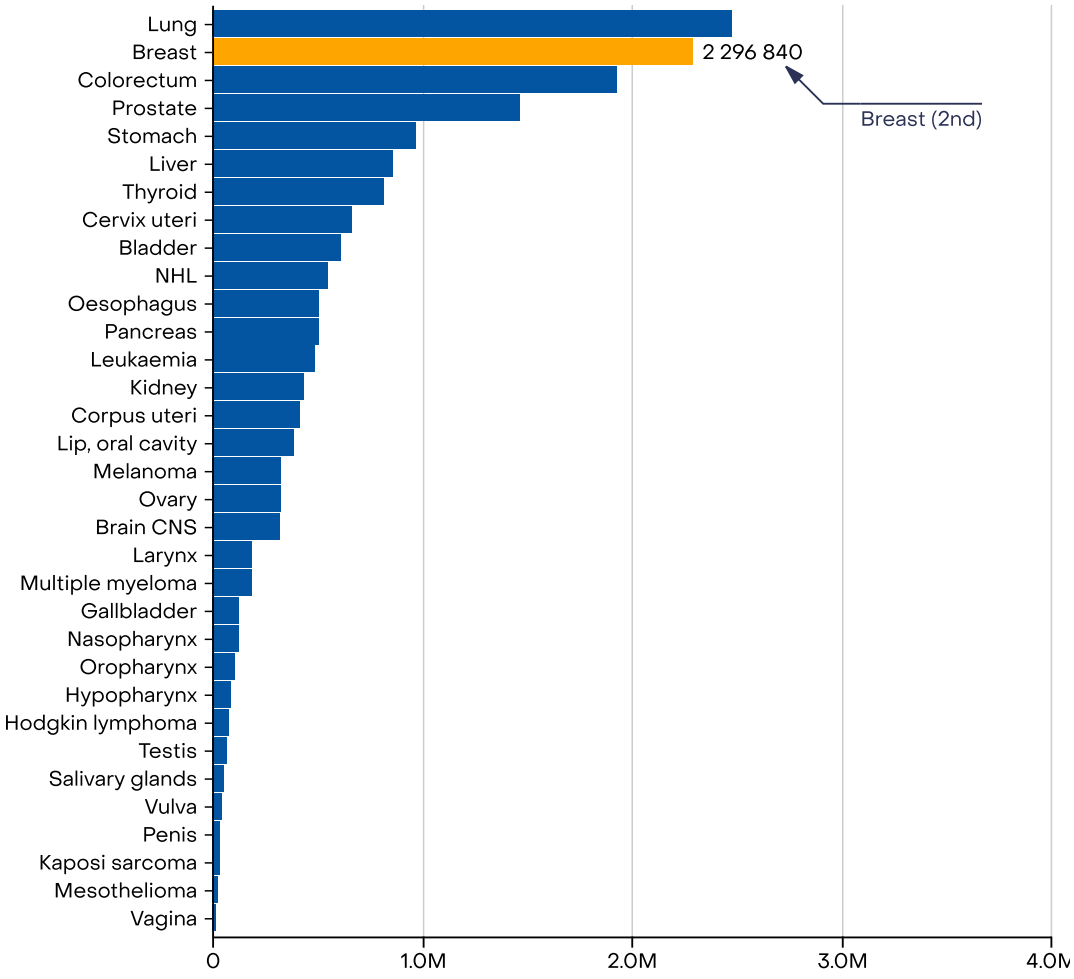
(a) Global age-standardized incidence rates



(b) Global age-standardized mortality rates

Figure 1.1 : Global age-standardized incidence and mortality rates of breast cancer based on GLOBOCAN 2022 estimates [1].

This division of labor is dictated by imaging physics and clinical realities. In women with dense breasts—who comprise a large fraction of screening populations—tissue superimposition can mask non-calcified masses on projection X-ray, and the

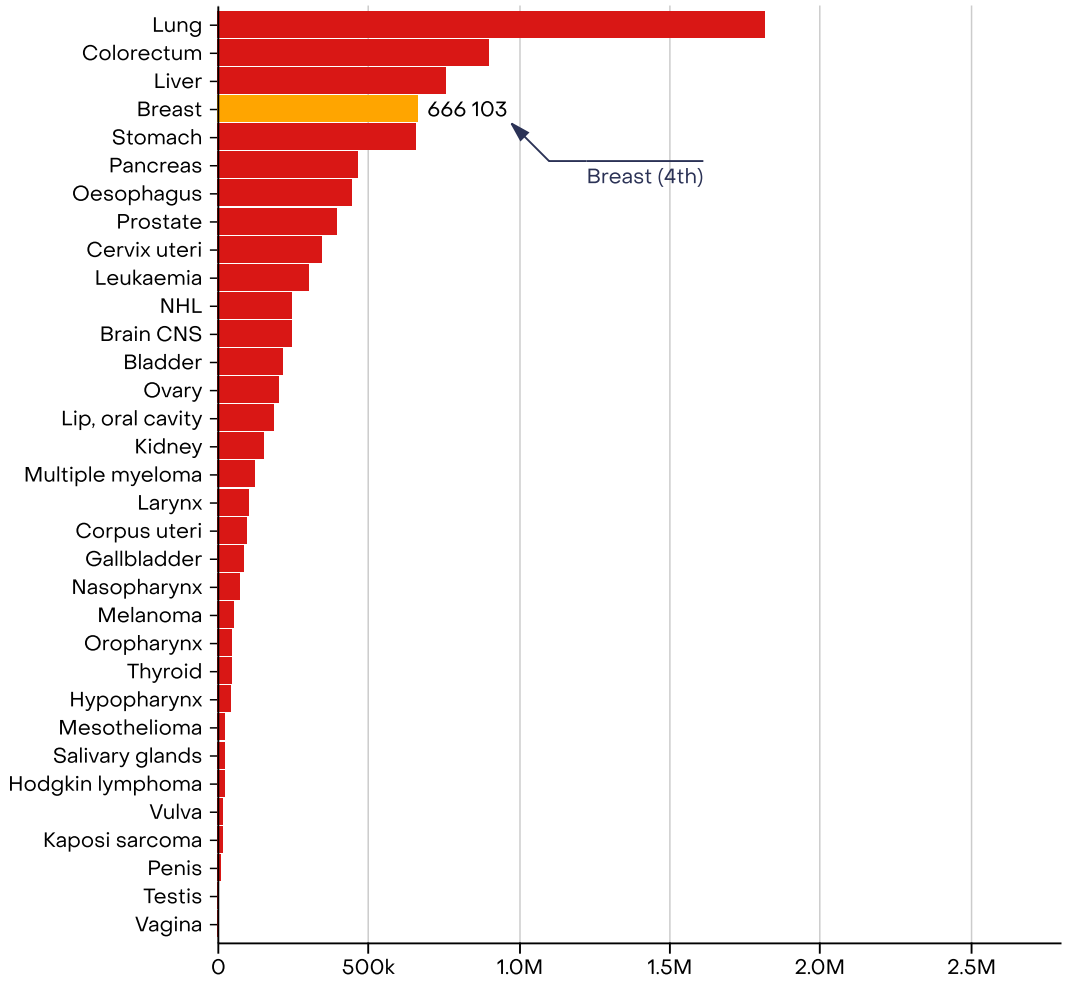


(a) Number of new cases by cancer type

Figure 1.2 : Global new cases and deaths by cancer type based on GLOBOCAN 2022 estimates [1].

sensitivity of mammography is known to drop compared with non-dense breasts [9]. As a standardized, radiation-free three-dimensional acquisition, automated breast ultrasound (ABUS) mitigates this superimposition and is used as a supplement to mammography in dense-breast screening to improve cancer detection while maintaining feasible workflows [10]. ABUS is therefore an appropriate substrate for the detection task in dense-breast screening: to recover tumors that would otherwise remain occult on X-ray.

Conversely, once a mass is visible on mammography (or tomosynthesis), the de-



(b) Number of deaths by cancer type

Figure 1.2 : Global new cases and deaths by cancer type based on GLOBOCAN 2022 estimates [1] (continued).

tector's sub-millimeter sampling (typically 50–100 μm) and standardized CC/MLO views provide a favorable canvas for quantifying morphology and margins—especially spiculation—which carry diagnostic value and support preoperative decisions. Pixel-level segmentation on mammography thus supplies reproducible size and boundary descriptors that are difficult to obtain with comparable precision from ultrasound volumes because of speckle, posterior acoustic shadowing, elevational-resolution anisotropy, and motion-related artefacts [11].

Guided by this clinically aligned split, this thesis focuses on mass tumors and

contributes two purpose-built models that address the hard constraints at each stage. For ABUS screening volumes, we introduce ABUSDet, a computationally efficient 2.5D detection framework that preserves through-plane context without incurring full-3D memory and latency. For mammography, we introduce TriNet, a progressive segmentation network that emphasizes boundary fidelity under tight compute budgets. Together, they form a complementary and clinically feasible pipeline: ABUS to detect, mammography to delineate.

1.1 Research Problems

Traditional machine learning delivered interpretable candidate generation and classification with hand-crafted descriptors but suffered from high false positives and poor cross-scanner transfer when appearance or protocols changed [12, 13]. Convolutional networks enabled end-to-end representation learning and became the backbone for detection and segmentation, yet their local receptive fields limit the modeling of global spatial dependencies [14]. Transformer-based designs introduced explicit long-range interactions and improved delineation on medical benchmarks, but significant optimization is still needed for model efficiency and inference speed [15, 16, 17]. Diffusion models have shown promise for generating high-quality medical images and handling ambiguous boundaries through probabilistic modeling, yet their iterative sampling process remains computationally prohibitive for real-time clinical deployment [18]. Against this backdrop, this thesis concentrates on two practical problems aligned with clinical workflow.

For ABUS volumetric detection, current methods fall into two paradigms with distinct trade-offs. The first processes volumes as adjacent slices using 2D detectors per frame, then fuses positions into 3D boxes [19]. While computationally efficient and compatible with transfer learning, this approach discards depth cues at decision time, causing tumor fragmentation across slices and inflating false positives

when similar tissue patterns repeat—particularly problematic for obliquely oriented masses and fibroglandular interfaces. The second paradigm applies sliding 3D windows to classify volumetric VOIs, preserving full context through 3D CNNs [20, 21]. However, this incurs prohibitive computational costs (67 seconds per volume), struggles with the wide tumor size distribution (2.9–73.8mm), and faces a fundamental trade-off: decreasing stride improves sensitivity but exponentially increases runtime. The problem targeted here is to preserve essential through-plane information under screening-grade memory and latency constraints, avoiding both full 3D computational burden and slice-wise fragmentation. This motivates the 2.5D approach that processes multiple adjacent slices simultaneously through channel compression, maintaining inter-slice continuity within the computational envelope of 2D frameworks.

For mammography segmentation, U-Net remains a standard architecture for pixel-level delineation, with skip connections preserving high-resolution detail while the decoder recovers object extent [14]. Building on this baseline, research has explored several strategies, each with practical limitations. One strand is multi-stage refinement: cascaded or connected designs (Connected-UNets/SegNets, DoubleU-Net) improve masks through coarse-to-fine iteration [22, 23]. Their main limitations are redundancy and error accumulation: duplicating a full encoder per stage inflates memory and latency, and treating the first-stage prediction as input to a second full network can amplify early mistakes downstream. A related line uses a presegmenter to produce a saliency prior that guides a subsequent refinement network [24]; while this reduces misses, it usually treats the prior as deterministic, so mislocalization in the prior can bias refinement without any mechanism to represent prior uncertainty. Deep supervision offers complementary benefits: UNet++ [25] and U-Net 3+ [26] attach multi-scale heads that stabilize optimization, speed convergence, and help recover small structures. However, mechanically adding more supervised heads can

introduce gradient competition and dilute learning signals, and applying the same loss to every output is suboptimal: early stages benefit from recall-oriented terms, whereas later stages require boundary-sensitive objectives. Viewed together, multi-stage refinement, prior guidance, and deep supervision can be unified in a single, efficient design that shares one encoder, employs sequential lightweight decoders, injects priors with explicit uncertainty modeling, and adopts stage-specific losses—this is the path we take in TriNet. Transformer-based global attention provides another advance: TransUNet [27], Swin-Unet [16], TransFuse [28], and TrEnD [29] show that long-range dependencies improve structural consistency; the practical trade-off is that tokenization and multi-branch fusion raise memory use and wall-clock time at mammography resolutions. Finally, boundary-guided methods—BAT and improved U-Net 3+, with recent density/size-aware reweighting (D-ASP)—sharpen contours [30, 31, 32, 33, 34, 35]; yet fixed-scale edges and hard targets can over- or under-penalize tumors at size extremes and destabilize gradients near the decision threshold, and most approaches ignore the uncertainty of intermediate boundaries. These observations motivate a decoder-centric and efficient TriNet: a shared MobileViT encoder, sequential decoders with lightweight row-column attention for global cues, uncertainty-aware prior injection via Gray–Scott perturbation, and scale-adaptive boundary supervision (ABALoss) with stage-specific objectives (Sec. 4.3).

These two problems reflect the complementary clinical roles of the modalities—ABUS to recover tumors that projection imaging may miss, mammography to quantify size and margins once visible—and anchor the methods developed in the next chapters.

1.2 Research Methodology

This thesis presents a comprehensive framework for breast tumor detection and segmentation that addresses distinct clinical scenarios through complementary tech-

nical innovations. The overarching objective is to develop computationally efficient yet accurate methods that enhance diagnostic capabilities across both established and emerging breast imaging modalities. This study pursues a dual-track research strategy: optimizing three-dimensional detection efficiency for advanced medical centers equipped with ABUS systems while advancing two-dimensional segmentation precision for ubiquitous mammography platforms.

For three-dimensional ABUS volumes, ABUSDet, a novel 2.5-dimensional detection framework that fundamentally reconceptualizes how volumetric medical data should be processed for object detection tasks, is proposed. Unlike existing approaches that either sacrifice depth information through pure 2D processing [19] or incur prohibitive computational costs through full 3D convolutions [21, 20], ABUSDet introduces an intelligent middle ground by processing multiple adjacent slices simultaneously through a channel squeeze mechanism. This design enables the model to capture inter-slice continuity patterns critical for distinguishing true tumors from artifacts while maintaining the computational efficiency necessary for clinical deployment. The framework further incorporates a deliberate training strategy inspired by cognitive learning theory [36], which dynamically prioritizes difficult samples during training to address the severe class imbalance inherent in screening applications where malignancies represent less than 1% of examined tissue.

For two-dimensional mammography images, this study introduces **TriNet**, a multi-stage iterative refinement network that treats segmentation as progressive boundary optimization. Instead of duplicating U-shaped encoders across stages, TriNet couples a single shared $MobileViT$ encoder (S1) with three unified decoders (S2/S3) to realize a coarse-to-fine pipeline in which an initial localization is progressively sharpened. Each decoder augments its lightweight ConvBlock (pointwise-depthwise-pointwise with optional ECA) with a **Lightweight Row-Column Attention (LRCA)** unit—a width-preserving axial self-attention that factorizes 2D

interactions into row/column components to inject long-range shape cues at low cost. To robustly leverage intermediate predictions, stage-wise priors are stochastically perturbed via Gray-Scott reaction-diffusion and selectively injected at intermediate resolutions (L_1/L_2) where they are fused with the convolutional stream and refined by LRCA. This design delivers crisp boundaries under tight compute budgets and mitigates error propagation from imperfect priors, aligning algorithmic performance with practical deployment constraints across diverse screening settings.

1.3 Main Contributions

This thesis addresses the dual challenges of volumetric detection efficiency and boundary segmentation precision in breast imaging. By developing specialized frameworks for ABUS and mammography, this research bridges the gap between advanced deep learning capabilities and the rigorous constraints of clinical deployment. The primary contributions are three-fold:

1. **Pioneering a Computationally Efficient 2.5D Framework for Volumetric ABUS Detection (ABUSDet)**

   To overcome the prohibitive computational costs of full 3D networks and the spatial incoherence of 2D slice-wise methods, this thesis introduces **ABUSDet**, a novel 2.5D detection framework. By fundamentally reconceptualizing how volumetric data is processed, the model utilizes a learnable **Channel Squeeze** mechanism to aggregate multi-slice context (28 adjacent slices) into compact representations. This approach preserves critical through-plane continuity for distinguishing tumors from artifacts while maintaining the inference speed of 2D detectors. Furthermore, to tackle the severe class imbalance inherent in screening, a cognitive-inspired **Deliberate Training** strategy is proposed. This mechanism dynamically prioritizes "hard" examples by actively filtering out well-trained batches, improving sensitivity to 77.94% at a

low false-positive rate (3.0 per volume) and achieving an inference time of just 11.0 seconds. Crucially, the 2.5D spatial aggregation is physically grounded in the scanner's spatial resolution, inherently preserving clinical interpretability by allowing physicians to trace predictions back to specific sonographic anatomical features—making routine ABUS screening not only clinically feasible but also highly trustworthy.

2. **Architecting a Resource-Efficient Progressive Refinement Network for Mammography (TriNet)**

   Addressing the limitation of standard U-Net architectures in capturing global context without exploding computational budgets, this thesis presents **TriNet**. This architecture treats segmentation as a coarse-to-fine iterative process, utilizing a single shared **MobileViT encoder** paired with sequential lightweight decoders to avoid the redundancy of cascaded networks. A key architectural innovation is the **Lightweight Row-Column Attention (LRCA)** module integrated within the decoders. By factorizing dense 2D attention into axial components, LRCA captures long-range global shape dependencies essential for defining tumor morphology while reducing computational complexity from $\mathcal{O}(L^4)$ to $\mathcal{O}(L^3)$. This design achieves state-of-the-art segmentation performance (88.40% Dice on INbreast) with only 2.83M parameters, demonstrating that high-precision segmentation does not require heavy, resource-intensive models.

3. **Establishing Uncertainty-Aware Prior Integration and Adaptive Boundary Supervision**

   To resolve the fragility of multi-stage networks where errors from early stages propagate downstream, this research introduces a robust uncertainty modeling mechanism. Instead of treating intermediate predictions as determinis-

tic ground truths, **Gray-Scott reaction-diffusion perturbations** are applied to generate diverse, realistic boundary variations, training the refinement stages to be resilient to imperfect priors. Complementing this, a novel **Adaptive Boundary-Alignment Loss (ABALoss)** is proposed to address extreme tumor scale variations. Unlike fixed-scale edge losses, ABALoss dynamically adjusts morphological gradients based on tumor size, ensuring that both microscopic lesions and large masses receive balanced supervision. This combination of uncertainty injection and scale-adaptive supervision significantly enhances boundary fidelity, particularly for spiculated and low-contrast tumors.

Overall, both tracks emphasize deployability: ABUSDet balances volumetric context with throughput for screening workflows, while TriNet refines boundaries without duplicating encoders and incorporates uncertainty-aware priors to bolster clinical trust.

1.4 Thesis Organization

The remainder of this dissertation is organized as follows. Chapter 2 reviews breast tumor analysis across imaging modalities (mammography, ultrasound, DBT, MRI), computational methods (from handcrafted features to CNNs and Transformers), and clinical requirements. This analysis identifies critical gaps in efficient 3D ABUS detection and boundary-accurate 2D segmentation, motivating the dual-track approach developed in subsequent chapters. Chapter 3 presents ABUSDet, a novel 2.5D detection framework for three-dimensional ABUS volumes, detailing the architecture design, deliberate training mechanism, and extensive experimental validation demonstrating superior performance compared to both 2D and 3D baseline methods. Chapter 4 introduces TriNet, a multi-stage segmentation network for mammography images, detailing the progressive refinement with a shared encoder

and sequential decoders, the decoder-side **Lightweight Row-Column Attention (LRCA)** that injects global shape context at low cost, the uncertainty modeling via Gray–Scott prior perturbations, and comprehensive ablations validating each architectural component. Chapter 5 concludes the dissertation by summarizing key findings, discussing clinical implications, and outlining future research directions, including multi-modal fusion approaches and deployment considerations for resource-constrained environments.

Chapter 2

Related Work

Building on Chapter 1's motivation and problem framing, this chapter surveys breast imaging modalities and traces the evolution of breast tumor detection and segmentation methods from handcrafted feature pipelines (2000–2012) through CNN dominance (2014–2021) to emerging Transformer and state-space architectures. By analyzing clinical requirements against computational advances, it identifies critical gaps—particularly efficient volumetric detection for ABUS and boundary-faithful segmentation for mammography—that motivate the ABUSDet and TriNet frameworks developed in Chapter 3 and Chapter 4.

2.1 Overview

Research on breast tumor analysis has progressed from handcrafted pipelines to deep learning, yet notable gaps remain in efficient volumetric detection (*e.g.*, ABUS/DBT) and boundary-accurate 2D segmentation. This chapter surveys modalities, methods, and clinical requirements to delineate those gaps and motivate the contributions. This study organizes prior work along three axes—imaging modalities, computational methods, and clinical requirements—to make the specific gaps that motivate the proposed ABUSDet (volumetric detection) and TriNet (2D segmentation) explicit.

Section 2.2 links imaging physics and clinical workflow to algorithmic requirements across mammography, ultrasound (HHUS/ABUS), DBT, and MRI. Section 2.3 traces methods from handcrafted features through CNNs to Transformers, diffusion,

and state-space models, emphasizing enduring limits in 3D detection and boundary fidelity. Section 2.4 contrasts detection and segmentation requirements. Section 2.5 shows why the field over-emphasizes 2D mammography segmentation while under-addressing 3D ABUS detection, motivating this study's dual-track approach.

Through this analysis, this study argues that tackling these complementary challenges—3D detection efficiency and 2D segmentation precision—offers a practical path to advancing AI-driven breast cancer diagnostics across diverse clinical settings.

2.2 Imaging Modalities for Breast Tumor Detection and Segmentation

The evolution of breast imaging reflects a sequence of trade-offs: each modality addresses specific limitations of its predecessors while introducing new algorithmic challenges. This study examines four modalities—mammography, ultrasound, DBT, and MRI—in order of clinical prevalence, focusing on (i) acquisition physics, (ii) clinical indications, and (iii) implications for detection and segmentation algorithms.

Digital Mammography maintains its position as the cornerstone of global breast cancer screening. Its predominance stems from practical advantages: low per-exam cost, rapid throughput, extensive standardization, and high sensitivity to clustered microcalcifications, a hallmark of early disease [37, 38]. In population screening, mammography is therefore prioritized as the first-line modality, particularly for visualizing microcalcifications that are less conspicuous on other modalities [10].

The extensive availability of annotated repositories has established mammography as the benchmark modality for algorithmic development. INbreast [2] provides high-quality expert annotations for several hundred studies, CBIS-DDSM [3] sup-

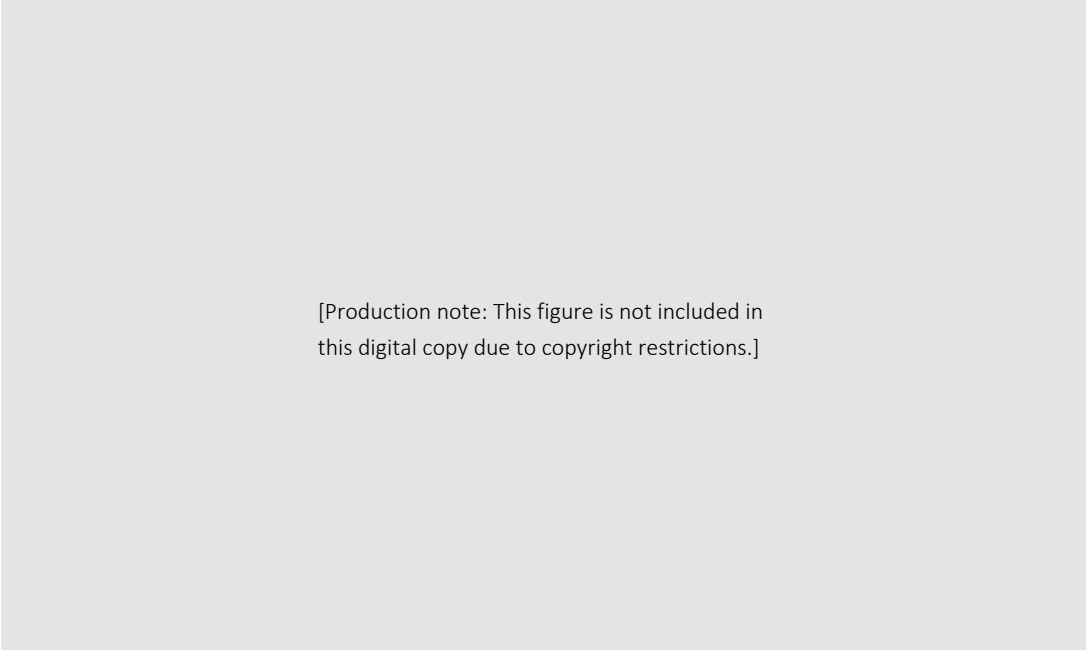


Figure 2.1 : Mammography examples from three public datasets (left to right: INbreast [2], CBIS-DDSM [3], DMID [4]). Yellow contours mark the tumor regions.

plies pixel-level annotations for thousands of images, and DMID [4] further broadens coverage with curated digital mammograms, enabling rigorous training and evaluation. Contemporary segmentation architectures report strong benchmark performance on these datasets, though generalization across institutions still depends on data diversity and harmonization [39, 40, 41]. Fig. 2.1 illustrates three public datasets (left to right: INbreast [2], CBIS-DDSMs [3], DMID [4]); all images are digital mammograms governed by the same X-ray physics. In practice, cross-site and vendor workflows introduce differences in acquisition and rendering (*e.g.*, exposure and compression at capture; reconstruction and contrast mapping during processing; noise/edge handling). These choices alter sharpness, local contrast, and noise texture, which directly influence tumor conspicuity and boundary delineation. In this example, CBIS-DDSM appears softer and noisier than INbreast and DMID, and the reduced micro-contrast can make the mass and its margins harder to appreciate.

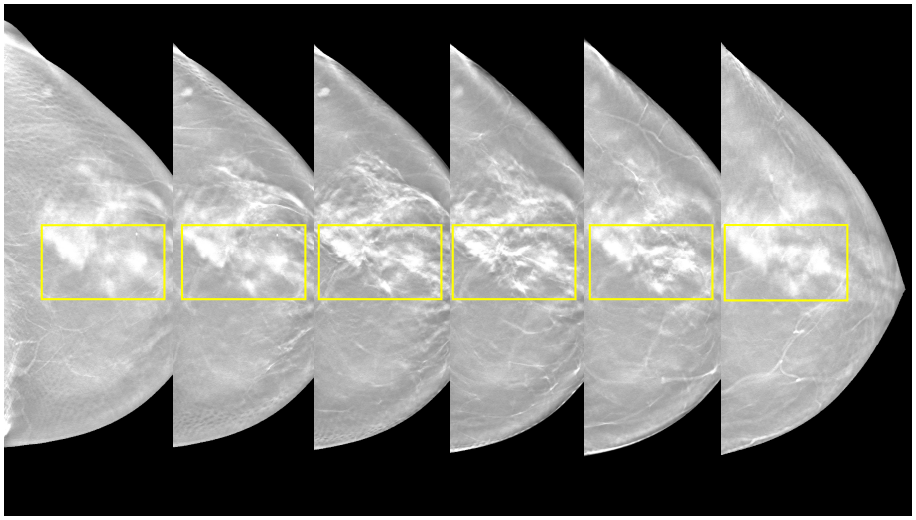


Figure 2.2 : Digital breast tomosynthesis (DBT) examples from the BCS-DBT dataset [5]. Six reconstructed slices from a single volume are shown, sampled every 10 slices in the through-plane direction to highlight appearance changes with depth. Yellow boxes indicate the region of interest containing the tumor.

Nevertheless, mammography is constrained by its two-dimensional projection: tissues at different depths are superimposed into a single plane, potentially concealing true tumors or creating artifactual findings. This limitation is most pronounced in radiographically dense breasts, where sensitivity decreases and recall may rise [9], motivating complementary modalities in dense-breast populations [10].

Digital Breast Tomosynthesis (DBT) aims to bridge the gap between two-dimensional and volumetric imaging by acquiring multiple low-dose projections over an arc and reconstructing quasi-3D image stacks. In clinical practice, DBT uses a limited-angle X-ray sweep that typically acquires 10–25 low-dose projections and reconstructs them into slices on the order of 1 mm thickness of the compressed breast, reducing tissue superposition and improving tumor conspicuity, particularly in dense breasts. Large clinical studies show that DBT can reduce recall rates and improve cancer detection compared with conventional mammography [42, 43, 44]. At the same time, DBT retains ionizing radiation and may increase total dose relative to full-field digital mammography depending on protocol and the use of synthetic 2D

images [45]; its interpretation typically requires additional reader training. Moreover, many validation studies have involved experienced readers and specific practice settings, which may limit generalizability to broader screening programs [46].

As an illustration, Fig. 2.2 shows six slices from one reconstructed DBT volume in the BCS-DBT dataset [5]. To emphasize through-plane evolution, the displayed slices are sampled every 10 slices; a fixed yellow box indicates the region of interest across depth. The progressive change in conspicuity and margins across adjacent slices is typical of DBT stacks and motivates detectors that aggregate 3D context while controlling false positives.

Handheld Ultrasound (HHUS) is a radiation-free modality suitable for younger and gravid patients [11]. Its diagnostic performance is less affected by mammographic density than X-ray-based imaging, supporting its use as a supplemental screening tool in dense-breast populations [47, 48]. However, HHUS suffers from operator dependence (probe angle, pressure), lacks standardized views, and exhibits high inter-observer variability [11]—factors that limit large-scale dataset curation and cross-exam comparability for algorithm development. In addition, HHUS is relatively limited in depicting microcalcifications, which reduces sensitivity for calcification-predominant tumors such as certain presentations of DCIS [10]. Together, these factors complicate large-scale, comparable datasets and motivate the use of complementary modalities in breast tumor detection workflows. As illustrated in Fig. 2.3, two ultrasound frames from the BrEaST dataset [6] demonstrate common field-of-view geometries—rectangular (linear array) and sector-shaped (curvilinear/phased array)—with radiologist freehand tumor contours.

Automated Breast Ultrasound (ABUS) standardizes whole-breast acquisition via automated, volumetric scanning and decouples image capture from interpretation. In routine workflows, a wide linear transducer sweeps each breast under

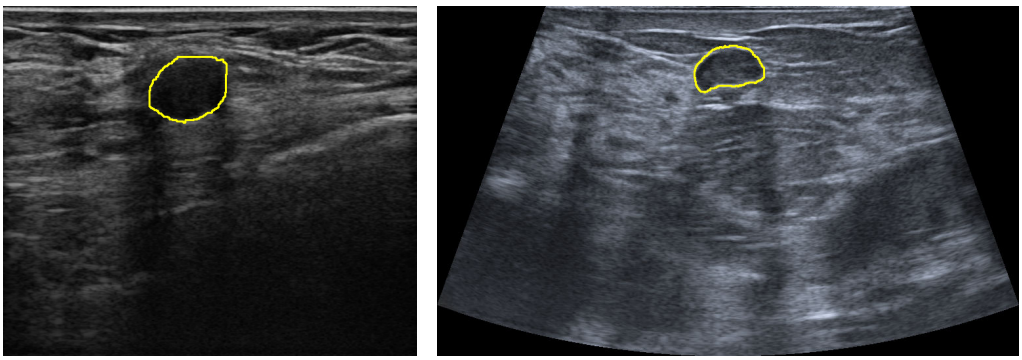


Figure 2.3 : HHUS examples from the BrEaST dataset [6]. Left: rectangular FOV from a linear-array probe; right: sector FOV from a curvilinear/phased-array probe. Yellow contours mark the tumor regions.

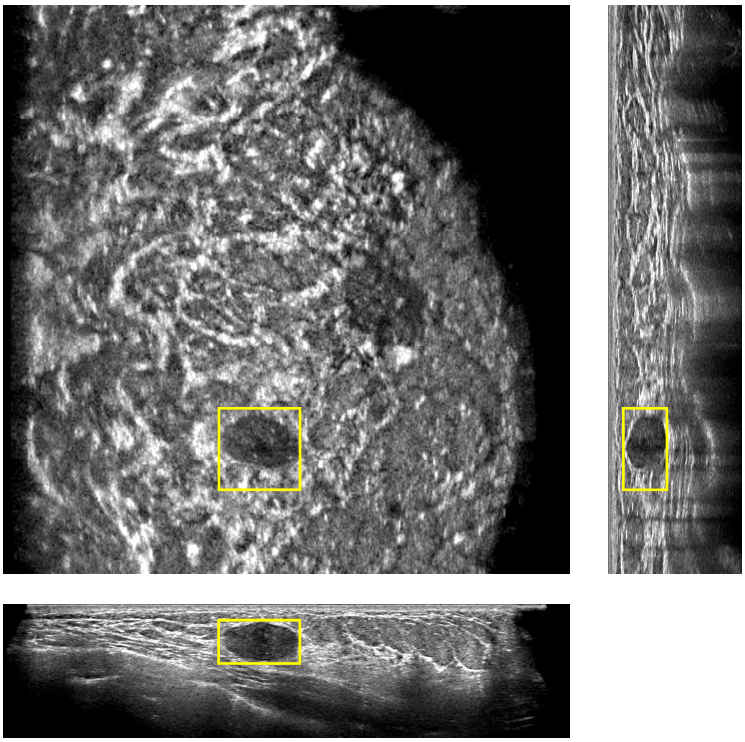


Figure 2.4 : Automated breast ultrasound (ABUS) example (de-identified). The same tumor is shown in the three standard planes from one volume: coronal (top-left), sagittal (right), and axial (bottom). Yellow boxes highlight the suspicious region of interest.

controlled compression and coupling, generating volumetric data—often anisotropic depending on vendor and reconstruction—that are reviewed in three orthogonal planes (axial, sagittal, and coronal). As illustrated in Fig. 2.4, a representative de-

identified case shows the tumor in the three orthogonal planes—coronal (top-left), sagittal (right), and axial (bottom)—with yellow boxes indicating the suspicious region. The coronal reformat is particularly valued for depicting architectural distortion patterns such as the "retraction phenomenon," which can aid malignancy assessment. Compared with handheld ultrasound (HHUS), ABUS offers more reproducible coverage and inter-exam comparability while preserving ultrasound's strengths for mass conspicuity in dense parenchyma [49, 19]. Commercial implementations differ in patient positioning, probe geometry, and reconstruction pipelines; these variations influence slice thickness, in-plane resolution, and coronal reformat algorithms and therefore should be explicitly considered when comparing studies or deploying models across vendors [49, 50].

Clinically, ABUS is most often deployed as a complement to mammography in women with dense breasts, where standardized whole-breast scanning helps reveal tumors obscured by tissue overlap on X-ray-based imaging. The gains in standardization come with practical trade-offs: large field-of-view sweeps demand consistent gel coupling and compression; reading time scales with volume size; and systematic multi-planar review requires reader training. Multi-site adoption also faces device-level and protocol-level heterogeneity that can introduce domain shift in both human interpretation and algorithmic outputs. Mitigation strategies include harmonized scanning protocols, device-aware preprocessing, and external validation across centers with different ABUS platforms [19, 50].

For algorithm design, ABUS's standardized 3D acquisition provides valuable context at the cost of higher memory and latency budgets. End-to-end 3D detectors can leverage volumetric cues and enforce inter-slice continuity, yet they often strain memory and inference time at clinical resolution; 2D slice-based or multi-view compromises reduce resource demands but risk losing cross-slice coherence [51]. Practical pipelines therefore adopt hybrid strategies: patch-wise or slab-wise processing with



Figure 2.5 : Breast MRI examples from the Duke-Breast-Cancer-MRI dataset [7]. Three orthogonal views from one exam are shown: *coronal* (top-left), *axial* (bottom), and *sagittal* (right). Yellow boxes highlight the suspicious region of interest.

feature pyramids, coronal-aware context aggregation, and post-processing that links detections across slices to control false positives per volume [52]. Because volumetric annotations are expensive, many datasets provide tumor centers or bounding boxes rather than full contours, which biases current work toward detection over full 3D segmentation and motivates evaluation that reports tumor-level sensitivity alongside the number of false positives per exam [53, 54].

Radiomics and deep-learning systems on ABUS benefit from modality-specific choices along the standard pipeline of acquisition, ROI delineation, feature learning/extraction, and validation. Speckle-aware denoising and intensity normalization can improve feature stability. ROI definitions should exploit both axial context and coronal-plane patterns. Feature sets or backbones should encode margin sharpness, posterior acoustic behavior, and architectural distortion. Robust study design further calls for clear reporting of vendor mix and reconstruction parameters, multi-center splits to quantify generalization, and calibration/interpretability analyses to

support clinical uptake [50].

Magnetic Resonance Imaging provides three-dimensional, contrast-enhanced visualization with superior soft-tissue contrast and is widely recognized for high sensitivity in selected populations, often exceeding 90% [55]. As illustrated in Fig. 2.5, three orthogonal reformats from a single case—coronal (top-left), axial (bottom), and sagittal (right)—depict the tumor; yellow boxes mark the suspicious region. Images are from the Duke-Breast-Cancer-MRI dataset [7]. Clinically, breast MRI is primarily used for screening of well-defined high-risk cohorts, preoperative assessment of disease extent, and evaluating tumors that are occult on mammography or ultrasound, particularly in dense breasts where X-ray-based conspicuity is reduced [10]. Its broader use is limited by longer acquisition and reading times, higher costs, the need for intravenous contrast, and tolerance issues (*e.g.*, claustrophobia), which collectively reduce suitability for general-population screening.

From an imaging standpoint, contrast-enhanced, time-resolved acquisitions emphasize tumor vascularity and morphology in volumetric stacks. However, tumor conspicuity is modulated by background parenchymal enhancement and motion, both of which vary across patients and sessions and can obscure or mimic pathology. These factors, together with trade-offs between spatial and temporal resolution, directly influence the detectability of small or non-mass enhancement patterns and introduce variability that learning-based systems must accommodate [56].

For detection and segmentation algorithms, MRI poses several practical implications. First, 3D volumes are large, and time-resolved imaging adds a temporal dimension; models therefore balance context utilization against memory and latency constraints. Second, background parenchymal enhancement and subtle non-mass enhancement create challenging edge cases: detection systems must maintain sensitivity without excessive false positives, while segmentation models must delineate

diffuse or heterogeneous boundaries with stability across cases. Third, annotation remains labor-intensive and exhibits inter-reader variability, particularly for non-mass enhancement; clear guidelines and consistent contouring protocols are necessary to obtain reliable tumor masks for training and evaluation. These characteristics make MRI a rich yet demanding substrate for breast tumor detection and segmentation.

Taken together, modality selection is task-specific and population-specific. HHUS serves as a targeted adjunct, ABUS provides standardized whole-breast 3D data that align with volumetric detection workflows, and MRI delivers high sensitivity for high-risk screening. This division of roles motivates algorithm designs and evaluations that account for modality-specific acquisition characteristics, annotation regimes, and deployment constraints.

2.3 Evolution of Breast Tumor Detection and Segmentation Techniques

The algorithmic trajectory of computer-aided diagnosis for breast imaging has unfolded in three broad waves that mirror both methodological innovation and clinical constraints. The first wave (*Traditional Machine Learning,* roughly 2000–2012) standardized a pipeline of handcrafted feature extraction and statistical learning that could be tuned to specific scanners and protocols but struggled with generalization and global context. The second wave (*CNN Dominance with a Transformer Transition,* 2014–2021; Sec. 2.3.2) replaced manual features with end-to-end representation learning, yielding practical gains in sensitivity and throughput but incurring memory/latency costs for 2D images and volumetric exams. The most recent wave (*Emerging Paradigms,* 2020–present; Sec. 2.3.3) pivots toward promptable foundations, efficient global models, generative refinement, and multimodal reasoning to reduce annotation burden and make global context affordable at a clinical scale. Throughout these shifts, two persistent pressures shape design choices: the need

to localize tiny or low-contrast findings (*e.g.*, microcalcifications, spiculated margins) at low false-positive rates, and the requirement to deliver stable, interpretable behavior across institutions and vendors.

2.3.1 Traditional Machine Learning Era (2000–2012)

This period consolidated a canonical workflow: delineate the breast region, generate candidates or regions of interest (ROIs), extract handcrafted descriptors, and train a statistical classifier; when delineations are required, add an energy-minimization segmenter before or after classification. The approach was attractive because it separated domain knowledge from learning and allowed each stage to be debugged and audited, but it also meant that performance hinged on the quality and transferability of the hand-designed features.

Feature engineering targeted complementary aspects of breast tumors. Texture descriptors summarized local or regional intensity organization. Gray-level co-occurrence matrices and Haralick statistics [57] captured spatial gray-level dependencies (contrast, correlation, homogeneity) that correlate with tissue regularity and spiculation. Local Binary Patterns (LBP) [58] provided rotation-invariant microtexture codes that are well suited to modeling ultrasound speckle. When coupled with early statistical classifiers such as Support Vector Machines [59] and Random Forests [60], these handcrafted descriptors formed the core of many pre-deep-learning pipelines. Morphological and shape descriptors—*e.g.*, compactness, eccentricity, circularity, and irregularity—were widely used to quantify mass geometry and margin characteristics, reflecting BI-RADS-inspired intuitions that malignant masses deviate from smooth, convex forms [61]. Intensity and histogram statistics (*e.g.*, first-order moments, peaks, skewness, and kurtosis) were also employed to capture brightness patterns and asymmetries relevant to microcalcifications and echogenicity differences [61], typically alongside edge- and shape-based cues. Be-

cause naively concatenating features often hurts generalisation, feature selection was routinely applied—most notably recursive feature elimination (RFE) [62] and mutual-information criteria such as mRMR [63]—to obtain compact, non-redundant subsets that mitigate overfitting on small datasets. In retrospect, these descriptors codified valuable priors about tissue appearance, yet they were sensitive to acquisition changes (compression, gain, reconstruction) and typically required per-modality retuning.

Detection moved from exhaustive patch classification toward shared-feature candidates to manage image scale. Multi-scale sliding windows [64] or simple candidate generators proposed suspicious regions, after which Support Vector Machines [59] or Random Forests [60] scored handcrafted features, and non-maximum suppression [65] consolidated overlapping hits. To stabilize local decisions, contextual models were introduced: pairwise or conditional priors via Markov Random Fields enforced spatial coherence [66], while graph-based formulations exploited relationships among neighboring regions through normalized or efficient graph cuts [67, 68]. In mammography, simple cross-view heuristics (consistency between CC and MLO) were sometimes added to downweight view-specific artifacts. These pipelines, while serviceable, required careful thresholding and per-institution calibration to achieve acceptable operating points on free-response ROC (FROC) analyses, and computation grew rapidly with window density and multi-scale settings [69, 70, 65].

Segmentation followed an energy-minimization framework with explicit priors on boundary smoothness and region homogeneity. Active contour methods [71] represent a closed curve explicitly and iteratively update it under image-derived forces (*e.g.*, gradient-based attraction) balanced by curvature or length regularization until convergence to a boundary. Level-set formulations [72, 73] encode the boundary as the zero level of a scalar field and evolve this field via partial differen-

tial equations obtained from variational energies; the implicit representation permits splitting and merging without additional bookkeeping. Graph-cut methods [74, 75] discretise binary labeling and minimize a submodular energy that combines data fidelity with pairwise smoothness terms; the global optimum for this class is obtained with an s-t min-cut/max-flow algorithm. Watershed segmentation [76] computes catchment basins on a gradient-derived image by identifying regional minima and performing marker-controlled flooding to delineate boundaries while reducing oversegmentation. Region growing [77] expands from user- or algorithm-defined seeds by iteratively adding neighboring pixels that satisfy intensity-similarity and connectivity constraints.

These techniques can perform reliably when edges are distinct and boundaries are approximately smooth, and their energy terms offer interpretable knobs for user guidance through seeds, markers, or regularization weights. In breast imaging, however, heterogeneous textures and speckle in ultrasound reduce edge salience; posterior acoustic shadowing and partial-volume effects obscure interfaces; and spiculated or ill-defined margins violate smoothness assumptions. Typical failure modes include active-contour evolution trapped in local minima, leakage across weak edges for graph cuts, and fragmentation into many small basins for watershed; region growing is sensitive to seed placement and similarity thresholds. Initialization and ROI quality exert strong influence—imprecise seeds or coarse masks propagate errors—particularly in dense mammography and speckled sonograms where tumor and background intensity distributions overlap substantially [78].

Evaluation and practice reflected the modular pipeline. Classifiers were benchmarked by ROC/AUC [79]; detectors by sensitivity versus false positives per image or exam using FROC analyses [69, 70]; and segmenters by Dice/Jaccard overlap and boundary distances (*e.g.*, Hausdorff). Data scarcity and limited public benchmarks hindered reproducibility; studies often reported internal cross-validation

without robust external-site testing, despite mounting evidence that vendor and protocol shifts degrade performance [80, 81, 78]. Preprocessing—breast region extraction, pectoral muscle suppression, intensity normalization—was essential to reduce variance but introduced additional hyperparameters that complicated technology transfer.

Synthesis: strengths, breakpoints, and the road to learned representations. Handcrafted pipelines yielded transparent systems whose errors could be traced to specific modules, and they established a common set of radiomics-style descriptors that remain informative covariates in multi-modal studies. Yet three structural limits persisted. First, feature locality and rigidity impeded recognition of complex, hierarchical patterns that distinguish malignant from benign tumors across scanners and populations; features tuned on one cohort frequently underperformed elsewhere [80, 78]. Second, latency and tuning burden rose with image size and candidate density, conflicting with screening throughput and increasing the engineering overhead of per-site recalibration [82]. Third, segmentation robustness suffered in exactly the regimes that matter clinically—heterogeneous, low-contrast, or spiculated tumors and speckled ultrasound. These pressures motivated a transition to learned representations that could internalize multi-scale context and reduce manual feature design, setting the stage for the deep-learning era discussed in Secs. 2.3.2–2.3.3.

2.3.2 CNN Dominance and the Transformer Transition (2014–2021)

The deep learning revolution began with AlexNet's ImageNet breakthrough [83], followed by VGG [84], ResNet [85], and other foundations that established convolutional feature hierarchies as the default paradigm. Transfer learning from ImageNet to medical imaging became standard practice [86], with early breast CAD studies demonstrating that pre-trained CNNs outperformed handcrafted features even with

limited annotations [87, 88].

The first wave of deep learning for breast tumor analysis was driven by convolutional architectures that translated successes from natural-image detection and segmentation into screening and diagnostic workflows. Compared with handcrafted pipelines, convolutional models offered end-to-end optimization, stable training recipes, and strong inductive bias for locality and translation, all of which were advantageous under small, class-imbalanced datasets and heterogeneous acquisition protocols. At the same time, clinical imaging imposed distinctive constraints: ultrahigh-resolution digital mammography (FFDM), volumetric digital breast tomosynthesis (DBT), and automated breast ultrasound (ABUS) stress memory and latency budgets; microcalcifications and spiculated margins require fine boundary fidelity; and deployment demands predictable throughput across scanners and institutions.

Detection evolved from patch classification to shared-feature two-stage systems and then to one-stage regression. Early clinical systems abandoned sliding-window patch classifiers in favor of region-based CNNs that share features across many candidate windows. Fast R-CNN [89] and Faster R-CNN [90] introduced ROI pooling and a Region Proposal Network, respectively, to couple region generation and classification within a single trainable backbone. These designs mattered for mammography because they amortized computation over extremely large images and, with feature pyramids, preserved sensitivity to small, low-contrast findings. Mask R-CNN extended the paradigm with an instance mask branch that is directly useful when delineations are required alongside tumor boxes [91]. In parallel, one-stage detectors such as YOLO [92] and SSD [93] reframed detection as direct bounding-box regression over grid anchors. In screening settings with tight throughput constraints, one-stage families often offered a favorable speed-sensitivity trade-off, while two-stage methods retained advantages at very low false-positive operating points. Several large-scale mammography studies reported that deep CNN detectors, trained

and evaluated with FROC protocols on retrospective test sets, surpassed classical CAD in sensitivity at fixed false-positive operating points, consolidating CNNs as the default detection backbone by the late 2010s [87].

Segmentation consolidated around encoder-decoder designs with skip connections, with boundary refinements as needed. Fully convolutional networks (FCNs) [94] established dense prediction for medical imaging, and U-Net [14] popularized a symmetric encoder-decoder with lateral skips that fuse coarse semantics and fine spatial detail. For breast tumors—whose appearances range from circumscribed masses to spiculated, low-contrast regions—skip fusion proved critical for preserving edges and thin structures that are otherwise lost to pooling. Variants such as nested skip pathways (U-Net^{++}) [25] and attention-gated decoders [95] mitigated background clutter and improved small-tumor sensitivity without excessive computational penalty. In practice, clinics and research groups paired these networks with modest post-processing (*e.g.*, connected-component constraints, dense CRFs) to sharpen boundaries when annotations emphasized contour fidelity. Evaluation conventions coalesced around Dice/IoU and boundary metrics (*e.g.*, Hausdorff), typically alongside exam-level AUC or Tumor-level FROC for companion detection/classification tasks.

Clinical realities of scale and dimensionality exposed the limits of locality. Ultra high-resolution images require reasoning across the entire breast to suppress false positives from benign structures, while DBT/ABUS introduces long-range through-plane dependencies. Purely local kernels can struggle to model such context efficiently. Workarounds—tiling with overlap, multi-view fusion for CC/MLO, and "2.5D" stacks that aggregate a few adjacent slices—help, but they complicate training and can break cross-slice consistency. End-to-end 3D CNNs capture volumetric context but demand prohibitively large memory and longer per-exam latencies on full-resolution DBT/ABUS, motivating architectural changes that increase the effective receptive field without quadratic cost.

The success of convolutional encoders on breast and broader oncologic imaging was followed by a gradual but substantial shift toward Transformers. The Transformer era began with Vaswani et al.'s attention-is-all-you-need architecture [96], followed by Vision Transformer (ViT) [97], which demonstrated that pure attention can match CNNs on ImageNet given sufficient data. DETR [98] reframed object detection as set prediction with bipartite matching, removing many hand-tuned components such as anchor design and non-maximum suppression. Several widely cited designs in medical imaging made these ideas practical for tumor segmentation and related dense prediction. TransUNet [15] coupled a ViT-style encoder with a U-shaped decoder and demonstrated that explicit long-range context improves delineation on medical benchmarks. For volumetric tumor analysis, UNETR [17] cast 3D segmentation as a sequence-to-sequence problem with a Transformer encoder and multi-scale skips to a CNN decoder, reporting strong results on brain-tumor tasks under standard protocols. In parallel, Swin-UNet [16] replaced the convolutional encoder-decoder with a hierarchical, shifted-window Transformer and showed that windowed attention sustains high-resolution medical segmentation with manageable compute. Task-specific evidence in oncology reinforced the recipe: TransBTS [99] integrated a 3D CNN with Transformer tokens for multimodal brain-tumor segmentation, mitigating slice-wise fragmentation; TransFuse [28] fused CNN and Transformer streams to obtain accurate polyp masks under tight latency.

Taken together, the pre-2022 period established a practical pattern for bringing global context to medical tumor segmentation: keep U-shaped decoders for boundary detail, apply Transformer encoders or token mixers for long-range dependencies, and control memory with windowing or hybrids. These choices motivated the post-2022 focus on scalable 3D context, medical-tailored pretraining, and modality-aware token interactions.

Training and evaluation practice matured alongside architectures. To

manage class imbalance and enrich tumor appearances, groups relied on tumor-centric sampling, cost-sensitive objectives, and augmentation portfolios (geometric/photometric jitter, elastic deformations). Multi-view fusion for CC/MLO pairs became standard in mammography classification/detection; breast-region extraction, pectoral muscle suppression, and intensity normalization reduced cross-institution variance. Reporting converged on external validation where possible, a response to mounting evidence that domain shift (scanner vendor, reconstruction, population) can erode apparent gains measured on single-site datasets [86]. Public datasets such as INbreast and CBIS-DDSM supported more reproducible comparisons, though the gap to true screening distributions remained a persistent caveat.

Discussion and critical comments. The CNN-dominated period delivered reliable, deployable baselines, especially when paired with careful data curation and operating-point selection. However, three structural limitations motivated the transition that follows. First, modeling global context on ultrahigh-resolution images and through-plane coherence remained cumbersome with purely local operators; early attention-based methods alleviated this but introduced nontrivial memory/latency overhead that must be managed via tiling or windowed designs. Second, sensitivity to domain shift persisted across paradigms: gains reported on internal data often attenuated under external-site validation, underscoring the need for robust preprocessing, calibration, and explicit reporting of out-of-distribution behavior. Third, the engineering cost of sustaining 3D context at clinical throughput (minutes per exam rather than seconds per image) left room for architectures that either make global reasoning cheaper or reduce annotation needs through interactivity. These pressures set the stage for the 2020s focus on promptable segmentation foundations and efficient long-range sequence models.

2.3.3 Emerging Paradigms (2020–Present)

Recent progress moves beyond purely convolutional pipelines toward models that capture global context efficiently, learn from sparse or weak labels, and support interactive or multimodal use. These directions are particularly relevant to breast imaging, where subtle tumors must be localized against complex backgrounds, annotations (especially volumetric) are costly, and deployment must respect strict memory and latency budgets.

Self- and weakly supervised pretraining. Self- and weakly supervised paradigms aim to reduce reliance on dense pixel-level labels by learning strong visual priors from unlabeled data. Masked autoencoders (MAE) [100] reconstruct heavily masked inputs and produce semantically rich encoders that transfer well to downstream segmentation with limited annotations. Consistency-based semi-supervised methods such as dual-task consistency [101] and uncertainty-rectified pyramid consistency [102] regularize multi-head or multi-scale predictions on unlabeled images, improving boundary stability and hard-case learning. In parallel, transformer-based volumetric encoders have been adapted to self-supervision: Swin-UNETR [103] pretrains hierarchical windowed transformers on 3D data, then fine-tunes them for organ and tumor segmentation. These ideas have also been explored in breast imaging settings where large unlabeled mammography or ultrasound archives can be exploited to bolster performance in label-scarce regimes.

Transformers after 2022. Work after 2022 has focused on making Transformer backbones practical for high-resolution *2D* and volumetric medical image analysis. Swin-UNETR [104], VT-UNet [105], UNETR++ [106], and mmFormer [107] illustrate a common pattern: hierarchical or windowed self-attention encoders capture long-range context, while U-shaped decoders with skip connections or pro-

gressive up-sampling restore fine spatial detail. These architectures achieve strong performance on diverse segmentation benchmarks and are conceptually well aligned with breast imaging, where models must reason over large fields of view in full-field mammography and across slices or modalities in DBT and ABUS without violating clinical throughput constraints. Design principles such as multi-scale tokenization, cross-scale feature fusion, and modality-aware context aggregation have influenced subsequent breast-focused work on both detection and segmentation.

Diffusion models. Denoising diffusion probabilistic models (DDPMs) [108] provide a powerful generative prior by learning a reverse denoising process from Gaussian noise to clean images. When conditioned on an input scan, diffusion models can sample segmentation masks from an implicit posterior, yielding accurate tumor delineations together with pixelwise uncertainty estimates [18]. Related work leverages "pseudo-healthy" synthesis to highlight anomalies as residuals [109], and MedSegDiffV2 [110] shows that transformer-enhanced conditional diffusion can act as a general-purpose medical segmenter across diverse datasets. For breast imaging, these techniques are attractive for handling heterogeneous, low-contrast lesions and for explicitly quantifying boundary uncertainty, complementing faster discriminative detectors and segmenters.

Vision–Language Models. Vision–language models (VLMs) align images with free text and open new ways to use weak supervision from reports and clinical notes. CLIP [111] demonstrated large-scale contrastive pretraining on image–text pairs, and medical variants such as ConVIRT [112] and BioViL [113] extend this idea to radiology, providing visual encoders that transfer well to detection and segmentation tasks. Open-vocabulary localization frameworks like GLIP [114] and CLIPSeg [115] further allow text prompts to guide bounding boxes or masks at inference time. In

breast imaging, these concepts support, for example, using mammography reports as weak labels for lesion localization or retrieving regions that match textual descriptions of suspicious findings, thereby reducing reliance on exhaustive pixel-level annotations.

Promptable SAM-style segmenters. The Segment Anything Model (SAM) [116] popularized promptable segmentation: a high-capacity image encoder is coupled with a lightweight decoder that turns points, boxes, or coarse masks into refined object masks without category-specific training. Medical adaptations such as MedSAM [117], SAM-Med2D [118], and SAM-Med3D [119] fine-tune SAM-like architectures on large collections of medical images and demonstrate accurate, easily editable masks for organs and lesions across modalities, including breast imaging. Task-specific variants such as Polyp-SAM [120] show that further specialization can match or surpass dedicated architectures while retaining interactive flexibility. For breast cancer diagnosis, SAM-style models are particularly appealing as interactive tools: radiologists or upstream detectors can provide coarse prompts around suspected masses, which are then refined into high-resolution, editable tumor masks.

State-space models. Selective state-space models (SSMs) such as Mamba [121] offer an alternative to self-attention by modeling long-range dependencies with linear-time complexity in sequence length. U-shaped and segmentation-oriented variants including U-Mamba [122], SegMamba [123], and VM-UNet [124] integrate SSM blocks into encoder–decoder architectures, achieving competitive results on high-resolution 2D and 3D segmentation benchmarks while maintaining favorable memory and runtime profiles. These designs are well suited to breast imaging scenarios such as full-field mammography, DBT, and ABUS, where each exam corresponds to a long, anisotropic sequence and efficient global context is essential for suppressing

false positives without sacrificing sensitivity.

Synthesis and relation to this thesis. The paradigms above address complementary bottlenecks in contemporary medical image analysis: self- and weakly supervised learning reduce annotation demands; Transformer and SSM backbones provide efficient global context for high-resolution and volumetric data; diffusion models offer strong shape priors and calibrated uncertainty around difficult boundaries; VLMs connect imaging with clinical language and weak supervision; and promptable SAM-style segmenters enable fast, interactive refinement of lesion masks. In this thesis, we build on these trends conceptually but focus our experimental study on methods that directly target breast tumor detection in ABUS volumes and breast mass segmentation in mammography, have accessible implementations or can be faithfully reproduced, and satisfy realistic memory and latency constraints in screening workflows. The broader families reviewed here are included to situate ABUSDet and TriNet within the evolving landscape of medical image analysis, but an exhaustive empirical comparison with every variant would require large-scale retraining and multimodal pretraining beyond the scope of this work.

2.4 Clinical Requirements for ABUS Detection and Mammography Segmentation

This section outlines the specific clinical requirements that motivate the two tasks addressed in this thesis: tumor detection in ABUS volumes and tumor segmentation in mammography images.

Detection requirements in ABUS volumes. In dense-breast screening, ABUS must detect tumors that may be obscured on mammography due to tissue superposition [10]. Clinicians require high sensitivity to avoid missing cancers while maintaining specificity to minimize unnecessary recalls [47, 48].

Detection systems must preserve through-plane coherence to distinguish true tumors from artifacts and to assess multi-focal disease [51]. The coronal reconstruction plane is particularly valued for visualizing architectural distortion patterns that aid malignancy assessment [49]. Systems must handle the computational demands of volumetric data while maintaining practical throughput for screening workflows [50].

Segmentation requirements in mammography. Accurate tumor boundaries are essential for TNM staging, which directly influences treatment decisions [37]. Margin characteristics (smooth, irregular, spiculated) are key BI-RADS descriptors that help distinguish benign from malignant lesions [10].

Segmentation must accommodate the wide spectrum of tumor presentations, from well-circumscribed masses to ill-defined spiculated lesions [34]. For surgical planning, precise boundary delineation helps determine excision margins and assess the need for neoadjuvant therapy [33].

Cross-modality deployment requirements. Clinical systems must maintain performance across different imaging equipment and protocols used at various institutions [80, 78]. Integration into existing workflows requires predictable processing times and minimal disruption to established reading patterns [82].

Interpretability and uncertainty communication are critical for radiologist acceptance, particularly when AI recommendations differ from initial clinical assessment [81]. Systems should provide clear visualization of detected regions and confidence levels to support clinical decision-making [117].

2.5 Research Gaps and Motivation

The preceding review across imaging modalities (Sec. 2.2), algorithmic evolution (Sec. 2.3), and clinical requirements (Sec. 2.4) indicates that current research emphasis is unbalanced. Methodological development is concentrated on 2D mam-

mography segmentation, while volumetric ABUS detection and clinically aligned boundary modeling in mammography are less fully addressed. This section summarizes the two core gaps that directly motivate ABUSDet and TriNet.

Gap 1: Volumetric ABUS detection under screening constraints. ABUS is specifically designed for standardized, whole-breast 3D acquisition and is routinely used as a supplement to mammography in dense-breast screening. However, existing ABUS detection methods either process slices independently with post-hoc 3D fusion or adopt fully 3D sliding-window strategies. Slice-wise approaches are efficient but weaken through-plane coherence and are prone to fragmented or inconsistent tumor representation across depth. Fully 3D approaches preserve volumetric context but incur prohibitive memory and runtime at clinical resolution and often assume dense volumetric annotations, which are expensive to obtain. In practice, there is a lack of ABUS-specific frameworks that preserve essential inter-slice context, operate within screening-grade memory and latency budgets, and remain compatible with the weak volumetric labels that are realistic in clinical datasets. ABUSDet is introduced to fill this gap by using a 2.5D detection design that aggregates adjacent slices through channel compression, maintains through-plane continuity, and keeps computation in the regime of strong 2D detectors.

Gap 2: Boundary-focused, prior-aware mammography segmentation within 2D budgets. For mammography, existing models typically optimize region overlap metrics and produce single-shot predictions, which improves average Dice but does not explicitly target the accurate delineation of complex boundaries that matter for staging and margin planning. Multi-stage refinement networks and prior-guided designs do exist, yet they commonly duplicate heavy encoders at each stage, treat intermediate priors as deterministic, or apply uniform losses across stages, leading

to redundancy, error accumulation, and suboptimal use of intermediate information. There is therefore a need for mammography segmentation frameworks that emphasize boundary fidelity rather than only global overlap, exploit progressive refinement with a shared encoder and lightweight decoders to stay within practical 2D compute budgets, and integrate prior information together with explicit uncertainty modeling instead of relying on fixed priors. TriNet is proposed in response to this gap: it treats segmentation as a coarse-to-fine process, shares a single encoder across stages, injects perturbed priors to encode prior uncertainty, and applies stage-specific supervision that separates recall-oriented and boundary-oriented objectives.

Synthesis and forward look. In summary, the related work motivates two complementary research directions. The first is to design an ABUS detection framework that makes volumetric, context-preserving analysis feasible under screening-grade constraints; this is realized by ABUSDet in Chapter 3. The second is to design a mammography segmentation framework that aligns with clinical emphasis on boundaries and prior information while remaining computationally lightweight; this is realized by TriNet in Chapter 4. Together, these models translate the identified gaps into task-specific, clinically oriented solutions.

Chapter 3

ABUSDet: A Novel 2.5D Deep Learning Model for Automated Breast Ultrasound Tumor Detection

Extending the gaps identified in Chapter 2, this chapter presents ABUSDet, a 2.5D framework that stacks adjacent slices for 2D detection and clusters in 3D to recover volumetric locations. On ABUS data, it attains 77.94% sensitivity at 3.0 FP/volume with ∼11 s/volume inference, balancing accuracy and throughput; these outputs set up the complementary segmentation tackled in Chapter 4.

3.1 Introduction

Breast cancer has surpassed lung cancer as the most commonly diagnosed cancer [125]. Early diagnosis and screening with suitable treatment can reduce mortality rates significantly in the long term [126]. On World Cancer Day 2021, the International Cancer Institute highlighted two key research projects, *i.e.*, reducing the risk of breast cancer and improving the survival rate of women with breast cancer. They aimed to emphasize the importance of taking action against breast cancer. On March 8, 2021, World Health Organization (WHO) introduced the Global Breast Cancer Initiative (GBCI), of which the primary objective is to reduce global breast cancer mortality by increasing access to early diagnosis and prompt comprehensive cancer management. After investigating the WHO Cancer Mortality Database from 1950 to 2016 and the literature review from 2010 to 2020, it was summarized in [127] that, among countries achieving sustained mortality reductions, at least 60%

of breast cancers are diagnosed at Stages 1 or 2, reinforcing the significance of early clinical diagnosis for improving the breast cancer outcomes.

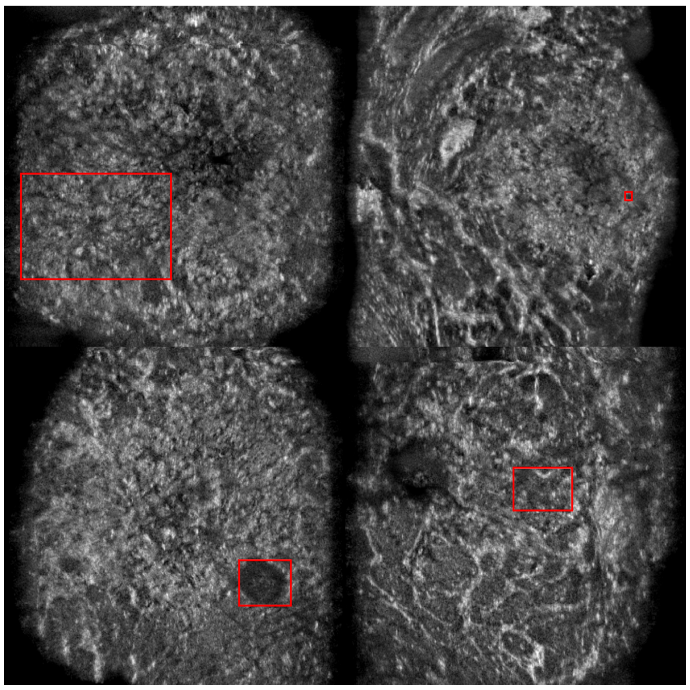


Figure 3.1 : Samples of tumors with different morphological patterns (Coronal View).

As the first generation of technology for breast tumor detection, mammography is the primary tool employed for early diagnosis and screening of breast cancer and has been well studied. However, mammography suffers from the undeniable problem of contraindication [128]. Besides, mammography's sensitivity and specificity are statistically significantly lower than ultrasound in dense breasts and for young women [129]. Therefore, Handheld Ultrasound (HHUS) and Magnetic Resonance Imaging (MRI) are often used for further evaluation of questionable findings [130] as adjuncts of mammography. However, HHUS also has limitations, such as operator dependence and imperfect reproducibility. MRI has superior sensitivity compared to other technologies, but it also has a high sensitivity to benign breast diseases [131]. Automated Breast Ultrasound System (ABUS) [132] is a new ultrasound screening technology for breast tumor diagnosis, which produces high-quality 3D volume data

with high screening efficiency and low operational dependence. It can produce hundreds of slices in each scan, producing data with more details. Nevertheless, it can be very time-consuming for physicians to review thousands of slices for each patient. In addition, the size, shape, and texture of tumors vary for different patients, so it further increases the difficulty of review. Fig. 3.1 shows four samples of ABUS slices with tumors from four patients. The tumor in the lower-left image is the easiest to detect, whereas the other tumors are not as easy to predict. In order to reduce the burden on physicians, speed up the screening, and improve the accuracy of tumor detection, researchers have proposed various computer-aided detection systems for ABUS tumor detection [12, 133, 13, 21, 20, 134, 19, 49]. The latest and most significant solutions for computer-aided ABUS tumor detection systems are briefly introduced in the next sub-section. This study then presents the key ideas of the proposed approach.

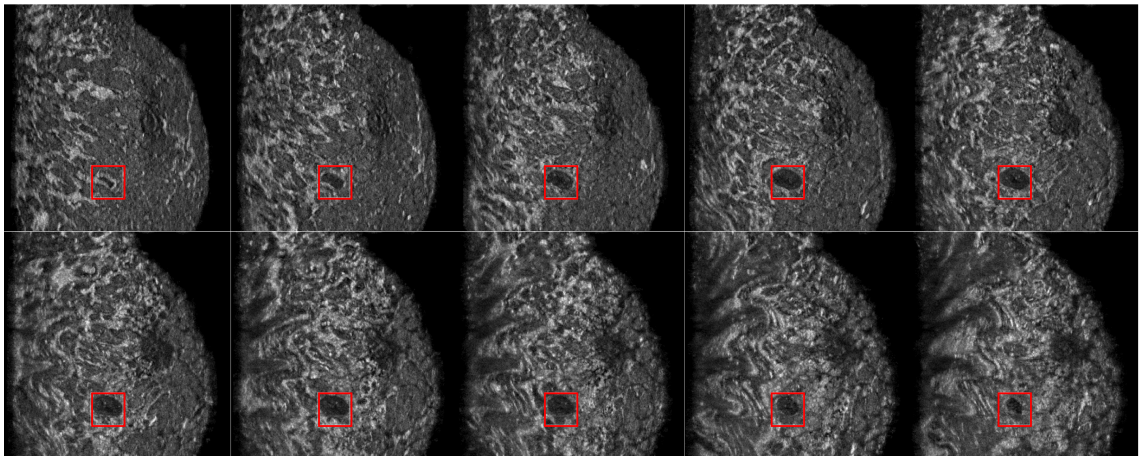


Figure 3.2 : A range of slices of data blocks showing one tumor.

3.2 Related Work

Machine learning was the primary tool for ABUS tumor detection before 2017. For example, Moon *et al.* [12] proposed a Hessian analysis with a multi-scale blob detection method, Tao *et al.* [133] proposed a multi-stage system, and Lo *et al.* [13]

proposed a Multi-Dimensional Tumor Detection system using Topographic Watershed. These traditional machine learning methods use hand-crafted features that require professional knowledge, and therefore, there is an accuracy bottleneck that cannot be broken through. Moreover, as the false-positive rate is high, it is not very practical for physicians.

Deep learning based solutions do not need hand-crafted features, and the detection rate outdistances the traditional machine learning methods. In [21], Wang *et al*. proposed a 3-dimensional (3D) CNN architecture to assist breast cancer tumor detection. A new threshold map was first designed to render the voxel-level threshold to classify cancer voxels from standard tissue regions to obtain a lower false-positive rate. Then, multi-scale discriminative features were designed to improve the sensitivity. Chiang *et al*. [20] proposed a computer-aided detection system that used the 3D CNN to estimate the tumor probability of the volume of interest (VOI) extracted by sliding windows. The VOI with a higher tumor probability is then selected as a candidate tumor to be clustered by a hierarchical clustering algorithm. Wang *et al*. [134] developed an innovative 3D CNN and a densely deep supervision method to augment the detection sensitivity by effectively using multi-layer features. Moreover, they used a threshold loss to present a voxel-level adaptive threshold for discriminating cancer against non-cancer. Zhou *et al*. [19] designed a 3D multi-view position analysis scheme to fuse 2D detection results and obtain the final 3D bounding boxes. They used a modified Faster R-CNN to extract 2D candidates. Zhang *et al*. [49] proposed a Bayesian YOLOv4 network based on Monte Carlo dropout (MC-Drop) to introduce uncertainty into the detection network.

The aforementioned deep learning methods can be classified into 2D object detection and 3D object detection. The first class uses 2D object detection methods to locate the tumor's position in each slice and then uses an aggregation algorithm to calculate the 3D coordinates. The second class uses 3D sliding windows to get

3D cubes and then uses 3D object classification methods to estimate the probability of tumors and calculate the 3D coordinates. Both of these methods have their advantages and disadvantages. While the 2D approaches work well for detecting objects in 2D images, when using them to deal with 3D volume data, their accuracy drops, and the false-positive rate increases, as the 3D depth information is lost. As for the 3D object detection methods, on one hand, each 3D sliding window contains three dimensions of information, making more accurate detection possible. On the other hand, the size and shape of tumors are random, while the 3D sliding window size is fixed, so it is powerless to detect unfixed-size tumors using a fixed-size sliding window. Moreover, using multiple 3D sliding windows with a small stride is time-consuming.

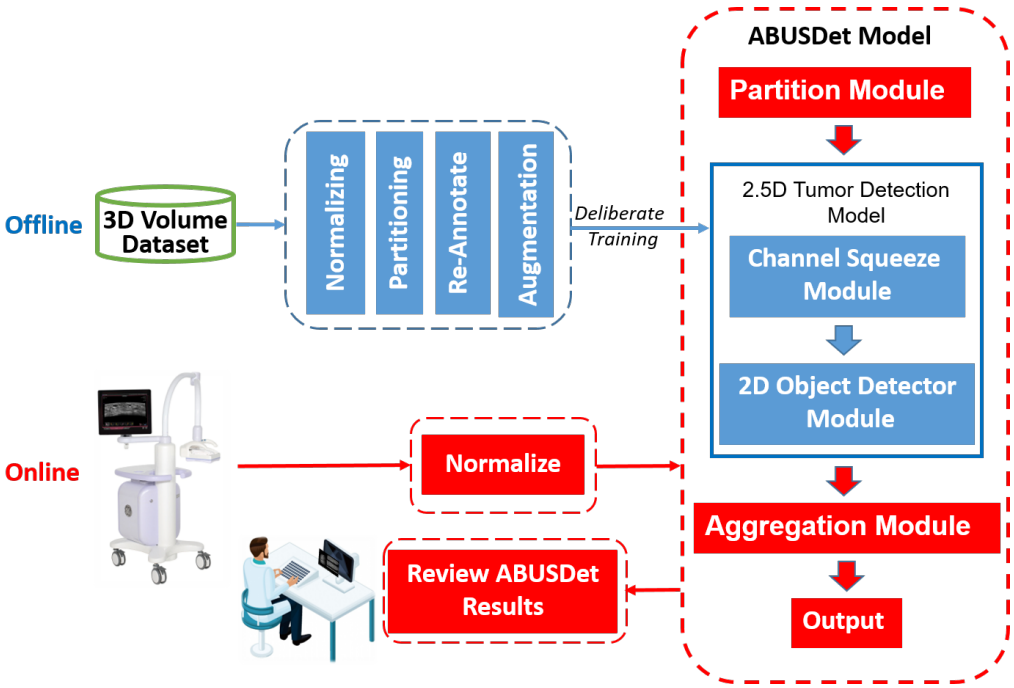


Figure 3.3 : The system architecture of the proposed ABUSDet.

As shown in Fig. 3.3, the ABUSDet system includes offline and online stages.

3.3 Methodology

In order to improve the sensitivity and suppress false positives (FPs), this study proposes a novel tumor detection method named ABUSDet. The 2.5D tumor detection model of ABUSDet perfectly inherits all the advantages of 2D object detection approaches, such as pre-training from models that have been trained with thousands of general image data. Moreover, the proposed 2.5D model detects tumors from 2.5D data that contains depth information, which improves detection accuracy.

3.3.1 ABUSDet Architecture

Table 3.1 : The hyperparameters of the proposed architecture.

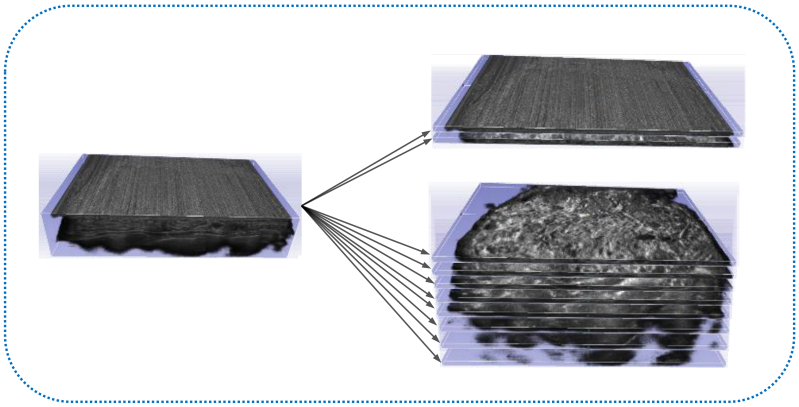
Stage	Hyper-parameter	Value
Model	2D object detection model	Faster R-CNN ResNet-50
	Channel Squeeze module	Neural Network
	Data Block Channels	28
Training	Deliberate Practice Loss	"phNum"
	Epochs	24
	Batch size	4
Testing	Confidence	0.2
	Stride	10
Aggregation	"min samples"	1
	"eps"	6mm
Optimizer	Type	"SGD"
	Learning Rate	0.005
	Momentum	0.9
	Weight Decay	0.0001
Learning Rate Policy	Type	"CosineAnnealing"
	Warmup	"exp"
	Warmup Epochs	8
	Warmup Ratio	0.1
	Min Lr	0.0001

In the offline stage, the datasets are pre-processed by normalizing, partitioning, re-annotation, and augmentation. Then, the datasets are used to train the 2.5D tumor detection model by a newly developed "deliberate training" method. In the online stage, the input ABUS image for tumor detection is first normalized and then

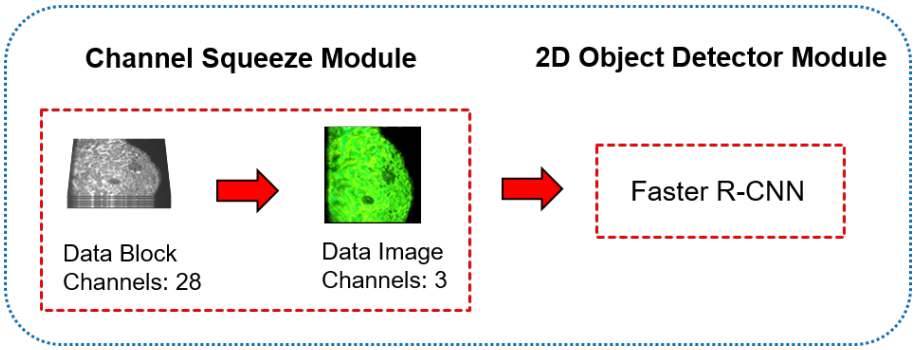
processed by the proposed ABUSDet for tumor prediction. Finally, the output of ABUSDet is sent to the physician. ABUSDet comprises a partition module, a 2.5D tumor detection module, and an aggregation module. The partition module splits the input 3D volume into a series of 2.5D data blocks by the given channel length and stride.

The 2.5D tumor detection module is used to detect the location of tumor candidates from the 2.5D data blocks. In the aggregation module, the 3D coordinates of each tumor candidate are calculated according to its 2D coordinates and depth information, and finally, all the detected tumor candidates are clustered. The 2.5D tumor detection module integrates a Channel Squeeze neural network with a 2D object detection module. The Channel Squeeze module squeezes multi-channel information into three channels, which are then delivered to the 2D object detection module to execute the detection task. The 2D object detection module chosen here is a Faster R-CNN model. Instead of feeding the model with one or three slices at a time, as in 2D object detection models, the 2.5D model can process multiple adjacent slices simultaneously. The depth information reflected by continuous slices provides additional information that helps for producing more accurate detection.

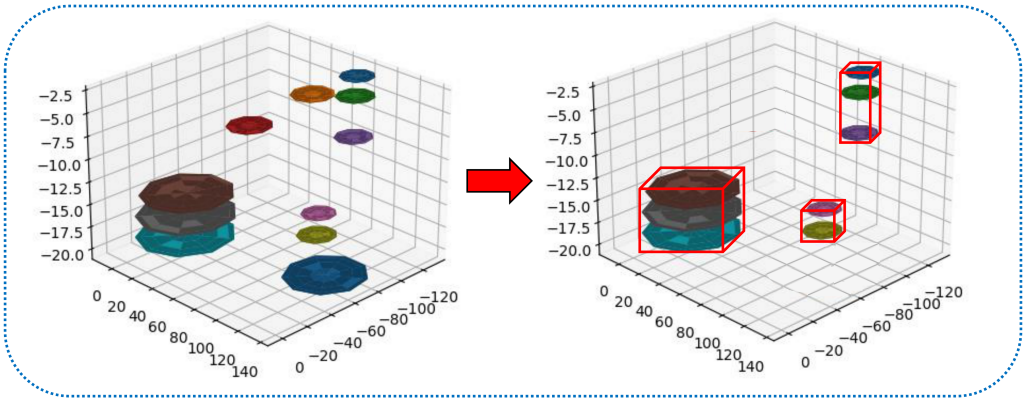
The Data Partition module is used to split 3D volume data into 2.5D data blocks, as shown in Fig. 3.4a. Then, a 2.5D tumor detection approach is proposed to locate the tumor from the 2.5D data blocks, as shown in Fig. 3.4b. Finally, when all 2.5D data blocks divided from the 3D volume are calculated, the aggregation algorithm fuses the localization results of all blocks and produces the location of the tumor on the volume level, which is illustrated in Fig. 3.4c. Table 3.1 presents the hyperparameters with the corresponding values at different stages of ABUSDet.



(a) The data partition module splits the 3D volume data into 2.5D data blocks (Coronal View).



(b) 2.5D tumor detection module



(c) Aggregation module

Figure 3.4 : Illustration of the key modules in the proposed ABUSDet breast tumor detection model.

3.3.2 Pre-processing

Data Partition Module

Each DICOM (Digital Imaging and Communications in Medicine) volume has 330 frames of slices (Axial View) in a three-dimensional array. The width and height

of each frame are not fixed due to the selected "Transducer Frequency". First, the 3D volume data is split into a series of 2.5D data blocks in the data processing stage. The shape of the 2.5D data block depends on three parameters, *i.e.*, the "splitting direction" represents the view that needs to be acquired after data partition (including coronal, sagittal, and axial), the "channel number" represents how many channels are contained in the 2.5D Data Blocks, and the "stride" represents the step length to cut 3D volume. Considering the range of the tumor size and verified by experiments, the values of "channel number" and "stride" are set to 28 and 10, respectively. The "splitting direction" is "Coronal View." Fig. 3.4a demonstrates how the 3D volume is divided into 2.5D data blocks from a coronal view.

For each volume, an experienced physician annotates each tumor. When the 2.5D data block is extracted in the offline stage, the counterpart bounding box will be generated in the format of 2D by the re-annotation procedure. Fig. 3.2 demonstrates several 2.5D data blocks containing tumors divided from a 3D volume.

Feature Scaling

In order to make sure all the pixels are on a similar scale, feature scaling, as shown below, is necessary:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{3.1}$$

For each 3D volume, before the partition module, the data is normalized using Eq. 3.1, where X is the input 2.5D data block, X' is the output of feature scaling, X_{min} is the minimum value among all pixels of input data, and X_{max} is the maximum value of the data. Since almost all pixel values in the data are between 0 and 200, this study sets X_{min} to 0 and X_{max} to 200.

Data Augmentation

The limitation of labeled data sets in the medical imaging area is a significant problem. Data augmentation is widely used to solve the dilemma of insufficient data. In the 2D object detection area, various image manipulation techniques have been used to enlarge the data set, such as rotation, scaling, and flipping. This study's 2.5D data blocks adopt these methods too. These data augmentation skills can be manifested online or offline. The offline method produces data in advance and stores it in memory. The online method manipulates data during training. This study uses rotation, scaling, and flipping by an online method, considering computation time and the limitations of memory space.

During data loading, each data block is flipped with a probability of 50%. In addition, it is rotated with the same probability to an angle randomly selected from 0° to 360° in 30° increments. For all rotation angles other than 0°, 90°, 180°, and 270°, the data block is scaled to 75% of its original size.

In the offline stage, there are two options when partitioning 3D volume data into 2.5D data blocks. First, only 2.5D data blocks containing the ground truth tumor are used for training. Second, data blocks without the Ground Truth tumor (emptyGT) are also used for training. The experiments revealed that the FPs decrease after adding data without tumors, and the sensitivity of the proposed model increases.

3.3.3 2.5D Tumor Detection Model Design

Object detection from 2D images has gained fantastic development, and many successful models and methods have been proposed for medical imaging. To fully utilize the 2D object detection's advantage, this study designs a 2.5D tumor detection model in a modular structure. A 2.5D tumor detection model is constructed by plugging a Channel Squeeze module in front of a 2D object detection model, so that the 2.5D model inherits the fast and accurate advantages of 2D models and obtains

depth information. It reduces the training time and improves training accuracy through transfer learning. To the best of our knowledge, using a channel squeeze module to compress multiple channels to adapt to a 2D object detection model is a novel contribution of this study. A similar concept appeared in [85], but it was to reduce and then increase dimensions to construct a bottleneck module with smaller input and output dimensions. Another related concept was proposed in [135], where a module was used to squeeze global spatial information into a channel descriptor.

After data partitioning and augmentation, several 2.5D data blocks can be obtained from each 3D volume. Essentially, a 2.5D data block is a multi-channel image, which differs from the traditional 2D pictures with either one or three channels. Therefore, a Channel Squeeze module is designed to convert a multi-channel data block to a 3-channel image. The Channel Squeeze module in the proposed model is a convolutional neural network with the size of input channels to be 28, the size of output channels to be 3, the size of kernel size to be 1, and the stride to be 1. It can be plugged into any of the 2D object detection models.

The 2D object detection model is the basis of the whole model. In this thesis, Faster R-CNN is employed as the 2D object detector module, as shown in Fig. 3.4b.

3.3.4 Aggregation Module and Post-processing Strategy

After the 2.5D tumor detection model calculation, a set of candidate detections is obtained for each volume, including 2D coordinates, depth information, and tumor probability scores. To balance screening sensitivity and computational efficiency, candidates with a tumor probability below a heuristic threshold of 0.2 are initially eliminated. The rationale for choosing this relatively low confidence threshold is rooted in the core clinical requirement of breast cancer screening: prioritizing high sensitivity to prevent fatal false negatives. A relaxed 2D confidence threshold preserves ambiguous but potentially malignant regions, while leaving the task of filtering

out isolated false positives to the subsequent 3D spatial aggregation phase.

Subsequently, the DBSCAN algorithm [136] is applied to cluster the remaining candidate sets and reconstruct the final predicted 3D tumor locations, as visually demonstrated in Fig. 3.4c. DBSCAN is explicitly chosen over other clustering methods (such as K-means) because it does not require the number of clusters to be predefined, perfectly aligning with the unpredictable number of tumors in a patient's breast. Furthermore, DBSCAN dynamically identifies arbitrarily shaped clusters and handles spatial outliers, making it highly suitable for the irregular morphology of breast tumors spanning multiple continuous slices.

The specific clustering hyperparameters, namely the neighborhood radius ("eps") and the minimum number of points ("min_samples"), are strictly calibrated based on the physical dimensions of the ABUS volume. As detailed in the dataset description, the inter-slice spacing is 0.506 mm, and the sliding stride is set to 10 slices. This results in a physical depth interval of exactly 5.06 mm between adjacent 2.5D data blocks. To successfully aggregate multiple 2D bounding boxes of a single tumor spanning across adjacent blocks (visualized as the vertically stacked discs on the left of Fig. 3.4c), the clustering radius "eps" must be slightly larger than this 5.06 mm stride. Therefore, setting "eps" to 6 mm provides a precise spatial margin to link detections of the same tumor, while effectively preventing the erroneous merging of distinct nearby lesions.

To ensure clinical sensitivity for early-stage screening, "min_samples" is deliberately set to 1. In DBSCAN, this parameter dictates the minimum number of points required to form a valid cluster; setting it to 1 allows an isolated detection to be recognized as a valid tumor rather than being discarded as noise. This parameter configuration is physically grounded in the scanner's spatial resolution and the biological morphology of the targets: because the smallest tumor depth in our

dataset (3.8 mm) is strictly less than the block stride (5.06 mm), an early-stage lesion may naturally manifest as a single valid detection in exactly one sliding block. Consequently, such a minute tumor may only be prominently captured and successfully detected in exactly one sliding block. When the window steps forward by 5.06 mm, the tumor is entirely bypassed. Setting "min_samples" to 1 ensures that these early-stage, single-hit tumors are preserved as independent clusters (as perfectly illustrated by the smallest red bounding box containing a single isolated detection in the right side of Fig. 3.4c), thereby effectively safeguarding the model's overall sensitivity.

Finally, for the volume-level prediction, the maximum probability value within each cluster subset is taken as the overall probability of the 3D tumor. As a tumor is scanned across slices, its most prominent cross-section yields the highest detection confidence. Taking the maximum score naturally selects the most reliable 2D slice representation for the entire 3D volume, whereas utilizing an average score would inappropriately dilute the overall confidence with lower-scoring predictions derived from the ambiguous boundary slices of the tumor.

3.3.5 Transfer Learning

Transfer learning plays a vital role in deep learning. For the 2.5D tumor detection model, transfer learning can help make it easier to train. In the medical image area, transfer learning includes the cross-domain model pre-trained on natural images and the cross-modal model pre-trained on medical images [137]. In the 2.5D tumor detection model, the 2D object detector is Faster R-CNN with a backbone ResNet-50 [85], while the ResNet-50 is pre-trained on ImageNet [138]. After using the cross-domain transfer learning, it needs only 24 epochs to fine-tune the 2.5D tumor detection model completely.

3.4 Deliberate Training

Deliberate training is a novel mechanism to train the deep learning model. When introducing a deep learning model into a computer-aided system, an important step is to make the model constantly updated. Physicians review images of patients and upload the latest data with annotations in real-time to the cloud for preservation and training of deep learning models. In this continuous learning process, how to allocate training data is a key problem that needs to be solved. If the model trained with all the data is inefficient, it will waste power and computing resources. In order to focus purely on the currently under-trained data in each training epoch, this study proposes a new deep learning training mechanism: deliberate training. This mechanism can work in online or offline methods. This study only demonstrates its offline method function.

The inspiration for deliberate training comes from the deliberate practice proposed by psychologist Ericsson [36]. Ericsson identified a set of conditions that can significantly improve one's skill acquisition performance. The conditions include 1) a given task with a well-defined goal, 2) motivation to improve, 3) provision with feedback, and 4) provision with ample opportunities for repetition and gradual refinements of their performance.

This study hypothesizes that the deliberate practice method used to train the human brain can also be used to train neural network models in deep learning. The existing training mechanism already meets the deliberate practice requirements in deep learning. However, inspired by the fourth condition (ample opportunities for repetition), this study proposes an improvement scheme to enhance the existing training process. The existing neural network training method trains all the data in every epoch, while the model cannot be well-trained, as the samples are hard to learn. However, under-trained data can be given more opportunities to repeat

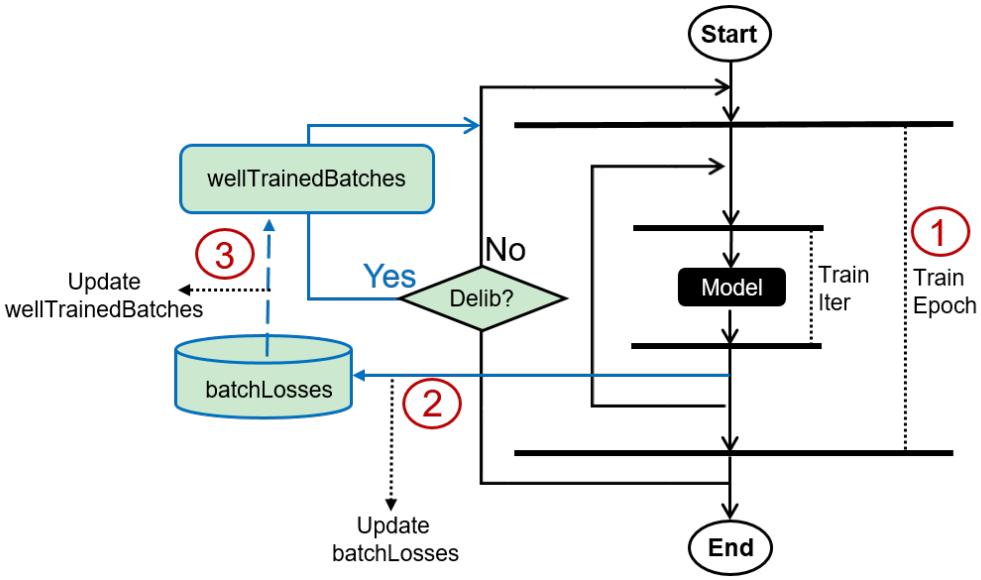


Figure 3.5 : The schematic diagram of the proposed deliberate training scheme for training the deep learning model.

training by actively ignoring current well-trained data.

Fig. 3.5 is the schematic diagram of the proposed deliberate training scheme for training the deep learning model. Algorithm 1 is the corresponding algorithm. In Fig. 3.5, the number in red circles represents the significant operation. Operation 1 is a Train Epoch, corresponding to lines 2 through 9 in Algorithm 1. Lines 2 through 4 in Algorithm 1 are executed before each Train Epoch. Before the training, this study defines which epochs need to be trained by deliberate training and then saves their indices in the set delibEpochs. Line 2 in Algorithm 1 determines whether the current epoch needs to perform deliberate training. If not required, set wellTrainedBatches to an empty set.

Operation 2 corresponds to lines 6 to 8 in Algorithm 1 and is performed once after each Train Iter in each Train Epoch. This operation calculates the batch delibLoss and updates it to batchLosses.

Operation 3 corresponds to lines 10 to 15 in Algorithm 1 and is performed once after each Train Epoch. This operation evaluates the training performance of each

Algorithm 1: Deliberate Training

Input: *dataset*: An ordered list saves data batches for training.
numberBatch: Number of batches to be trained.
numberEpoch: The number of times to completely train all data.
delibEpochs: A set of epoch indices that need to be executed in deliberate practice.

Variable : *batchLosses*: An ordered list to save delibLoss of each batch.
wellTrainedBatches: A set of Batch indices that have been trained relatively well.

Function: *median*: Returns the median of the array elements.
trainGetLoss: Train and return delibLoss calculated from Eq. 3.1.

```
for currentEpoch ∈ [0, numberEpoch] do
    if currentEpoch ∉ delibEpochs then
        wellTrainedBatches ← ∅
    end
    for currentBatch ∈ [0, numberBatch] do
        if currentBatch ∉ wellTrainedBatches then
            batchLosses[currentBatch] ←
             trainGetLoss(dataset[currentBatch])
        end
    end
    wellTrainedBatches ← ∅
    for currentBatch ∈ [0, numberBatch] do
        if batchLosses[currentBatch] < median(batchLosses) then
            wellTrainedBatches ← wellTrainedBatches ∪ {currentBatch}
        end
    end
end
```

batch by using the batch losses updated in each training Epoch. Then, a part of the samples with minor training losses is selected not to participate in this round of training. The median value of batchLosses is used to decide which batchLosses are well-trained in the current epoch. This study compared three loss functions, *i.e.*, the positive hit number (phNum), the positive hit number with Non-maximum Suppression [65] (phNum+NMS), and the positive missed number (pmLoss). The phNum value means all the true-positive numbers of candidates in a batch. On the basis of phNum, this study applies NMS to the candidate of each image and gets the

positive hit numbers as the phNum+NMS. The larger the phNum or phNum+NMS in a batch, the better trained the batch. The pmLoss indicates the false-negative numbers in a batch. The less the pmLoss in a batch, the better the batch is trained.

3.5 Experiments and Results

The ABUS dataset used In this thesis was acquired between 2017 and 2019 in the Traditional Chinese Medicine Hospital of Guangdong Province from an Automated Breast Ultrasound Scanner (Invenia1.2) made by GE Healthcare. The device produces 330 frame slices, and the "Spacing Between Slices" is 0.506 mm. Each slice's "Pixel Spacing" is 0.082 mm and 0.2 mm along posterior-anterior and left-right axes. The "Transducer Frequency" used in the ABUS data in this study can be 8180 Hz, 9000 Hz, or 10000 Hz. Due to the different "Transducer Frequency," each slice's width and height are not fixed. Due to strict patient confidentiality regulations, ethical guidelines, and the proprietary data protection policies of the collaborating hospital, the ABUS dataset used in this study cannot be made publicly available.

Three regions need to be scanned in each breast for complete coverage for each patient. This study gets six DICOM images for each patient, called RAP, RLAT, RMED, LAP, LLAT, and LMED, but only those containing tumor images are used in this study. The dataset consists of 259 DICOM images from 97 patients between 23 to 65 years old, including 315 tumors. The histogram of tumor size (length of the longest side) is shown in Fig. 3.6. From this chart, it has been seen that most of the tumors are not bigger than 50 mm. The mean size of tumors is 19.6 mm. The most extensive tumor is $73.8mm \times 20.5mm \times 51.1mm$. The smallest tumor is $2.9mm \times 3.1mm \times 3.8mm$.

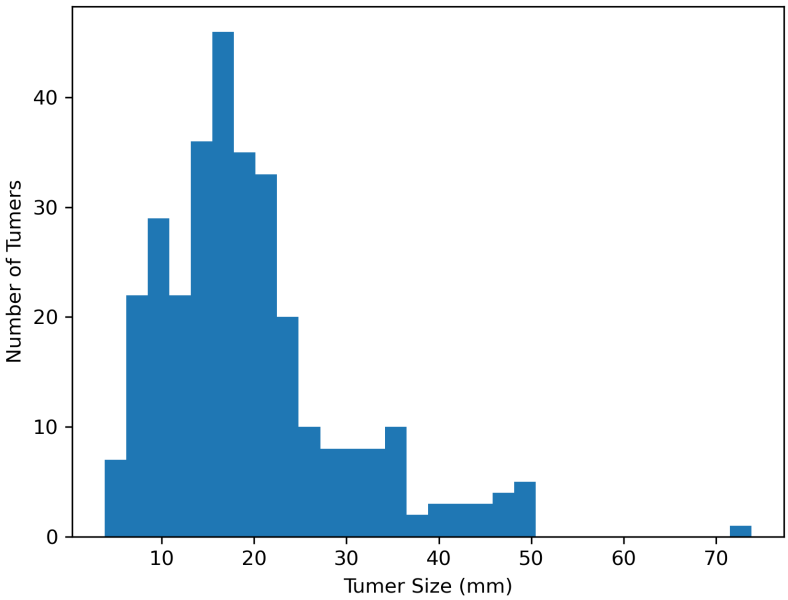


Figure 3.6 : Histogram of tumor size distribution in the ABUS dataset.

3.5.1 Implementation Details

This study randomly divides all patient data into the Training Sets, Validation Sets, and Test Sets. Only volumes containing tumors are used. The Training, Validation, and Test sets include 35, 17, and 35 patients with 90, 44, and 101 volumes and 102, 66, and 120 tumors, respectively. To strictly prevent data leakage, the dataset split was performed exclusively at the patient level. Therefore, the training, validation, and test sets are completely mutually exclusive in terms of patient identity, meaning no volume data from the same patient appears in more than one subset. All of the experiments are evaluated by the free-response receiver operating characteristics (FROC) curves [139] and the area under the curve (AUC) [140]. The ordinate of FROC represents Sensitivity, and the abscissa represents the average value of FPs/volume. A predicted box will be treated as a true positive if its center is contained in the ground truth box and the distance between the center of the predicted box and the center of the ground truth box is less than 15mm.

3.5.2 Results and Analysis

Experiments on Data Augmentation:

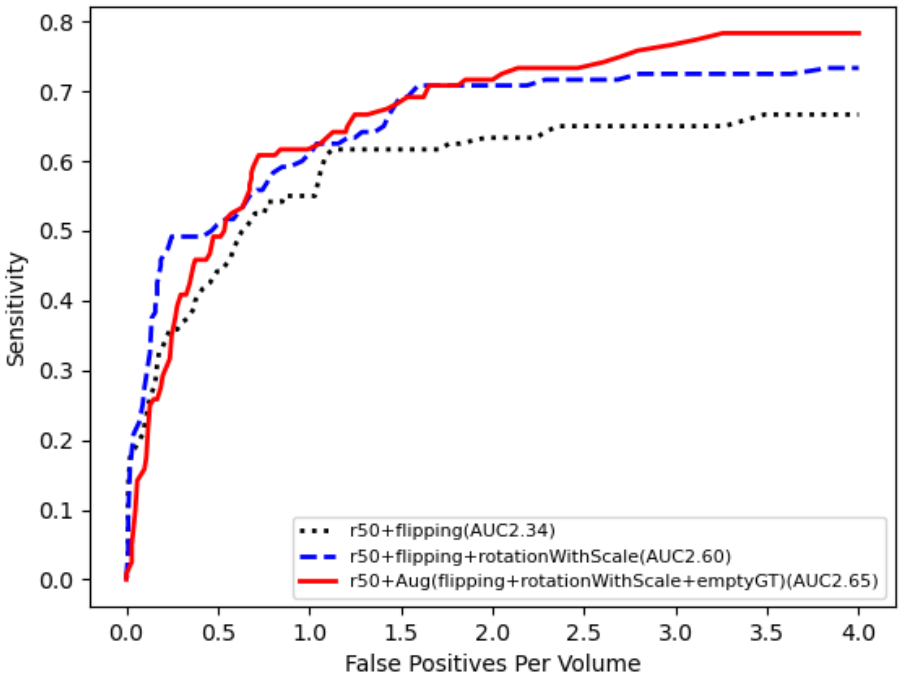


Figure 3.7 : Using data augmentation and emptyGT data

Firstly, this study designs a comparative experiment to compare the effects of different data augmentation techniques on the results. This study compares three augmentation methods of "flipping," "flipping and rotation with scale," "flipping, rotation with scale, and adding emptyGT" (Aug). The experimental results, shown in Fig. 3.7, show that the results are improved significantly after using the combination of flipping, rotation with scale, and emptyGT (Aug). Therefore, the ABUSDet-related experiments are based on augmentation (Aug).

Experiments on Deliberate Training.

Fig. 3.8 shows the results of regular training (without using deliberate loss) and deliberate training with different deliberate losses, *i.e.*, pmLoss, phNum or phNum+NMS. Experimental results have verified that the sensitivity is improved from

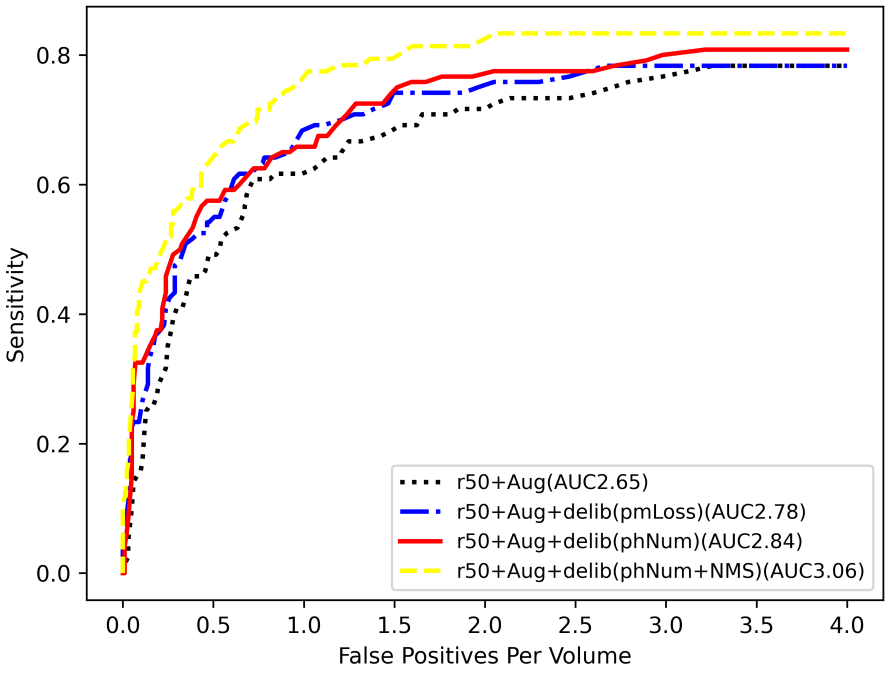


Figure 3.8 : Comparison of different deliberate losses

78.33% to 83.33%, and the AUC increased from 2.65 to 3.06 after using deliberate training with phNum+NMS as a loss function. Moreover, it can be seen from the curve in Fig. 3.8 that the results of deliberate training are better than those of ordinary training methods for the sensitivity corresponding to any FPs/volume value. It should also be noted that the control experiments below are conducted based on phNum as a loss function instead of the optimal phNum+NMS.

Experiments on 2.5D Tumor Detection Parameters.

Fig. 3.9 shows the results of 2.5D tumor detection when selecting different channel numbers. From this experiment, it can be seen that when the value of stride remains unchanged, both sensitivity and AUC increase with the increase of channel numbers, and they reach the optimal value when the number of channels is 28. The experimental results prove that 2.5D tumor detection with multiple channels is superior to 2D object detection. When the number of channels is too large, such as 35, small and thin tumors are easily omitted, resulting in decreased sensitivity.

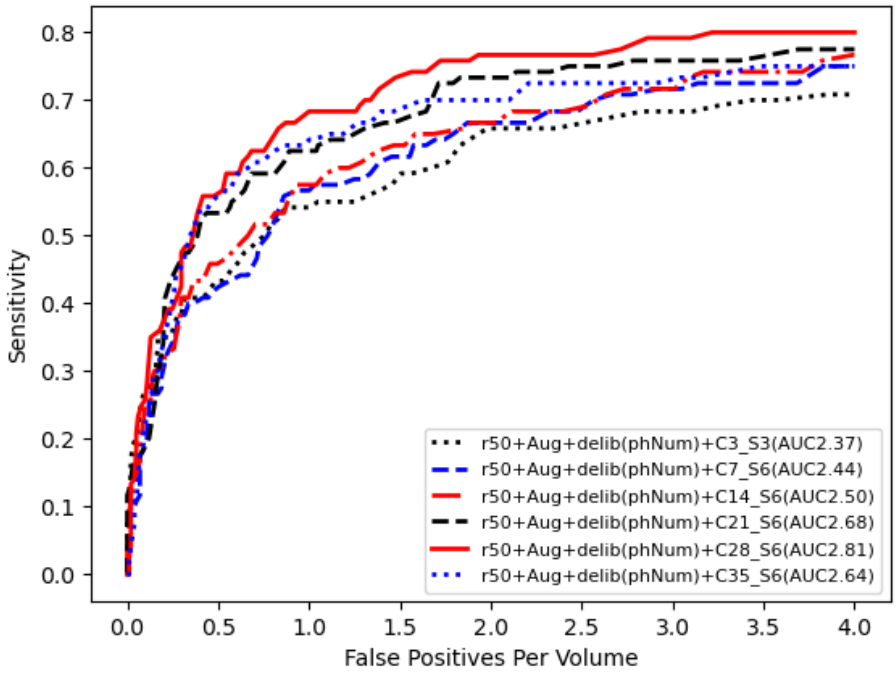


Figure 3.9 : Comparison of different channel numbers of the Coronal View.

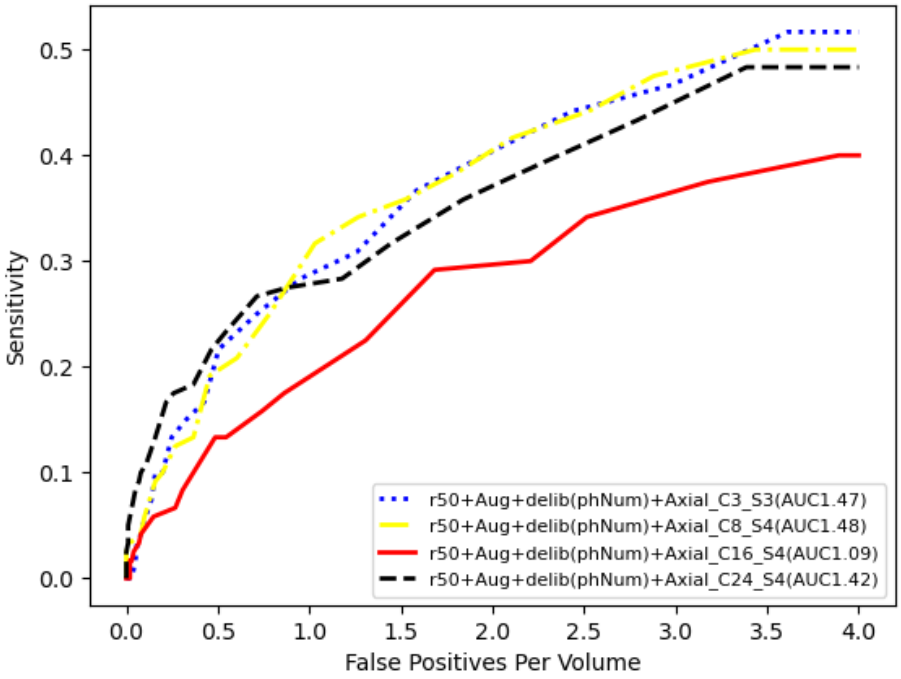
Fig. 3.10 (a) and (b) show the tumor detection results by cutting data blocks from Axial View and Sagittal View, respectively. Their results are not as good as the Coronal View.

Fig. 3.11 shows the results of different strides under the condition that deliberate training is adopted and channel numbers are set to 28.

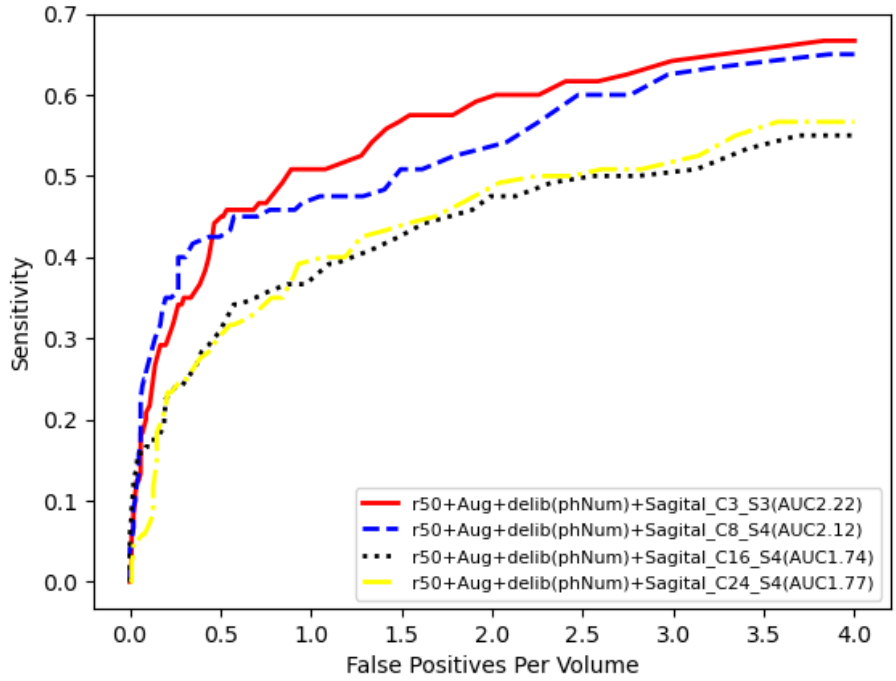
Table 3.2 : Execution time per volume of different strides in the ABUSDet model.

Stride	3	6	10	14	21	28
Time	35.5s	17.7s	11.0s	6.6s	5.7s	**3.8s**

The experimental results show that maximum sensitivity is higher when the stride is 10. When the stride is set to 14, the AUC is the highest. Moreover, the larger the stride is, the faster calculating a volume is. The execution time of these methods is listed in Table 3.2. When channels are set to 28, the test results under different strides show similar performance. The execution time per volume largely



(a) Axial View



(b) Sagittal View

Figure 3.10 : Comparison of different Channel numbers of the Aerial View and Sagittal View.

depends on the partitioning strategy. When the channel number is fixed, the larger the stride is, the shorter the execution time is.

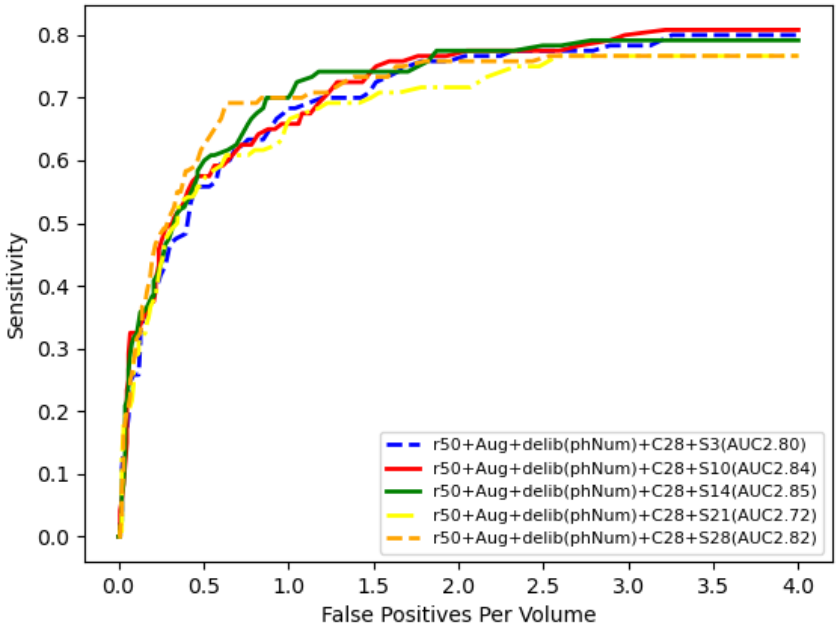
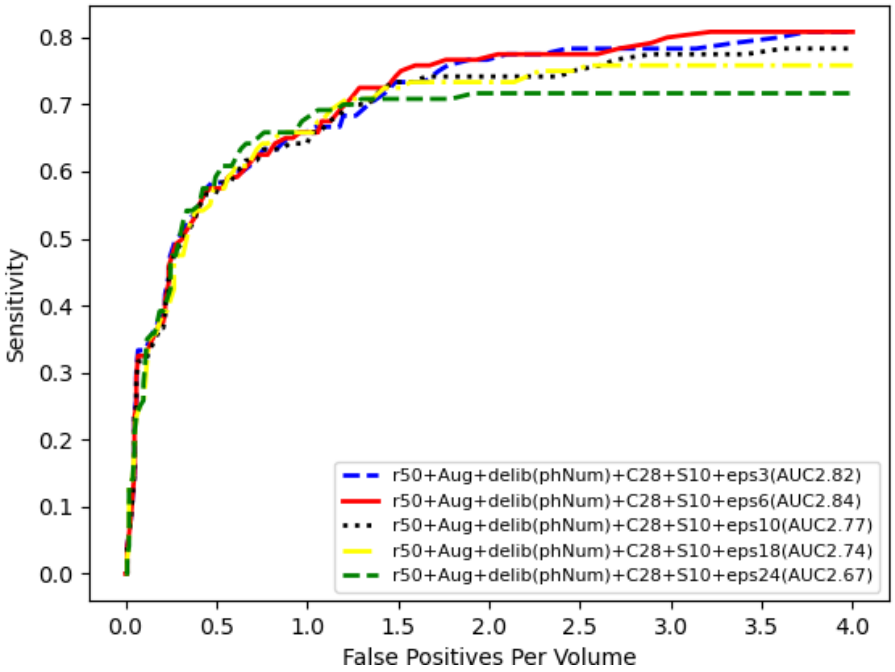


Figure 3.11 : Sensitivity obtained with different strides.

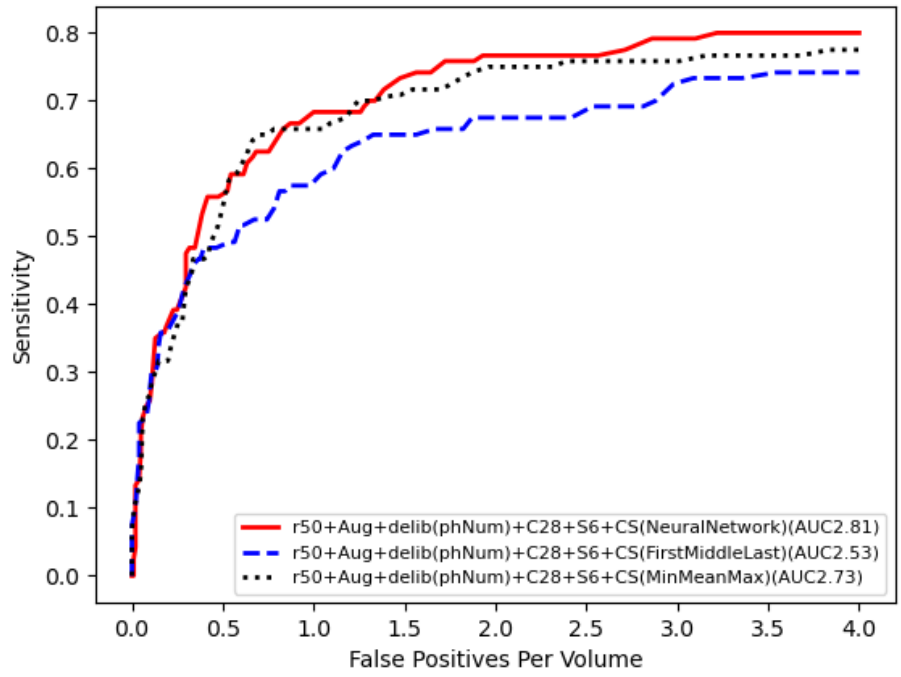
During the test, the selection of the parameter "eps" of the clustering algorithm affects the final result. If the distance between two points in candidates is lower than or equal to "eps", these points are considered in the same cluster. It can be seen from Fig. 3.12 (a), the smaller the value of "eps" is, the higher the sensitivity is. Overall, the best "eps" is 6.

This study uses the neural network model as the Channel Squeeze module and compares it with other methods. FirstMiddleLast extracts the 1st, 14th, and 18th of the 28 channel outputs. MinMeanMax is three channels that extract the minimum, average, and maximum values of 28 channels per pixel. As shown in Fig. 3.12 (b), the neural network is the best Channel Squeeze method with superior performance.

Based on the above experiments, the optimal parameters of the ABUSDet model can be obtained, as shown in Table 3.1. Some samples for tumor detection on the test dataset using optimal parameters are shown in Fig. 3.13. The figure shows the results of tumor detection in 2.5D data blocks using the ABUSDet model.



(a) Different aggregation parameters “eps”



(b) Different CS (Channel Squeeze)

Figure 3.12 : Comparison of different aggregation parameters “eps” (a) and different CS (Channel Squeeze) methods in (b).

Here, the images in the first row are samples with accurate tumor detection. Looking at the slices in the second row, the nipple was wrongly recognized as a

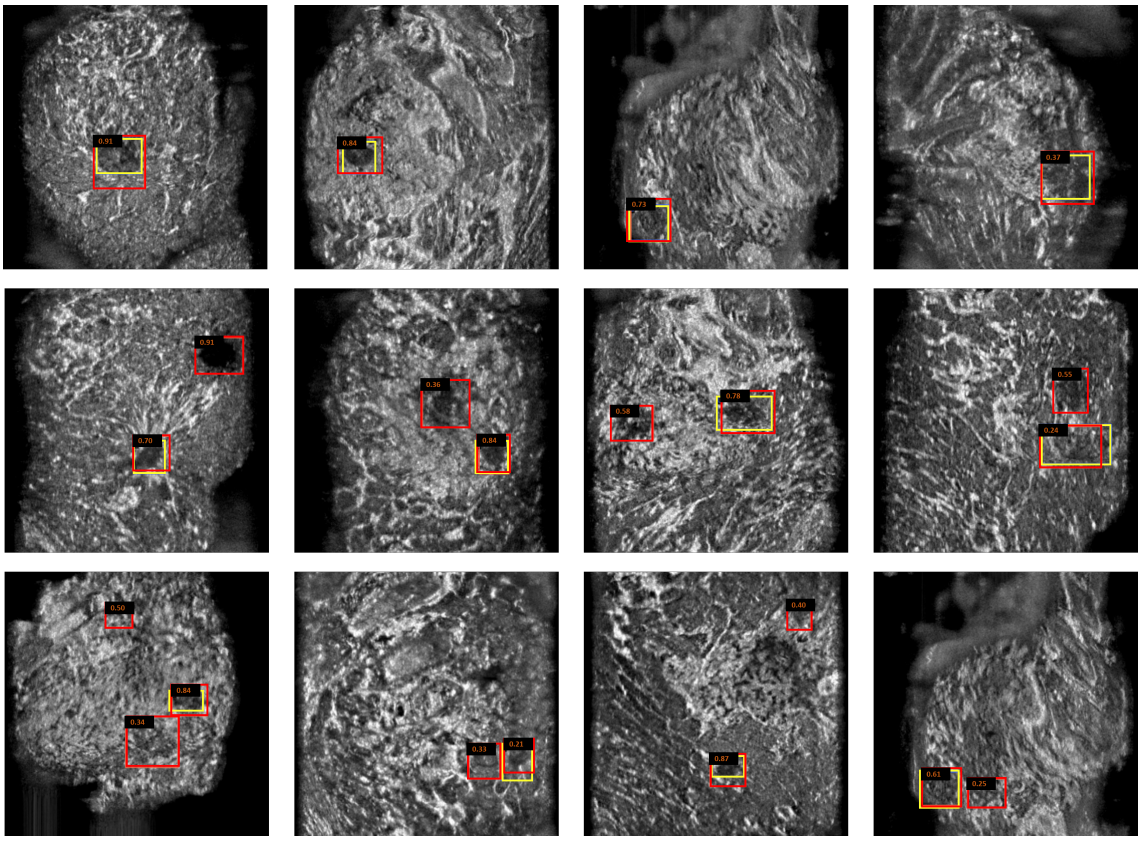


Figure 3.13 : Selected examples of tumor detection results on 2.5D data block of the Test set using the proposed ABUSDet model. The yellow boxes are the Ground Truth tumors, and the red boxes are the predicted boxes of tumor candidates.

tumor, while ground truth tumors were detected. The ABUS image of the nipple is easily confused with the tumor, even if recognized by the naked eye. Therefore, if the model uses insufficient data to train, it is easy to produce a high number of FPs. The samples in the third row show the detected actual tumors and some non-nipple FPs.

Stratified Performance Analysis by Tumor Size

To comprehensively evaluate the detection capability of the proposed ABUSDet model across different tumor dimensions, a stratified performance analysis was conducted on the test set. Based on the maximum side length, the tumors were categorized into three groups: Small (< 15 mm), Medium ($15 - 25$ mm), and Large

(> 25 mm). Table 3.3 summarizes the sensitivity at various FPs/volume thresholds for these three categories, as well as the total number of tumors in each group.

Table 3.3 : Stratified performance analysis of ABUSDet with respect to different tumor sizes.

FPs/volume	Sensitivity (%)		
	Small (< 15 mm)	Medium ($15 - 25$ mm)	Large (> 25 mm)
0.5	32.35	55.36	43.33
1.0	41.18	78.57	60.00
1.5	47.06	80.36	83.33
2.0	47.06	82.14	86.67
2.5	47.06	85.71	90.00
3.0	50.00	87.50	90.00
3.5	50.00	87.50	90.00
4.0	50.00	87.50	90.00
Total number	34	56	30

As shown in Table 3.3, ABUSDet demonstrates excellent and robust performance in detecting Medium and Large tumors, achieving sensitivities of 87.50% and 90.00% respectively at 3.0 FPs/volume. The detection of Small tumors (< 15 mm) remains a relatively challenging task, yielding a sensitivity of 50.00% at 3.0 FPs/volume. This performance discrepancy is primarily due to the subtle morphological features and low contrast of small tumors in ABUS images, making them easily confused with normal fibroglandular tissues or acoustic shadowing. Nevertheless, the high detection rate for Medium and Large tumors confirms the significant clinical value of the proposed model in capturing clinically actionable lesions.

Evaluation of Model Interpretability

In medical image analysis, model interpretability is paramount for clinical adoption. The interpretability of the proposed ABUSDet framework can be evaluated

from two primary perspectives: structural mechanism interpretability and spatial outcome interpretability.

Structural Mechanism Interpretability: The architecture of ABUSDet is inherently designed to mimic the clinical diagnostic workflow of radiologists. In clinical practice, physicians rarely diagnose a tumor based on a single 2D slice or a whole 3D volume at once; instead, they scroll through adjacent slices to observe the continuity and morphological changes of a lesion. The proposed 2.5D data blocks (specifically configured with 28 channels) emulate this cognitive process by feeding a continuous sequence of slices into the model. Furthermore, the novel *Channel Squeeze* module acts as a learned projection mechanism. Rather than operating as a black-box 3D convolution, it explicitly compresses the 3D depth information (28 channels) into a 3-channel feature map, dynamically extracting the most salient volumetric variations (e.g., depth-wise acoustic shadowing and boundary continuity) into a format interpretable by the 2D Faster R-CNN backbone.

Spatial Outcome Interpretability: Beyond structural design, the model provides explicit spatial interpretability through its bounding box outputs and multi-slice aggregation. The aggregation module leverages the DBSCAN algorithm to cluster 2D predictions into 3D coordinates, allowing physicians to trace the exact slices and spatial regions that triggered a detection. The qualitative visualization presented in Fig. 3.13 serves as a direct evaluation of the model's behavior. By examining the true-positive cases (the first row of Fig. 3.13), it is evident that the network has successfully learned to focus on clinically relevant morphological patterns, such as irregular boundaries and hypoechogenicity. Moreover, analyzing the false-positive cases reveals the model's decision-making logic. For instance, the nipple region (shown in the second row of Fig. 3.13) was mistakenly recognized as a tumor because its geometric shape and posterior acoustic shadowing share striking sonographic similarities with actual tumors. This specific error pattern paradoxically

validates the model's interpretability: it proves that the network makes decisions based on learned structural anatomy rather than random noise. By highlighting these regions, clinicians can easily understand the source of the model's confusion and confidently dismiss such false positives based on anatomical context.

Experiments Comparing Different Faster R-CNN Models

Fig. 3.14 shows the results of the Faster R-CNN model with different layer numbers. The experimental results show that the 101-layer model is not as good as the 50-layer model. Theoretically, the 101-layer model needs more data for training. The 50-layer Faster R-CNN achieves the best results with the same training data and rounds.

Experiment Comparing Different Methods

Table 3.4 : Sensitivity and FPs/volumes of different models.

FPs/volume	Sensitivity (%)		
	2D Faster R-CNN	3D Tumor Detection [20]	ABUSDet
0.5	44.02	5.45	**54.31**
1.0	53.73	6.67	**65.19**
1.5	57.74	10.39	**73.12**
2.0	63.04	12.58	**75.49**
2.5	64.03	13.95	**77.45**
3.0	66.63	14.58	**77.94**
3.5	68.75	14.58	**77.94**
4.0	70.41	14.58	**77.94**
Execution time per volume	**8.5s**	67.0s	11.0s

In order to compare the performance of 2D, 2.5D, and 3D object detection methods on ABUS data, this study implements a traditional Faster R-CNN-based 2D object detection method and replicates the 3D tumor detection method [20] on the data set. The training and clustering parameters of the 2D model are consistent

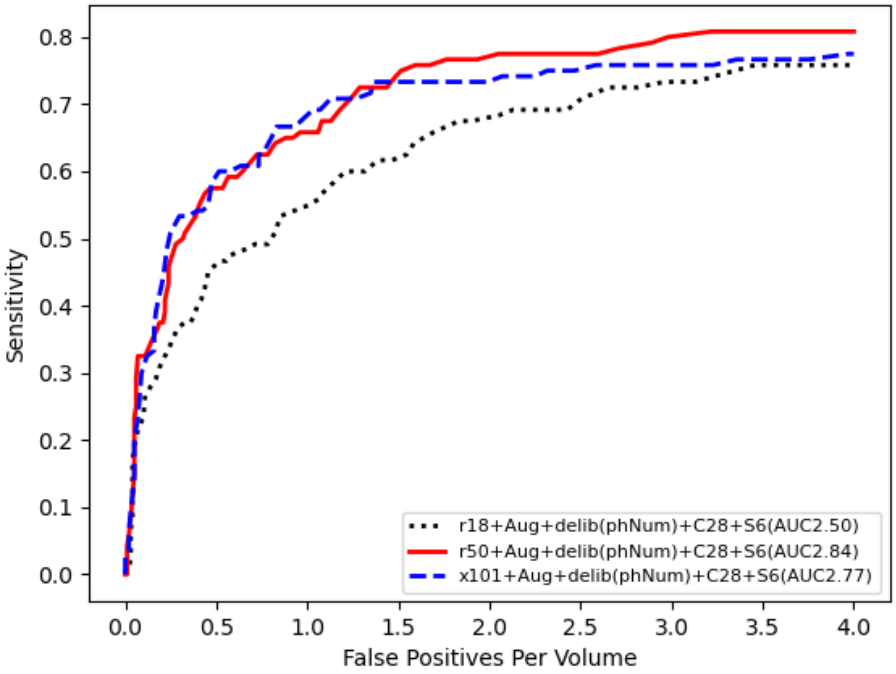


Figure 3.14 : Compare different Faster R-CNN models.

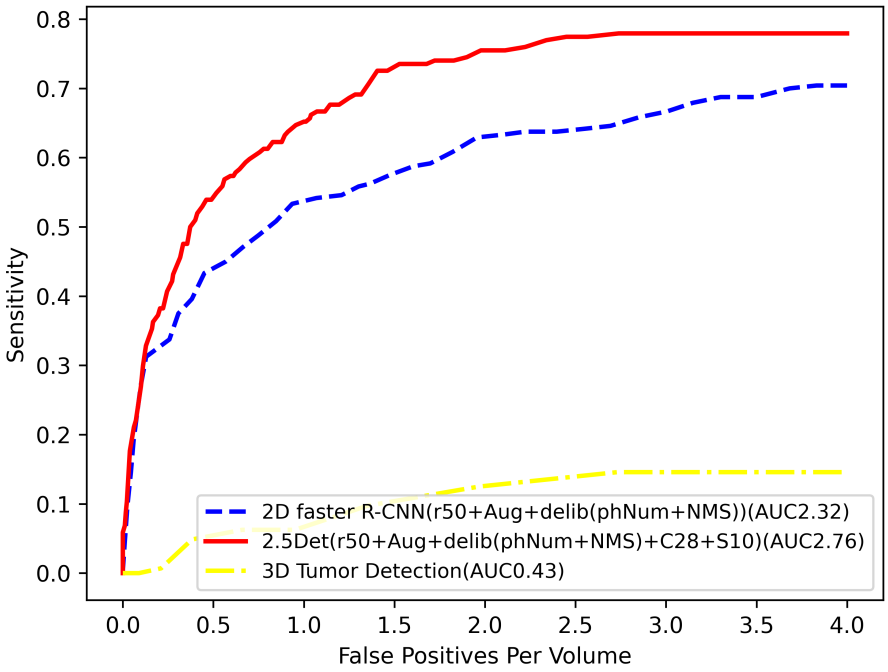


Figure 3.15 : Comparison among 2D, 3D, and 2.5D methods

with the proposed ABUSDet model. At the same time, the evaluation parameters are the same among these methods. In this experiment, a 2-fold cross-validation is

adopted. In addition to the original allocation method, the second method is the exchange of the training set and the test set. In Fig. 3.15, each curve represents the merged results of two cross-validations. The 2.5D tumor detection obtains the highest sensitivity, while the results of 3D tumor detection are not satisfactory. The reason is that under insufficient training data, the 2D object detection and 2.5D tumor detection methods apply transfer learning instead of 3D methods. In addition, the model and training parameters proposed in the paper [20] are not universal. It may work well in one dataset but not in others.

Table 3.4 lists the sensitivity and execution time of different methods. The execution time of 2.5D tumor detection and 2D Faster R-CNN is significantly shorter than that of 3D approaches.

3.6 Summary

This chapter has proposed a novel 2.5D tumor detection model ABUSDet, which effectively detects tumors from 3D volume ABUS data without involving high computation costs. By integrating a Channel Squeeze module into the 2D object detector module, ABUSDet can deal with multi-channel data blocks and utilize the trained weights from other 2D object detection models. Finally, in the aggregation module, the 3D coordinates of each tumor candidate are recalculated according to its 2D coordinates and depth information, and then the 3D coordinates are clustered by the DBSCAN algorithm. In addition, this study has devised a deliberate training mechanism in the training process. The experimental results demonstrate that the proposed 2.5D tumor detection with multiple channels is superior to 2D and 3D object detection methods. The results of the deliberate training mechanism outperform the traditional training process.

In summary, the proposed tumor detection model, ABUSDet, is a reliable and efficient assistant tool for reviewing breast tumors in ABUS volume. In the future,

a transformer-based detection approach will be adopted as the primary detection structure; Additionally, unsupervised learning will be introduced for feature extraction, enabling large-scale unlabeled data to contribute to the detection performance.

Chapter 4

TriNet: Multi-Stage Refinement Networks for Mammographic Breast Tumor Segmentation with Uncertainty Modeling

Accurate tumor detection and delineation in mammographic images is crucial for breast cancer screening, directly impacting patient diagnosis and treatment decisions. Therefore, complementing Chapter 3's volumetric detection, this chapter further expands to breast tumor segmentation and introduces TriNet, a triangle-shaped progressive refinement network with a shared MobileViT encoder and lightweight row–column attention for boundary-faithful mammographic tumor masks. It achieves state-of-the-art results on three widely-used benchmark datasets, completing the detection–segmentation pipeline ahead of the concluding synthesis in Chapter 5.

4.1 Introduction

Problem statement and gaps. Deep learning has substantially advanced mammographic tumor segmentation, with encoder-decoder architectures such as U-Net outperforming traditional pipelines [14]. Nevertheless, modern approaches (and their incremental variants) still face three fundamental limitations that are most salient in clinical scenarios: (i) Boundary precision: prevailing architectures struggle to recover crisp boundaries in low-contrast regions, where errors near tumor margins are clinically consequential; (ii) Global context and scale adaptability: canonical U-Net-style designs are CNN-centric and thus biased toward local receptive fields, making it difficult to capture long-range dependencies and to adapt to extreme tumor-size variations; (iii) Robust prior utilization: intermediate predictions or ex-

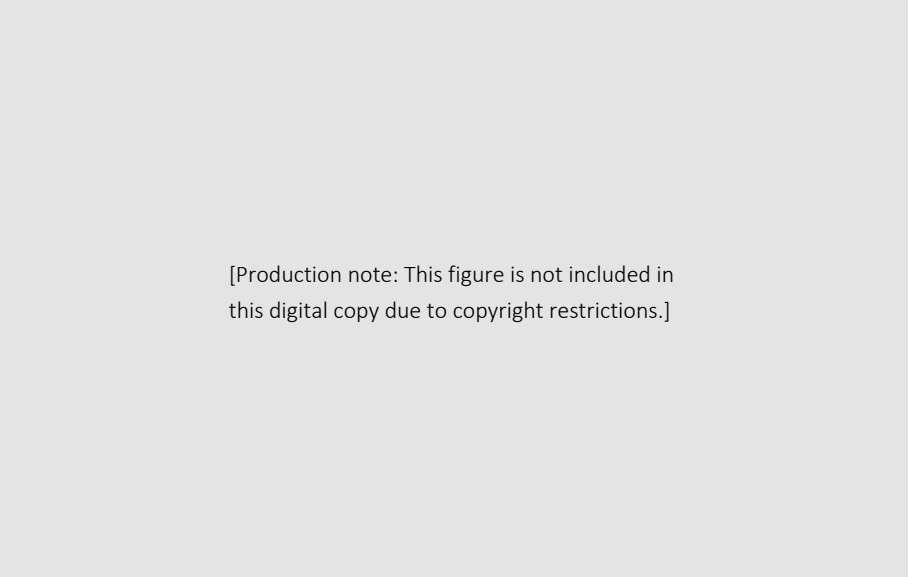


Figure 4.1 : Examples of mammographic images showing low tumor-tissue contrast, ambiguous boundaries, and substantial size variability from the INbreast dataset [2]. Left: a small, low-contrast tumor with indistinct margins. Right: two tumors of different sizes within the same mammogram. Blue contours indicate the ground-truth tumor boundaries.

ternal priors are often used heuristically and treated as deterministic, so errors tend to propagate across stages without explicit uncertainty handling.

Segmenting tumor pixels precisely from mammographic images is challenging due to low tumor-tissue contrast, ambiguous boundaries, and substantial size variability. Representative examples are shown in Fig. 4.1. These characteristics require architectures that preserve fine details while integrating large-scale context.

Recent advances and their limitations. Contemporary research has targeted these issues from three complementary directions, but important gaps remain. First, multi-stage coarse-to-fine frameworks (*e.g.*, Connected-UNets and Connected-SegNets) refine predictions by concatenating multiple full encoder-decoder networks, demonstrating the benefits of progressive refinement [22]. However, direct network concatenation introduces substantial computational redundancy, approximately doubling parameters and inference time, which hinders deployment in time- and resource-

constrained clinical environments. Second, hybrid CNN-Transformer architectures [29] enrich global context modeling and long-range dependency capture, while attention mechanisms [141] can highlight tumor-relevant regions and multi-task formulations [142] can couple localization with classification. Yet, these hybrids are frequently heavy in both parameters and FLOPs; efficiency remains a central bottleneck. Third, prior- and uncertainty-aware formulations have emerged. Still, most methods either discard intermediate masks instead of iteratively refining them or model output uncertainty without explicitly accounting for the imperfection of priors themselves. As a result, the integration of imperfect priors (*e.g.*, coarse masks) can still be fragile and prone to error accumulation across stages.

The proposed approach in a nutshell. This study introduces **TriNet**, a progressive, efficiency-oriented refinement network tailored to mammographic tumor segmentation. TriNet employs a single MobileViT encoder (Stage 1, S1) and a configurable stack of sequential decoder stages starting at S2. By default, TriNet adopts a three-stage configuration (S1–S3), with an optional fourth refinement stage (S4) also evaluated in the experiments. All decoders reuse a **unified Decoder** primitive that (a) fuses same-level encoder skips with top–down features, (b) applies a lightweight **ConvBlock** (pointwise–depthwise–pointwise with ECA), and (c) appends a width-preserving **Lightweight Row-Column Attention (LRCA)** unit after the ConvBlock. LRCA factorizes 2D attention into row- and column-wise components, computes them in a single Scaled Dot-Product Attention (SDPA) call, and maps tokens back via unpatchify, thus supplying global shape cues at every decoding level while reducing complexity from $\mathcal{O}(L^4 d)$ to $\mathcal{O}(L^3 d)$ without altering the decoder topology.

Each decoder integrates two synergistic pathways: (1) a CNN-based stream for local feature extraction and alignment; and (2) the proposed **LRCA** module, which

provides holistic shape reasoning as a width-preserving residual unit. To robustly leverage intermediate predictions, refinement stage(s) (*i.e.*, $n \geq 3$) are prior-guided: the stage-$(n-1)$ mask at level L_0 (*e.g.*, S_2L_0 when refining with S3) is first perturbed by Gray–Scott reaction–diffusion noise and then selectively injected at intermediate resolutions (L_1/L_2) of the current stage, where it is fused with the convolutional stream and subsequently refined by LRCA.

Contributions. The contributions made in this chapter are:

- **TriNet architecture:** This chapter designs a compact coarse-to-fine framework that couples a single shared encoder (S1) with a configurable number of sequential decoders starting at S2 (this study's default setting uses two decoders, *i.e.*, S2 and S3). All stages are built from the same Decoder primitive that fuses same-level skips with top–down features and keeps the decoding interface consistent across scales. This avoids re-encoding costs in multi-stage refinement (as in concatenated U-shaped networks), reduces parameters and latency, and focuses capacity where detail recovery actually happens—in the decoder.

- **LRCA, decoder-side lightweight global context:** To complement convolutional edges/textures with holistic shape cues, this chapter devises a width-preserving **Lightweight Row-Column Attention (LRCA)** module after each ConvBlock. LRCA patchifies feature maps into tokens, performs row- and column-wise self-attention in parallel via a single Scaled Dot-Product Attention (SDPA) call, then unpatchifies back to pixels as a residual unit. This factorization reduces attention cost from $\mathcal{O}(L^4 d)$ to $\mathcal{O}(L^3 d)$.

- **Uncertainty-aware prior integration:** Progressive stages reuse the previous stage's L_0 mask as spatial prior, but, instead of treating it as determin-

istic, this chapter perturbs it using Gray–Scott reaction–diffusion to emulate realistic boundary variability. The perturbed prior is selectively injected at intermediate resolutions (L_1/L_2) of the refinement stage(s)—where spatial detail is forming yet features remain stable—then fused with the convolutional stream and refined by LRCA within the same Decoder. Through modeling uncertainty, this design improves robustness to imperfect intermediate predictions and curbs error propagation without incurring the overhead of a second encoder.

- **Adaptive boundary supervision:** Beyond region overlap, **ABALoss** is proposed to align boundaries in a scale-aware and probabilistic manner. For each instance, a scale-adaptive morphological gradient yields thin edges for small tumors and thicker contours for large ones; alignment is computed on $\sigma(\hat{y})$ to remain stable gradients near decision thresholds. Training is staged: S2 optimizes DiceBCE and MissPenalty losses to prioritize recall, whereas S3 (and S4, when present) and the fused output adopt DiceBCE and ABALoss losses to consolidate fine boundaries across extreme size variations.

Extensive experiments on three benchmark datasets show that TriNet advances mammographic tumor segmentation while retaining clinical efficiency, achieving state-of-the-art performance across multiple evaluation metrics.

Next, a thorough review of existing methodologies that inform the proposed TriNet framework is presented. This is followed by a detailed exposition of TriNet for mammographic tumor segmentation, accompanied by extensive experiments conducted on three benchmark datasets and evaluated against state-of-the-art approaches.

4.2 Related Work

Breast tumor segmentation in mammography has witnessed significant advances through deep learning approaches, yet challenges persist in accurately delineating tumor boundaries and handling large inter-tumor size variation. This section provides a comprehensive review of existing methodologies that inform this study's proposed TriNet framework. This study organizes the existing literature into five key areas: foundational architectures for medical image segmentation, multi-stage approaches, attention mechanisms, prior knowledge integration with uncertainty modeling, and boundary-aware methods. Through this systematic review, we identify the limitations of existing approaches and establish the theoretical foundation for this study's multi-stage architecture.

4.2.1 Medical Image Segmentation Architectures

The encoder–decoder (U-shaped) paradigm has been the dominant choice for medical image segmentation since U-Net [14]. This seminal architecture established the importance of skip connections for preserving fine-grained spatial information during downsampling and upsampling, and it has achieved strong performance across diverse medical imaging tasks.

Building on this foundation, U-Net++ [25] introduces nested and dense skip connections that narrow the semantic gap between encoder and decoder features, yielding gains especially for objects with varying sizes—a crucial consideration in mammographic tumor segmentation. For whole-mammogram settings, AUNet [143] further demonstrates that attention-guided dense upsampling within an asymmetrical encoder–decoder can better handle the wide variety of mass shapes and sizes. Complementing these designs, ARF-Net allocates scale-appropriate receptive fields via a Selective Receptive Field Module (SRFM) that combines multi-branch receptive fields with a learned selection mechanism, improving sensitivity to small tumors

without abandoning the encoder–decoder prior [144].

Beyond pure CNN-based U-shaped variants, hybrid architectures combine Convolutional Neural Networks (CNNs) with Vision Transformers to capture long-range dependencies while retaining high-resolution detail. TransFuse [28] exemplifies this trend with parallel CNN and Transformer branches fused by a BiFusion module employing self-attention to selectively integrate multi-level features, achieving state-of-the-art performance with competitive efficiency. In a related vein, MCTrans [145] employs proxy embeddings to enhance cross-scale interactions for multi-organ segmentation, offering insights into effective feature fusion strategies that inform multi-stage decoder design.

Taken together, these developments point to a key principle: effective medical image segmentation must balance global context with precise local detail. Pure CNNs excel at edges and textures but may miss long-range structure, whereas Transformer-based components capture global relationships but can struggle with boundary precision. This insight motivates this study's TriNet design, which aims to leverage both paradigms within a unified, decoder-centric framework.

Outside the U-shaped families, general-purpose baselines such as DeepLab [146, 147, 148], PSPNet [149], and HRNet [150] have been adapted to medical images via multi-scale context aggregation (ASPP/pyramid pooling) and high-resolution feature maintenance. While strong on heterogeneous scenes, their reliance on coarse output strides or heavy multi-branch context can underperform on faint, small tumors and razor-thin boundaries in full-field mammograms, where preserving fine spatial detail across scales is critical. These observations further motivate decoder-side refinements tailored to mammography.

4.2.2 Multi-Stage Segmentation Frameworks

Multi-stage segmentation has proved effective for breast mass analysis because it mirrors how clinicians work: obtain a reliable coarse localization and then refine boundaries with increasing detail. In mammography, this strategy is often instantiated by connecting two segmentation networks so that information from the first stage guides the second. Representative examples are Connected-UNets [22] and Connected-SegNets [23], which link two full encoder–decoder models through inter-network connections to improve mask quality on breast datasets. While these designs validate progressive refinement, they duplicate both the encoder and the decoder, resulting in substantial parameter growth and latency—an obstacle for practical deployment. More recently, a presegmenter-cascaded framework generates a saliency prior in Stage 1 and injects presegmenter attention in Stage 2, reducing false negatives across datasets while remaining backbone-agnostic [24].

Motivated by these observations, the proposed architecture follows a different path: we keep a single shared encoder and organize the decoders across stages into one integrated, gradually shallower stack. Conceptually, the overall shape is closer to an inverted triangle than a repeated U-Net: the Stage 2 decoder directly reuses and connects to the shared encoder feature pyramid from Stage 1, avoiding another round of feature extraction while preserving the benefits of stage-wise refinement. Evidence that multi-stage or multi-path decoders strengthen feature fusion and reuse includes DoubleU-Net—running a second U-Net improves detail recovery but pays the cost of a second encoder [151]—and LadderNet, where U-Net-style multi-path decoders enhance information flow through repeated fusion while reusing parameters [152]. This study's design inherits the refinement power of these lines of work but removes the redundant encoder, focusing capacity where it matters—the decoding pathway.

Decoder-side refinement is further supported by advances in generic semantic

segmentation. RefineNet exemplifies a principled approach to multi-path refinement at high resolution, repeatedly fusing semantic and fine details to recover sharp object boundaries [153]. Following this principle, we place this study's lightweight row-column attention (LRCA) inside the decoder to inject global shape context alongside local convolutional cues. The goal is not to replace convolutions, but to complement them: convolutions excel at edges and textures, whereas LRCA provides long-range structure awareness within the same decoding stream.

From a training perspective, multi-stage decoders naturally support deep supervision. Deeply-Supervised Nets established that supervising intermediate outputs can stabilize optimization and improve generalization [154]. This practice has been widely adopted in medical segmentation: UNet++ and U-Net 3+ attach supervision to multiple decoder outputs to better exploit multi-scale information [25, 26]. TriNet follows this paradigm: each stage produces a single-channel prediction that is directly supervised, aligning learning signals with the intended coarse-to-fine behavior and improving boundary quality in later stages.

Finally, broader evidence supports sharing one backbone while using multiple decoding paths. Multi-decoder U-Net variants use a shared encoder with distinct decoders to model uncertainty or alternative refinements without replicating the feature extractor [155]. Related ideas in multitask pixel prediction also split or share decoders to efficiently allocate capacity across outputs [156]. TriNet adopts a sequential version of this principle: Stage 2 and Stage 3 decoders are tightly coupled, and the second stage is guided by a perturbed prior from the first, achieving progressive correction with a compact, deployment-friendly footprint.

4.2.3 Attention Mechanisms in Medical Image Segmentation

CNN-based Attention Modules

Attention in convolutional networks has proven effective for mammographic segmentation, where tumors can be small, low-contrast, and easily confounded by dense tissue. Channel attention is a natural starting point: Squeeze-and-Excitation (SE) [135] adaptively rescales channels to emphasize informative responses while keeping computational cost modest. Spatial attention tailored to medical images was explored by Attention U-Net [95], which introduced learnable gates in skip connections to suppress irrelevant regions and strengthen localization of target structures.

A complementary line of work pursues lightweight designs that preserve accuracy while minimizing overhead. ECA-Net [157] replaces heavy channel mixing with a local 1D convolution over the channel dimension, avoiding dimensionality reduction and maintaining efficiency. FcaNet [158] interprets channel attention through a frequency lens, enriching channel descriptors without large parameter growth. Guided by these trends, this study's decoder's ConvBlock optionally includes an ECA unit (pointwise–depthwise–pointwise with ECA) to improve tumor–background contrast at negligible cost. Overall, lightweight channel/spatial attention provides a practical means to enhance contrast in low-visibility tumors without incurring notable computational burden.

Transformer-based Self-Attention

Self-attention entered vision by viewing an image as a sequence of patches. ViT [97] operationalized this idea and demonstrated that long-range interactions improve recognition. To contain quadratic complexity and reintroduce locality, Swin Transformer [159] organized tokens into hierarchical maps with shifted windows, later proving powerful for dense prediction. For semantic segmentation, SETR [160]

reframed the task as sequence-to-sequence learning, confirming that Transformer encoders can drive pixel-level predictions when paired with appropriate upsampling. These developments established key ingredients subsequently adapted to medical images: tokenized features, windowed or axial attention for efficiency, and hierarchical decoding for resolution recovery.

Within U-shaped architectures for medical segmentation, attention has been injected at different locations. TransUNet [27] fuses Transformer features with high-resolution decoding to provide global context for fine structures; Swin-Unet [16] extends this to a largely Transformer-based U-shape with windowed attention in both encoder and decoder; and hybrid models such as TransFuse [28] fuse CNN and Transformer streams before decoding. Several works place attention explicitly inside the decoder: UTNet [161] integrates efficient self-attention at multiple upsampling levels to capture cross-scale dependencies with limited overhead; MedT [162] applies gated axial attention by decomposing 2D interactions along rows and columns, well suited to medical textures; and MISSFormer [163] strengthens the segmentation Transformer block with improved multi-scale interaction and feed-forward design, leading to better boundaries. Collectively, these studies indicate that decoder-side attention can add global shape cues precisely where details are reconstructed.

For whole-mammogram settings, recent transformerized decoders further systematize decoder-side attention. TrEnD [29] introduces a superpixel-guided adaptive patch embedding to align tokenization with anatomy and an attention-gated decoder coupled to dual global–local branches, jointly suppressing irrelevant background while preserving fine structure. Complementarily, SCSGNet [164] combines a coarse global guidance map with series–parallel fusion, dynamic long-range correlation capture, and a triplet-attention update to iteratively refine ambiguous mass boundaries.

Domain evidence corroborates the value of long-range context for tumor segmentation, where shapes vary widely and may be partially obscured. In brain tumor benchmarks, TransBTS [165] and Swin-UNETR [166] show that attention helps separate subregions with overlapping intensities. In breast imaging, studies report consistent gains from attention modules: Swin-Net [167] augments U-Net with Swin blocks for ultrasound breast tumor segmentation; HAU-Net [168] employs hybrid attention to improve mass boundaries in challenging ultrasound scenes; and CSwin-PNet [169] and BTS-ST [170] explore hierarchical/windowed attention for breast tumor analysis and multimodal segmentation–classification pipelines. Axial-transformer [171] variants have also been explored for mammography. While encoder- or bottleneck-centric placements are common, decoder-side attention across multiple upsampling levels has been comparatively less emphasized and merits further study in full-field settings.

A parallel literature targets efficiency for mobile or real-time use. MobileViT [172] and MobileViTv2 [173] show that small attention blocks can coexist with depth-wise separable convolutions, delivering strong accuracy–efficiency trade-offs. EfficientViT [174], TinyViT [175], and FastViT [176] further reduce cost through windowing or linearized interactions, distillation, and structural re-parameterization, while remaining transferable to segmentation. These trends motivate decoder designs that add just enough global context without changing feature shapes or incurring large memory or latency.

From ViT, Swin, and SETR to TransUNet, Swin-Unet, and axial or efficient variants, the literature consistently shows that attention complements convolution by supplying long-range information and improving boundaries. Positioned in this landscape, a decoder-side, lightweight, and shape-preserving self-attention is a natural design choice to inject global cues precisely where details are reconstructed. This study follows this pattern and provides the specific variant and placement rationale

in Section 4.3.

4.2.4 Integrating Prior Knowledge and Uncertainty Modeling

Prior knowledge integration has emerged as a fundamental strategy for enhancing segmentation accuracy in mammography, where detection networks, coarse segmentation stages, or radiologist annotations provide valuable spatial guidance. However, existing approaches to prior integration commonly treat these priors as reliable inputs without adequately addressing their inherent uncertainties. In practice, such priors frequently exhibit imperfections—*e.g.*, inaccurate tumor boundaries, missed small tumors, and false positives—yet current pipelines rarely model or mitigate these issues.

The predominant approaches employ deterministic frameworks that propagate errors from imperfect priors directly into final segmentation results. While the Probabilistic U-Net [177] represents an advancement by modeling distributions over segmentation hypotheses, it primarily targets predictive uncertainty rather than explicitly modeling the uncertainty embedded in the prior itself; moreover, its dedicated Prior and Posterior networks introduce substantial architectural overhead and training complexity. This omission is particularly problematic in multi-stage pipelines, where intermediate outputs serve as priors for subsequent stages: any uncertainty or error compounds across stages and degrades the final mask quality.

A representative example of deterministic prior use is the presegmenter-cascaded framework, which first produces a saliency prior and then guides a refinement network via presegmenter attention to focus on informative regions [24]. Yet the saliency prior is consumed as a fixed guidance signal—its uncertainty is neither estimated nor propagated—so mislocalized or overconfident priors can still bias downstream refinement.

Consequently, the absence of explicit modeling for prior uncertainty remains

a significant gap for mammography. Real-world priors are inherently imperfect (coarse segmentations with rough boundaries, detector outputs with false positives, inter-observer variability in annotations); without mechanisms to account for these uncertainties, models remain vulnerable to prior-induced errors, limiting clinical applicability.

In contrast to treating priors as error-free guides, we explicitly consider the uncertainty of intermediate priors within a multi-stage scheme. Concretely, we perturb coarse priors during training to expose the model to realistic boundary variability, and we integrate them selectively in the decoding process; implementation details and variants are deferred to Section 4.3.

4.2.5 Boundary-Aware and Edge-Enhanced Segmentation

Precise boundary delineation is critical for breast tumor segmentation, directly impacting clinical decision-making and treatment planning. Recent advances, therefore, incorporate explicit edge cues into the segmentation process to address ambiguous tumor margins in mammographic images.

A first line of work strengthens boundary modeling within the network architecture. Wang et al. [33] introduced Boundary-Aware Transformers that inject boundary-wise attention gates into transformer blocks; by leveraging boundary priors to guide spatial attention, their method preserves object structure and boundary detail, illustrating the value of combining global context modeling with boundary enhancement. Decoder-centric refinements in improved U-Net 3+ variants likewise add edge-aware modules and hybrid losses (*e.g.*, Dice+BCE) to alleviate class imbalance while retaining multi-scale detail in full-field settings [34]. Complementing these architectural ideas, [35] proposed density- and size-aware losses (Adaptive Sample-Level Prioritizing), dynamically reweighting pixels based on mass size and breast density to reflect the varying difficulty of boundary delineation across mam-

mographic presentations.

Beyond architectural changes and sample-level reweighting, boundary- and topology-aware objectives explicitly emphasize contours and shape and are effective complements to region terms. Representative examples include Boundary Loss based on distance transforms [30], Hausdorff-derived objectives [31], and topology-preserving clDice [32]. Such objectives improve boundary faithfulness and help mitigate class imbalance; in this study, we likewise combine region criteria with boundary-aware supervision tailored to extreme tumor-size variation.

Despite these advances, limitations remain. Most boundary-aware methods employ fixed-scale edge operators that cannot adapt to the extreme size variations observed in breast tumors. Moreover, few approaches explicitly account for the uncertainty inherent in boundary delineation, often treating all edges with equal confidence regardless of image quality or tumor characteristics. A broader literature estimates predictive uncertainty via Monte Carlo dropout [178], aleatoric/epistemic decomposition [179], and evidential/Dirichlet formulations [180], but these methods primarily target uncertainty in the final prediction rather than the reliability of boundary cues themselves. In mammography pipelines—especially multi-stage ones where intermediate outputs serve as priors—this distinction is crucial: inaccuracies and overconfidence in intermediate boundaries can bias subsequent refinement if not modeled.

These observations motivate an adaptive, uncertainty-aware treatment of edges. In the proposed framework, we develop an adaptive edge-aware loss that scales with tumor size and operates on probabilistic predictions to accommodate boundary ambiguity. By coupling such supervision with decoder-centric refinement, we aim to sharpen contours across a wide size spectrum while remaining robust to uncertainty in boundary evidence.

4.2.6 Summary and Motivation for TriNet

Prior work validates progressive refinement and attention mechanisms, yet three gaps persist for mammography: (i) concatenating full encoder–decoder networks duplicates compute; (ii) priors are consumed as if they were perfect, letting errors propagate; and (iii) boundary supervision is often fixed-scale and brittle under extreme tumor-size variation. TriNet addresses these gaps with design choices tied directly to these failure modes:

- **Efficient multi-stage architecture.** Instead of chaining two complete U-shaped models, we share one MobileViT encoder (S1) and deploy a configurable stack of sequential decoders starting at S2. This study's default configuration uses two refinement decoders (S2/S3), with an optional fourth stage (S4) evaluated in experiments. Each stage uses the same unified Decoder primitive, so refinement happens where details are reconstructed and multi-scale skips can be reused without re-encoding. This preserves the benefits of coarse-to-fine correction seen in Connected-UNets, SegNets, and Presegmenter cascades [24] while cutting parameter and latency overhead, yielding a deployment-friendly footprint.

- **LRCA (decoder-side lightweight global context).** Global context is most needed when upsampling restores details. LRCA therefore sits inside the decoder as a width-preserving residual after the ConvBlock, following a patchify & tokenize → parallel row/column self-attentions in one SDPA → unpatchify pipeline. This factorization reduces attention complexity from $\mathcal{O}(L^4 d)$ to $\mathcal{O}(L^3 d)$ without altering the decoder topology, making it practical for every decoding level under tight compute budgets.

- **Uncertainty-aware prior integration.** Intermediate masks are informative but imperfect. This study explicitly models prior uncertainty by evolving

Gray–Scott fields on the previous stage's L_0 mask to generate plausible boundary perturbations. The perturbed prior is then selectively injected at intermediate resolutions (L_1/L_2) of the current refinement stage (*i.e.*, any stage $n \geq 3$, such as S3 or S4 when present)—levels that balance spatial fidelity with feature stability—merged with the convolutional stream, and refined by LRCA. This selective, uncertainty-aware injection improves resilience to coarse-stage errors without heavy Bayesian machinery.

- **Adaptive boundary supervision.** Fixed-scale edges under- or over-penalize tumors at the extremes of size. **ABALoss** computes a scale-adaptive morphological gradient on ground-truth and aligns it with probabilistic prediction edges $\sigma(\hat{y})$, producing informative, stable gradients near boundaries. Training is staged: S2 optimizes DiceBCE + MissPenalty to prioritize recall and avoid misses, whereas refinement stages (S3 and S4, when present) and the fused output adopt DiceBCE + ABALoss to consolidate fine boundaries across extreme size variation.

By synthesizing insights from multi-stage processing, LRCA, and uncertainty modeling, TriNet advances mammographic tumor segmentation while maintaining the computational efficiency required for clinical deployment.

4.3 Methodology

This section details the methodology of the proposed TriNet for mammographic tumor segmentation. It first presents the multi-stage architecture and its design rationale, followed by a description of the unified decoder—the sole decoding primitive shared across stages and levels—which illustrates how prior knowledge is injected and how the newly introduced Lightweight Row-Column Attention (LRCA) enhances the convolutional pathway. A subsequent subsection provides the algorith-

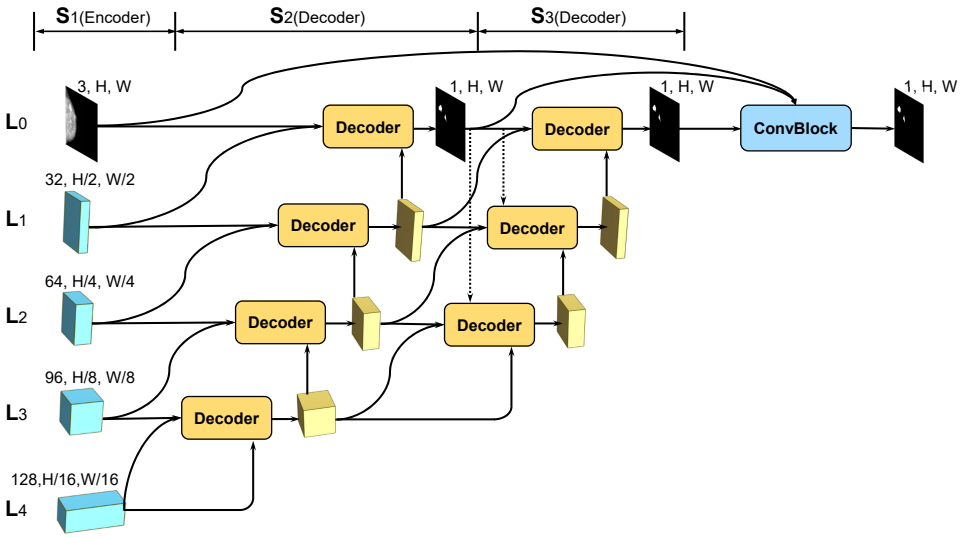


Figure 4.2 : Overall architecture of TriNet. The three stages progressively refine the mask: S1 (feature extraction), S2 (initial decoding), and S3 (prior-guided refinement). The 3D cuboids represent the output feature maps of the encoder and decoder; orange boxes denote decoder modules, and light-blue boxes denote ConvBlocks.

mic details of LRCA.

4.3.1 Overall Framework

Fig. 4.2 provides a high-level view of TriNet. The left column of stacked cuboids corresponds to the encoder output pyramid $\{L_0, \ldots, L_4\}$ produced by a single MobileViT backbone [172]. These multi-scale features span high-resolution, low-channel signals at L_0 to low-resolution, high-channel representations at L_4. The right side shows the level-wise outputs of the decoder stack. In the default TriNet configuration (3 stages in total), there are two decoder columns (S2 and S3). Orange blocks denote Decoder instances and light-blue blocks indicate the internal ConvBlock submodule (see Fig. 4.3).

Stage 1 (S1): Encoder. S1 extracts hierarchical, context-rich features by combining efficient convolutional operators with lightweight transformers. This hy-

brid design captures long-range tissue dependencies while preserving fine, edge-level structure that is crucial for precise tumor delineation.

Progressive refinement design. TriNet employs a gradually shallower decoder stack across stages: S2 uses four decoder blocks (processing $L_3 \rightarrow L_0$), while S3 uses three blocks (processing $L_2 \rightarrow L_0$), and an optional S4 would use two blocks. This design concentrates computational capacity in early stages where coarse localization benefits most from deep processing, while later refinement stages require fewer layers to polish boundaries given strong spatial priors from preceding stages. The result is an inverted-triangle topology that balances refinement quality with parameter efficiency.

Stage 2 (S2): Initial decoding. Conditioned on S1's pyramid, a sequence of Decoder instances generates level-wise representations $\{S_2L_m\}$ and an initial mask $\{S_2L_0\}$ as prior knowledge. S2 focuses on reliable localization and approximate boundaries, forming a strong starting point for refinement. During training, S2 is supervised with DiceBCE combined with MissPenalty to explicitly suppress false negatives and maximize recall (detailed in Sec. 4.3.5).

Prior-guided refinement. S3 reuses the same decoding mechanism to progressively refine the mask and serves as the prior-guided refinement stage in the three-stage TriNet; the design is readily extendable to additional refinement stages if desired. The stage-$n{-}1$ mask at level L_0 (*e.g.*, S_2L_0 when refining with S3) is first perturbed with Gray–Scott reaction–diffusion noise to model boundary uncertainty. The perturbed prior is then injected as spatial guidance into selected intermediate-resolution decoders of the current stage (only uses L_1 and L_2 in practice), where it is fused with the top-down feature and further refined by LRCA. In this multi-stage setting, S3 corrects coarse errors from S2 and consolidates fine boundaries, yielding

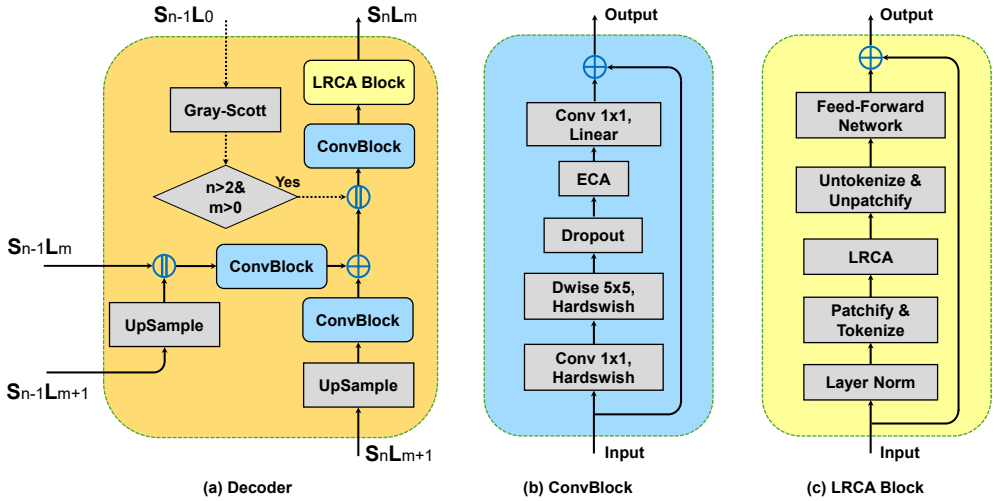


Figure 4.3 : Core building blocks: (a) **Decoder** with optional prior injection in refinement stage; (b) **ConvBlock** (pointwise–depthwise–pointwise with ECA); (c) **LRCA** placed after the ConvBlock.

a more robust prediction under imperfect priors. S3 and final output are supervised with DiceBCE combined with ABALoss to emphasize boundary fidelity (Sec. 4.3.5).

Unlike "connected" multi-stage networks (*e.g.*, Connected-UNets [22] and Connected-SegNets [23]) that replicate a full encoder-decoder per stage, TriNet employs one shared encoder (S1) followed by sequential decoders (S2, S3, and optionally S4 in extended configurations). This preserves the benefits of progressive refinement without re-encoding costs.

4.3.2 Decoder Module

As shown in Fig. 4.3a, each **Decoder** fuses a same-level encoder skip with a top–down feature from the higher level of the current stage. The backbone inside the block is a lightweight **ConvBlock** (pointwise–depthwise–pointwise with ECA) for strong local edge/texture cues. To complement this locality with holistic shape reasoning, the proposed **LRCA** (Fig. 4.3c) unit is placed after the ConvBlock at *selected* decoder levels. Concretely, for computational efficiency, we activate LRCA only in the level-1 and level-2 decoders of each stage (denoted S_nL_1 and S_nL_2); other

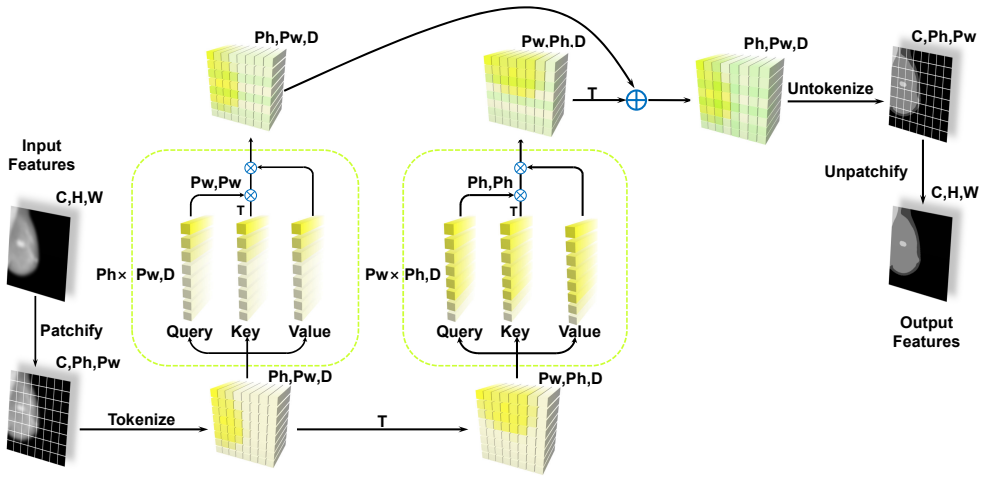


Figure 4.4 : **LRCA** pipeline. Patchify & Tokenize produces a square token grid; row-wise and column-wise self-attentions operate on axial sequence views; results are fused and restored by Untokenize & Unpatchify. The implementation computes the two axial attentions in a single SDPA call by concatenating the row/column batches.

levels use the ConvBlock without LRCA. In Stage 3, a prior mask can be injected at intermediate resolutions, but the decoding primitive itself is identical across stages. LRCA is configured with width-preserving in the implementation, so it serves as a drop-in residual unit without altering the decoder topology.

4.3.3 LRCA: Lightweight Row-Column Attention

Fig. 4.4 illustrates the design. Given a decoder feature map $x \in \mathbb{R}^{B\times C\times H\times W}$, this study first patchifies and tokenizes it into a square token grid, followed by a token-wise normalization. The token grid is then processed by two axial self-attention operators: one along rows and one along columns. These two directional interactions capture long-range dependencies that convolutions alone cannot. Their outputs are fused and mapped back to the pixel space by Untokenization & Unpatchification. A lightweight convolutional FFN refines the result. This design keeps the attention budget small while supplying global shape cues to the decoder.

Let t denote the patch size and $D = C\,t^2$ the token dimensionality. The patchify operator $\mathcal{P}_t$ yields

$$T \;=\; \mathcal{P}_t(x) \in \mathbb{R}^{B\times P_h\times P_w\times D}, \qquad P_h = \frac{H}{t},\; P_w = \frac{W}{t}. \tag{4.1}$$

In our implementation, images are resized to square inputs, and t is chosen to evenly divide the spatial size, so $H = W$, and we denote $L \;=\; \frac{H}{t} \;=\; \frac{W}{t}$, *i.e.*, a square $L \times L$ token grid.

A single linear map produces queries, keys and values per token:

$$[Q, K, V] = TW_{\text{qkv}} + \mathbf{1}b_{\text{qkv}}^{\top}, \qquad W_{\text{qkv}} \in \mathbb{R}^{D\times 3D}. \tag{4.2}$$

Row-wise and column-wise attentions are defined on axial sequence views of the same token grid. For each row index $r \in \{1, \ldots, L\}$ the method forms a length-L sequence and compute

$$Y_r^{\text{row}} = \text{Softmax}\Big(\tfrac{Q_r K_r^{\top}}{\sqrt{d}}\Big)\, V_r, \tag{4.3}$$

and for each column index $c \in \{1, \ldots, L\}$,

$$Y_c^{\text{col}} = \text{Softmax}\Big(\tfrac{Q_c K_c^{\top}}{\sqrt{d}}\Big)\, V_c, \tag{4.4}$$

where d is the per-head width after a standard multi-head reshaping. The LRCA output in token space is obtained by element-wise averaging the two axial attentions. This balances their contributions and keeps the activation scale comparable to a single branch:

$$Y_{\text{tok}} = \tfrac{1}{2}\Big(Y^{\text{row}} + Y^{\text{col}}\Big). \tag{4.5}$$

Eqs. (4.3)–(4.5) constitute the conceptual algorithm shown in Fig. 4.4: two directional self-attentions that decompose dense 2D interactions into row/column com-

ponents and then fuse them.

For efficiency, this method computes Eqs. (4.3) and (4.4) in parallel by concatenating the row and column sequence batches and invoking the fused scaled dot-product attention once. The result is then split and fused as in Eq. (4.5). This preserves the conceptual design while reducing kernel launches and memory traffic.

A linear projection maps Y_{tok} back to the required token dimensionality, followed by

$$y = \mathcal{P}_t^{-1}(Y_{\text{tok}} W_{\text{out}} + \mathbf{1} b_{\text{out}}^{\top}), \tag{4.6}$$

where $\mathcal{P}_t^{-1}$ is the inverse unpatchify. A residual connection to x and a lightweight convolutional FFN complete the block.

Practical configuration. LRCA exposes two hyperparameters: the number of attention **heads** and the **patch size** t. We fix the number of heads to 4 for all LRCA instances in TriNet. Across decoder stages, LRCA is enabled only at the intermediate resolutions: we set the patch size to t=2 for S_nL_1 (the level-1 decoder in stage n) and t=1 for S_nL_2; other decoder levels omit LRCA. This configuration offered a strong accuracy–efficiency trade-off in whole-image mammography.

Let $L = \frac{H}{t} = \frac{W}{t}$ and $N = L^2$ denote the tokens per side and the total number of tokens, respectively. Let $N = L^2$ be the number of tokens. A dense 2D attention costs $\mathcal{O}(N^2 d) = \mathcal{O}(L^4 d)$. LRCA requires

$$\mathcal{O}(L \cdot L^2 d) + \mathcal{O}(L \cdot L^2 d) = \mathcal{O}(L^3 d) = \mathcal{O}\big(N\sqrt{N}\, d\big). \tag{4.7}$$

This is strictly lower than dense 2D attention, up to a constant factor due to the two axial passes, and is well-suited to multi-scale decoding.

LRCA follows the pipeline in Fig. 4.4: patchify & tokenize (Eq. (4.1)) → row

and column self-attentions (Eqs. (4.3)–(4.5)) → Untokenization & Unpatchification (Eq. (4.6)). Conceptually, it factorizes 2D interactions into two axial components; in practice, this method executes both with a single SDPA call. Placed after the ConvBlock, LRCA supplies global shape context with substantially reduced complexity (Eq. (4.7)), complementing the decoder's local evidence.

4.3.4 Incorporating Prior Knowledge with Gray–Scott Noise

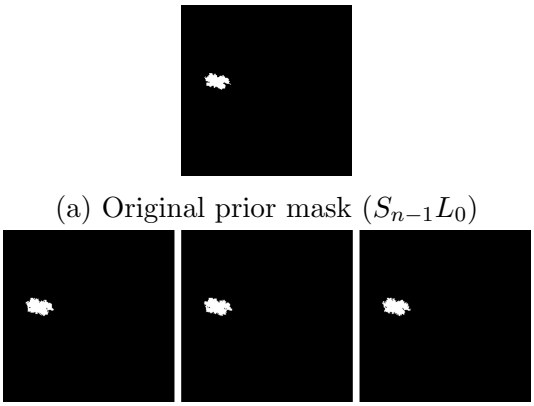
This method reuses the level–0 mask from the previous stage, denoted $S_{n-1}L_0$ (*e.g.*, S_2L_0 when refining with S3), as spatial prior for stage n. Before injection, this prior is stochastically perturbed to model boundary uncertainty. As shown by the Decoder schematic, the perturbed prior is only injected at intermediate levels L_m with $m > 0$ of the current stage, where it is fused with the convolutional stream and refined by LRCA.

To emulate realistic, small yet clinically meaningful boundary variations, this study evolves a Gray–Scott reaction–diffusion field [181] on a narrow band around the predicted contour:

$$\begin{aligned} \frac{\partial U}{\partial t} &= D_u \nabla^2 U - UV^2 + F(1-U), \\ \frac{\partial V}{\partial t} &= D_v \nabla^2 V + UV^2 - (F+k)V. \end{aligned} \tag{4.8}$$

Implementation is band-limited and lightweight: (i) form a 5-pixel dilation band of $S_{n-1}L_0$; (ii) initialize random seeds within this band and integrate (4.8) with D_u=0.16, D_v=0.08, F=0.060, k=0.062; (iii) threshold at τ=0.25 and lightly smooth to preserve contour continuity. These settings yield coral-like edge fluctuations that mimic mammographic irregularities while keeping global shape intact.

Fig. 4.5 visualizes the effect of Gray–Scott perturbation: panel (a) shows the original prior mask $S_{n-1}L_0$; panel (b) shows multiple perturbed variants with subtle



(a) Original prior mask ($S_{n-1}L_0$)

(b) Masks after Gray–Scott perturbation

Figure 4.5 : Effect of Gray–Scott noise on the prior mask. (a) Original prior mask $S_{n-1}L_0$. (b) Masks after Gray–Scott perturbation. The perturbation adds plausible boundary uncertainty while preserving overall tumor structure, which improves robustness during stage-n refinement (*e.g.*, S3).

but meaningful boundary deviations while preserving the tumor's overall structure. In the network, these perturbed $S_{n-1}L_0$ masks are selectively injected into the refinement decoder(s) at levels L_1 and L_2 (implemented with S3 in the three-stage TriNet; see the injection branch and fusion nodes in Fig. 4.3(a)). Concretely, the prior enters the Decoder's side branch, passes through a *ConvBlock*, is merged with the top-down feature of that level, and is subsequently refined by LRCA within the same block.

During training, the Gray–Scott seeds are resampled, so the refinement stage encounters diverse boundary configurations. This stochastic exposure builds robustness to imperfect priors while retaining their spatial guidance, leading to higher accuracy and fewer propagated errors compared with treating $S_{n-1}L_0$ as a deterministic, noise-free input.

4.3.5 Loss Functions

TriNet produces segmentation maps from three supervision heads, denoted **S2**, **S3**, and the final fusion head. Each stage is supervised with a tailored composite loss, and the resulting stage-wise losses are aggregated via a weighted sum to balance

optimization across stages.

Let Ω denote the set of pixels in a 2D mammogram, with cardinality $|\Omega| = N$. For each image in a mini-batch, $y \in \{0,1\}^N$ denotes the vectorized ground-truth mask and $\hat{y} \in \mathbb{R}^N$ the corresponding logits predicted by a given head. All losses are computed per sample and then averaged over the mini-batch. We write $\sigma(\cdot)$ for the sigmoid function and use a small constant $\varepsilon > 0$ for numerical stability.

We adopt three basic losses: the Dice–BCE composite loss (DiceBCE) in Eq. (4.11), the recall-oriented MissPenalty term in Eq. (4.12), and the proposed Adaptive Boundary-Alignment Loss (ABALoss) in Eq. (4.15).

Dice-BCE (DiceBCE). To encourage both region overlap and per-pixel calibration, we combine a soft Dice loss with a binary cross-entropy (BCE) term. The Dice component is applied to probabilities, whereas the BCE component is implemented on logits to match common practice.

$$\mathrm{DiceLoss}(y, \hat{y}) = 1 - \frac{2\sum_{i\in\Omega} y_i\,\sigma(\hat{y}_i) + \varepsilon}{\sum_{i\in\Omega} y_i + \sum_{i\in\Omega} \sigma(\hat{y}_i) + \varepsilon}. \tag{4.9}$$

Eq. (4.9) defines the soft Dice loss over vectorized tensors, where the numerator measures the intersection between prediction and ground truth, and the denominator is the sum of their areas.

$$\mathrm{BCE}(y, \hat{y}) = -\frac{1}{N}\sum_{i\in\Omega}\Big(y_i \log \sigma(\hat{y}_i) + (1 - y_i)\log\big(1 - \sigma(\hat{y}_i)\big)\Big). \tag{4.10}$$

Eq. (4.10) corresponds to the standard binary cross-entropy between the ground-truth labels y_i and the predicted probabilities $\sigma(\hat{y}_i)$. For clarity, Eq. (4.10) is written in terms of probabilities, while in implementation we use the usual numerically stable

formulation that directly operates on logits.

$$\text{DiceBCE}(y, \hat{y}) = 0.6\,\text{DiceLoss}(y, \hat{y}) + 0.4\,\text{BCE}(y, \hat{y}). \tag{4.11}$$

As shown in Eq. (4.11), DiceBCE blends the Dice and BCE components with weights 0.6 and 0.4, respectively, which empirically provide a stable gradient signal by slightly emphasizing region-level overlap while still regularizing pixel-wise predictions.

MissPenalty. Breast tumor screening is highly recall-sensitive: missing a positive region is more critical than over-segmenting the background. To explicitly discourage false negatives, we introduce a miss-only term that approximates one minus the recall of the positive class.

$$\text{MissPenalty}(y, \hat{y}) = 1 - \frac{\sum_{i\in\Omega} y_i\,\sigma(\hat{y}_i)}{\sum_{i\in\Omega} y_i + \varepsilon}. \tag{4.12}$$

Eq. (4.12) measures the fraction of ground-truth tumor pixels that are not covered by the prediction and is therefore equivalent to $1 - \text{Recall}$ up to the stabilizing constant ε. Since the normalization involves only the foreground cardinality $\sum_{i\in\Omega} y_i$, this loss is insensitive to false positives; over-segmentation of the background is controlled separately by the DiceBCE term.

Adaptive Boundary-Alignment Loss (ABALoss). While DiceBCE and MissPenalty enforce region coverage, they do not explicitly constrain boundary sharpness. We therefore propose a boundary-focused loss that compares edge maps extracted from the ground truth and the predicted probability map, with an edge thickness that adapts to the tumor scale.

Formally, let $x \in [0,1]^{\Omega}$ be a probability map. We define a differentiable mor-

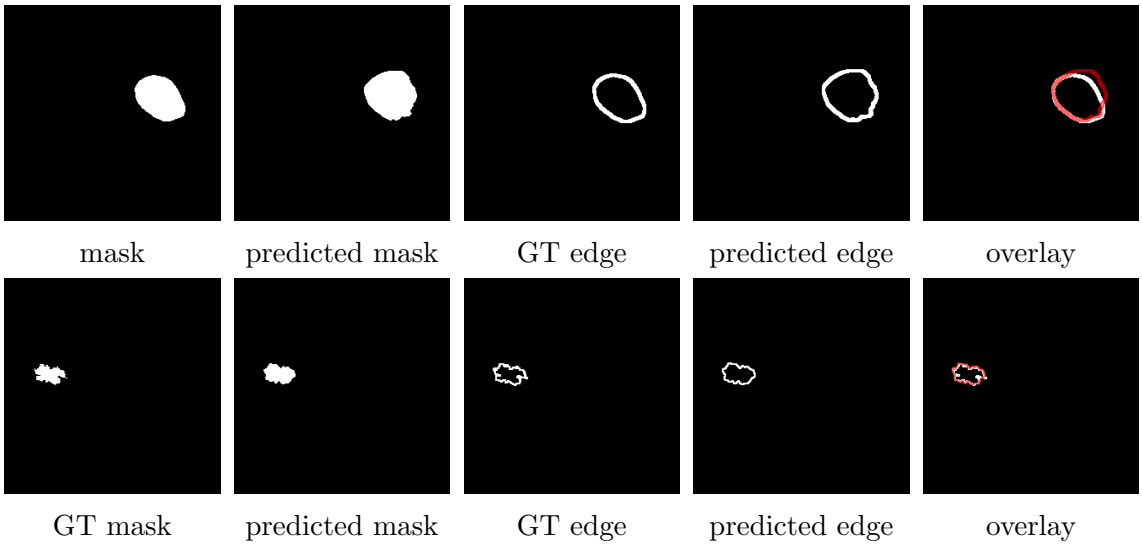


Figure 4.6 : Qualitative illustration of the proposed ABALoss on two mammograms. The top row shows a case with a large tumor, and the bottom row shows a case with a small tumor. Columns show the ground-truth mask, predicted mask, their automatically extracted edge maps, and the edge overlay maps. Note how the edge thickness adapts to the tumor scale, providing balanced supervision across lesion sizes.

phological gradient

$$E_k(x) = \delta_k(x) - \varepsilon_k(x), \tag{4.13}$$

where δ_k and ε_k denote dilation and erosion with a square structuring element of size k. In our implementation, both operators are realized via max-pooling operations (and their dual via negation) with kernel size k and unit stride, which makes $E_k(\cdot)$ fully differentiable with respect to x.

To handle the extreme tumor-size variation (14.6–3682.6 mm^2), the kernel size k is chosen adaptively for each sample. Given a binary mask y, we first compute its minimum-area enclosing rectangle and take the longer side length L_i (in pixels). The kernel size is then set to

$$k \;=\; \text{clip}_{[3,9]}\big(\text{round}(L_i/15)\big), \tag{4.14}$$

and further forced to be an odd integer so that the structuring element is symmetric. Small tumors therefore, use thin edges, whereas large tumors use thicker contours, yielding balanced boundary supervision across scales.

Let $E(y)$ denote the edge map of the ground-truth mask obtained by applying $E_k(\cdot)$ to the binarized mask, and let $E(\sigma(\hat{y}))$ be the edge map extracted from the predicted probabilities using the same kernel size k.

$$\mathrm{ABALoss}(y, \hat{y}) = 1 - \frac{2 \sum_{i \in \Omega} E(y)_i \, E(\sigma(\hat{y}))_i}{\sum_{i \in \Omega} E(y)_i + \sum_{i \in \Omega} E(\sigma(\hat{y}))_i + \varepsilon}. \tag{4.15}$$

Eq. (4.15) defines ABALoss as a Dice loss between the two edge maps, encouraging accurate alignment of tumor boundaries across a wide range of lesion sizes by focusing on scale-adaptive edges rather than full regions. Additional qualitative examples of the extracted edge maps for large and small tumors are provided in Fig. 4.6

The three losses are combined differently at each stage of TriNet.

S2 (initial decoding). This stage prioritizes high recall and explicitly penalizes missed tumor regions.

$$\mathcal{L}_{\mathrm{S2}} = (1 - \lambda_{\mathrm{mp}}) \, \mathrm{DiceBCE}(y, \hat{y}^{\mathrm{S2}}) + \lambda_{\mathrm{mp}} \, \mathrm{MissPenalty}(y, \hat{y}^{\mathrm{S2}}). \tag{4.16}$$

Eq. (4.16) combines DiceBCE with MissPenalty, where λ_{mp}=0.30 in all experiments controls the strength of the recall-oriented term.

S3 (prior-guided refinement). Once a coarse tumor region has been identified, we emphasize region overlap and boundary fidelity.

$$\mathcal{L}_{\mathrm{S3}} = (1 - \lambda_{\mathrm{edge}}) \, \mathrm{DiceBCE}(y, \hat{y}^{\mathrm{S3}}) + \lambda_{\mathrm{edge}} \, \mathrm{ABALoss}(y, \hat{y}^{\mathrm{S3}}). \tag{4.17}$$

Eq. (4.17) augments DiceBCE with ABALoss, where λ_{edge}=0.10 controls the strength

of the boundary constraint.

Final fusion output. The final fusion head is trained with the same boundary-aware supervision:

$$\mathcal{L}_{\text{Out}} = (1 - \lambda_{\text{edge}})\,\text{DiceBCE}(y, \hat{y}^{\text{Out}}) + \lambda_{\text{edge}}\,\text{ABALoss}(y, \hat{y}^{\text{Out}}). \tag{4.18}$$

Eq. (4.18) mirrors Eq. (4.17), enforcing boundary alignment at the final prediction.

The stage-wise losses are aggregated with weights $(w_{\text{S2}}, w_{\text{S3}}, w_{\text{Out}}) = (0.4,\ 0.4,\ 1.0)$:

$$\mathcal{L}_{\text{total}} = w_{\text{S2}}\,\mathcal{L}_{\text{S2}} + w_{\text{S3}}\,\mathcal{L}_{\text{S3}} + w_{\text{Out}}\,\mathcal{L}_{\text{Out}}. \tag{4.19}$$

Eq. (4.19) reflects stronger supervision at the final output while still providing deep supervision to intermediate stages. For the extended 4-stage variant TriNet (+S4) with an additional refinement stage S4, we similarly use $(w_{\text{S2}}, w_{\text{S3}}, w_{\text{S4}}, w_{\text{Out}}) = (0.4,\ 0.4,\ 0.4,\ 1.0)$.

4.4 Experiments

To validate TriNet's performance on challenging breast tumor segmentation, this study conducted comprehensive experiments on three benchmark datasets: INbreast [2], CBIS-DDSM [3], and DMID [4], and then compared TriNet against existing state-of-the-art approaches using established evaluation protocols. This section introduces the datasets, evaluation metrics, and experimental settings, followed by comprehensive quantitative and qualitative results.

4.4.1 Datasets and Pre-processing

This study evaluates three public mammography datasets that collectively span acquisition technology, population, and annotation style: INbreast (full-field digital mammography, FFDM) [2], CBIS-DDSM (screen-film mammography, SFM) [3],

and DMID (FFDM) [4]. These three datasets collectively cover both full-field digital mammography (FFDM) and digitized screen-film mammography (SFM), heterogeneous tumor presentations (masses, calcifications, asymmetries, architectural distortions), and differing label granularities.

The INbreast benchmark [2] comprises 410 high-resolution mammograms from 115 cases, with native image sizes ranging from 3328×2560 to 4084×3328 pixels. Of these, 107 images include pixel-wise tumor annotations provided by expert radiologists. INbreast presents significant segmentation challenges due to three factors: (i) limited annotated samples, (ii) severe class imbalance where tumor areas typically occupy less than 5% of the field of view, and (iii) broad phenotypic variability spanning masses, calcifications, asymmetries, and architectural distortions encountered in clinical practice.

CBIS-DDSM [3] is a curated modernization of the original DDSM resource, offering standardized case organization, tumor annotations, and metadata for benchmarking. Following common practice in the literature, this study focuses on the *mass* subset comprising 1,231 images for the segmentation experiments. Each case is accompanied by tumor-level labels and supporting information (*e.g.*, view/laterality and pathology), facilitating consistent experimental setup and cross-paper comparability. Unlike FFDM, CBIS-DDSM images are derived from digitized screen-film mammograms, which exhibit distinct noise texture, contrast behavior, and intensity distribution compared with native digital acquisitions. These modality-related differences partly explain the overall lower performance on CBIS-DDSM. This study trains on the official training portion and reports results on the official test split to avoid overlap between training and evaluation data and to enable fair comparison with prior work.

DMID [4] is a recently released single-institution FFDM collection acquired at

Samved Hospital (Ahmedabad, India) using a Siemens MAMMOMAT 3000 Nova system. Images are distributed in DICOM/TIFF with high native resolutions (typically around 4743×6000 or 3520×4784), and are accompanied by structured metadata (laterality, view, background tissue composition) and tumor-level clinical information (abnormality type, BI-RADS category, and approximate coordinates/radius). For abnormal studies, binary masks delineate the target tumor, enabling pixel-level supervision. This study uses the tumor-bearing subset comprising 269 images with available masks, and adopts patient-level splits and evaluation consistent with INbreast.

Because neither INbreast nor DMID provides an official test partition, this study employs five-fold *patient-level* cross-validation on these two datasets. Specifically, the subjects are split to ensure that images from the same patient never appear in both training and test sets within any fold, thereby preventing data leakage and yielding realistic performance estimates for clinical deployment. Splits are stratified to preserve the original tumor prevalence and view/laterality distributions across folds. In contrast, CBIS-DDSM includes an official split; this study trains on the corresponding training portion and reports results on its official test set. This unified protocol ensures fair comparisons across datasets of differing sizes, acquisition technologies (FFDM vs. SFM), and label granularities.

Offline Preprocessing. This study applies a single, deterministic preprocessing pipeline to all data partitions (training and test sets) to avoid distribution shift. The pipeline first performs contrast equalization using Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clip limit of 2.0 and a 4×4 tile grid configuration. This enhancement strategy improves local tissue contrast while preventing noise over-amplification in low-signal regions, thereby facilitating better discrimination between tumor and normal tissues. Following contrast enhancement, spatial normalization is applied by resizing all images to a uniform resolution of 512×512

pixels. The resizing operation employs bilinear interpolation for intensity images to maintain smooth gradients, while nearest-neighbor interpolation is used for corresponding segmentation masks to preserve label integrity without introducing spurious intermediate values. This standardized preprocessing ensures consistent appearance and resolution across all datasets while retaining the fine boundary details essential for accurate tumor delineation.

Online Data Augmentation. During training only, this study implements a dynamic augmentation pipeline applied at data-loading time to increase training data diversity and improve model generalization. The augmentation strategy comprises geometric transformations that are applied jointly to both images and their corresponding masks to maintain spatial correspondence. The primary augmentation consists of affine transformations incorporating in-plane rotation within the range of $[-45°, 45°]$, isotropic scaling between 0.85 and 1.15 of the original size, and translation in both horizontal and vertical axes within $\pm 10\%$ of the image dimensions. Exposed regions resulting from these transformations are filled with zero values to maintain consistent boundary conditions. Additionally, horizontal flipping is applied with a probability of 0.5 to further increase data variability and account for bilateral symmetry in mammographic imaging. This study deliberately excludes vertical flips and 90° rotations from the augmentation pipeline to preserve the anatomical constraints inherent to standard mammographic views. Such transformations would violate the natural positioning and gravitational effects present in clinical mammography acquisition, potentially introducing anatomically implausible configurations that could compromise the model's clinical applicability. The selected augmentation strategy thus balances data diversity with anatomical realism, ensuring that the model learns from variations that reflect actual clinical positioning variability rather than artificial transformations.

Table 4.1 : Comprehensive training and model configuration parameters used across all experiments.

Parameter	**Value**
General training configuration	
Batch size	12
Maximum number of epochs	300
Early stopping patience (epochs)	20
Optimizer	AdamW
Initial learning rate	1×10^{-3}
Weight decay	1×10^{-4}
LR scheduler	StepLR
LR decay factor (γ)	0.3
LR decay epoch (INbreast)	150
LR decay epoch (CBIS-DDSM, DMID)	70
Input image size	512×512
Loss function configuration	
Baseline loss function	DiceBCE
Weight of Dice loss term in DiceBCE	0.6
Weight of BCE term in DiceBCE	0.4
TriNet stage-wise loss weights ($w_{\mathrm{S2}}, w_{\mathrm{S3}}, w_{\mathrm{Out}}$)	$(0.4, 0.4, 1.0)$
ABALoss weight (λ_{edge})	0.1
MissPenalty weight (λ_{mp})	0.3
LRCA configuration	
Number of attention heads (LRCA)	4
Patch size at decoder level L_1 (t)	$t = 2$
Patch size at decoder level L_2 (t)	$t = 1$
Decoder levels with LRCA	L_1 and L_2 only
ABALoss configuration	
Morphological kernel size range (k)	$[3, 9]$
Morphological kernel size formula	$k = \mathrm{clip}_{[3,9]}\big(\mathrm{round}(L_i/15)\big)$
Regularization	
Dropout rate (TriNet)	0.3

4.4.2 Evaluation Metrics

This study employs five complementary metrics to comprehensively evaluate segmentation performance, each capturing distinct aspects of tumor boundary delineation accuracy that are critical for clinical deployment of computer-aided diagnosis systems.

Dice Similarity Coefficient (DSC). The DSC measures the spatial overlap between the predicted segmentation P and the ground truth G, providing a balanced assessment of segmentation quality that accounts for both precision and recall:

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}, \tag{4.20}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The DSC ranges from 0 (no overlap) to 1 (perfect overlap), with values above 0.7 often considered acceptable for medical image segmentation. As an overlap-based metric, DSC is strongly influenced by the alignment of the predicted and ground-truth boundaries, making it well suited for evaluating the accuracy required in treatment planning and surgical guidance applications. The factor of 2 in the numerator ensures that the metric treats the predicted and ground-truth regions symmetrically.

Intersection over Union (IoU). Also known as the Jaccard Index, IoU provides a stricter overlap measure compared to DSC:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{4.21}$$

IoU penalizes both over-segmentation and under-segmentation equally, offering a more conservative assessment than DSC. For any given overlap, IoU yields lower values than DSC, with the relationship $\text{DSC} = \frac{2\times\text{IoU}}{1+\text{IoU}}$. In breast tumor segmentation, IoU is particularly valuable for evaluating the model's ability to delineate irregular tumor boundaries without including surrounding healthy tissue, which is crucial for minimizing collateral damage during interventional procedures.

Accuracy (ACC). Pixel-wise accuracy quantifies the proportion of correctly

classified pixels across the entire image:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{4.22}$$

Here, TN denotes true negatives. While accuracy can be dominated by the large proportion of background pixels in mammographic images, it provides an important overall assessment of classification performance. In the context of breast imaging, high accuracy ensures that both tumor regions and healthy tissue are correctly identified, supporting comprehensive diagnostic evaluation. However, this metric should be interpreted cautiously in cases of severe class imbalance, where it may mask poor performance on the minority tumor class; therefore, it must be considered in conjunction with sensitivity and specificity.

Sensitivity (SEN). Also referred to as recall or true positive rate, sensitivity measures the model's ability to detect tumor pixels:

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{|P \cap G|}{|G|}. \tag{4.23}$$

This metric directly addresses the critical clinical requirement of minimizing missed diagnoses. In breast cancer screening, high sensitivity is paramount, as false negatives can delay treatment and adversely affect patient outcomes. A sensitivity approaching 1.0 indicates that the model successfully identifies nearly all tumor regions, although this must be balanced against specificity to avoid excessive false positives.

Specificity (SPEC). Specificity quantifies the model's ability to correctly identify non-tumor regions:

$$\text{SPEC} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{4.24}$$

High specificity reduces false positive rates, thereby minimizing unnecessary

biopsies, follow-up procedures, and patient anxiety. In practice, sensitivity and specificity must be jointly considered, as improving one often comes at the expense of the other depending on the decision threshold.

Together, these five metrics provide a comprehensive evaluation framework that addresses both the technical performance requirements and clinical deployment considerations for breast tumor segmentation systems. The combination of overlap-based metrics (DSC, IoU), pixel-wise classification metrics (ACC), and clinically oriented performance indicators (SEN, SPEC) ensures a thorough assessment of model capabilities across diverse clinical scenarios.

4.4.3 Experimental Environment and Training Setup

Computational Environment. All experiments were conducted on a Linux-based computing cluster equipped with two NVIDIA L40 GPUs, each providing 48 GB of VRAM. The implementation utilized the PyTorch 2.0 framework with Python 3.9 and CUDA 12.8 for GPU acceleration. This configuration provided sufficient computational resources for training deep segmentation models while maintaining reasonable training times across all experimental conditions.

Training Protocol. To ensure fair comparison across all evaluated architectures, we established a unified training protocol with consistent hyperparameters. All model backbones were initialized with *ImageNet-pretrained* weights. Table 4.1 summarizes the comprehensive configuration employed throughout our experiments, encompassing the general training configuration, loss function configuration, LRCA configuration, ABALoss configuration, and regularization settings. The training procedure incorporated early stopping with a patience of 20 epochs to prevent overfitting while ensuring convergence. The learning rate scheduling strategy was adjusted based on dataset characteristics: for the smaller INbreast dataset, the StepLR decay was triggered at epoch 150 to allow sufficient initial optimization, whereas for

the larger CBIS–DDSM and DMID datasets, decay commenced at epoch 70. This dataset-specific scheduling ensured stable convergence patterns across varying data scales.

The model configuration parameters reflect key architectural decisions specific to TriNet. The Lightweight Row–Column Attention (LRCA) module employs four attention heads in all instances, providing a balance between computational efficiency and representational capacity. The patch size t is configured by decoder level to accommodate different spatial resolutions: the decoder at level L_1 uses a patch size of $t = 2$ for moderate downsampling, while the decoder at level L_2 employs $t = 1$ for pixel-level attention. This hierarchical patch-size strategy ensures appropriate granularity at each resolution level while maintaining efficiency by activating LRCA only at intermediate decoder levels L_1 and L_2.

The Adaptive Boundary-Alignment Loss (ABALoss) dynamically adjusts its morphological kernel size k based on tumor dimensions, with k constrained to the range $[3, 9]$ pixels. The kernel size is computed as a function of the tumor's longer axis, following the formula $k = \text{clip}_{[3,9]}\big(\text{round}(L_i/15)\big)$, which provides balanced boundary supervision across the extreme size variations observed in mammographic tumors. The loss-function weights, particularly the 0.6:0.4 ratio between the Dice loss and BCE loss terms in DiceBCE, were empirically determined to balance regional overlap optimization with pixel-wise classification accuracy.

For TriNet's multi-stage supervision, the stage-wise loss weights $(w_{\text{S2}}, w_{\text{S3}}, w_{\text{Out}}) = (0.4, 0.4, 1.0)$ reflect the hierarchical importance of each refinement stage. The final output receives the highest weight to emphasize accurate final predictions, while intermediate stages S2 and S3 receive equal weights to provide sufficient gradient signals for progressive refinement without overwhelming the primary objective. The ABALoss weight $\lambda_{\text{edge}} = 0.1$ and MissPenalty weight $\lambda_{\text{mp}} = 0.3$ jointly balance

boundary precision with recall performance, ensuring that the model maintains high sensitivity to small tumors while preserving accurate boundary delineation.

This configuration, with dataset-specific adjustments limited to learning rate scheduling, ensures reproducible results and enables meaningful performance comparisons across different architectural designs and datasets. All experiments strictly adhered to these parameters to maintain experimental consistency and scientific rigor.

4.4.4 Experimental Results and Discussion

This study compared TriNet with classical architectures such as U–Net [14] and DeepLabV3 [147], a widely adopted medical segmentation baseline nnU-Net [182], as well as more recent models including Segformer [183] and MobileUNETR [184] under a unified preprocessing and optimization protocol (Sec. 4.4). Full metrics are reported in Tables 4.2, 4.3, and 4.4. Across all three benchmarks, TriNet achieves the strongest overlap–based performance with an order of magnitude fewer parameters than classical encoders, highlighting its suitability for resource–constrained deployment. On the small–sample INbreast benchmark [2], five–fold patient–level evaluation yields notably consistent results for TriNet, indicating that its coarse–to–fine refinement and edge–aware, stage–specific supervision are well aligned with whole–image tumor delineation.

Quantitative Comparisons

Table 4.2 presents the segmentation performance on the INbreast dataset [2]. TriNet demonstrates substantial advantages across all critical metrics. Both TriNet (3 stages) and TriNet (+S4) (4 stages) achieve superior DSC, IOU, and Sensitivity compared to all baseline methods, with TriNet (3 stages) attaining the best overall performance. Notably, TriNet accomplishes this using only 2.83M parameters and

Table 4.2 : Quantitative comparisons of segmentation performance of different models on the INbreast dataset [2] implemented in this research. Results are reported as mean ± std over five folds. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; ACC: Accuracy; SEN: Sensitivity; SPEC: Specificity; Params: Parameter count; FLOPs: Floating Point Operations; Time: Inference time per image; Best results are shown in boldface.

Model	**Params↓ (M)**	**FLOPs↓ (G)**	**Time↓ (ms)**	**DSC↑ (%)**	**IOU↑ (%)**	**ACC↑ (%)**	**SEN↑ (%)**	**SPEC↑ (%)**
U-Net	18.25	58.13	4.6	76.36 ± 7.61	63.22 ± 8.14	99.59 ± 0.15	70.65 ± 5.67	99.87 ± 0.09
nnU-Net	25.24	85.8	8.1	79.07 ± 7.89	66.35 ± 9.63	99.62 ± 0.13	77.04 ± 10.41	99.83 ± 0.04
DeepLabV3	58.63	242.1	18.0	78.50 ± 4.28	65.42 ± 4.99	99.61 ± 0.11	73.79 ± 6.63	99.86 ± 0.05
Segformer	29.7	31.71	15.8	79.16 ± 6.20	66.28 ± 8.09	99.60 ± 0.17	74.58 ± 5.52	99.85 ± 0.11
MobileUNETR	14.16	26.88	20.1	81.13 ± 5.66	68.74 ± 7.83	99.62 ± 0.09	75.20 ± 9.65	**99.90**± 0.02
TriNet(3 Stages)	**2.83**	**14.2**	22.3	**88.40**± 0.68	**79.31**± 1.06	**99.76**± 0.06	**88.20**± 1.54	99.88 ± 0.04
TriNet(4 Stages)	2.91	16.16	28.5	85.52 ± 1.23	74.90 ± 1.75	99.72 ± 0.06	83.56 ± 3.58	99.88 ± 0.04

Table 4.3 : Quantitative comparisons of segmentation performance of different models on the CBIS-DDSM dataset [3]. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; ACC: Accuracy; SEN: Sensitivity; SPEC: Specificity; Params: Parameter count; FLOPs: Floating Point Operations; Time: Inference time per image; Best results are shown in boldface.

Model	**Params↓ (M)**	**FLOPs↓ (G)**	**Time↓ (ms)**	**DSC↑ (%)**	**IOU↑ (%)**	**ACC↑ (%)**	**SEN↑ (%)**	**SPEC↑ (%)**
U-Net	18.25	58.13	4.6	34.59	21.42	98.85	**59.24**	99.07
nnU-Net	25.24	85.8	8.1	47.63	33.18	99.42	52.32	99.68
DeepLabV3	58.63	242.1	18.0	51.65	36.63	99.54	50.61	99.80
Segformer	29.7	31.71	15.8	48.72	33.66	99.47	49.08	99.75
MobileUNETR	14.16	26.88	20.1	41.09	27.03	99.48	36.61	**99.82**
TriNet(3 Stages)	**2.83**	**14.2**	22.3	**55.94**	**40.51**	**99.56**	56.46	99.79
TriNet(4 Stages)	2.91	16.16	28.5	53.03	37.38	99.49	56.76	99.72

Table 4.4 : Quantitative comparisons of segmentation performance of different models on the DMID dataset [4]. Results are reported as mean ± std over five folds. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; ACC: Accuracy; SEN: Sensitivity; SPEC: Specificity; Params: Parameter count; FLOPs: Floating Point Operations; Time: Inference time per image; Best results are shown in boldface.

Model	**Params↓ (M)**	**FLOPs↓ (G)**	**Time↓ (ms)**	**DSC↑ (%)**	**IOU↑ (%)**	**ACC↑ (%)**	**SEN↑ (%)**	**SPEC↑ (%)**
U-Net	18.25	58.13	4.6	70.14 ± 4.79	55.01 ± 5.14	99.04 ± 0.30	64.19 ± 6.14	99.75 ± 0.03
nnU-Net	25.24	85.8	8.1	64.93 ± 7.74	49.98 ± 8.11	98.87 ± 0.43	59.91 ± 11.06	99.72 ± 0.07
DeepLabV3	58.63	242.1	18.0	70.37 ± 7.40	55.57 ± 8.25	98.84 ± 0.78	69.01 ± 7.50	99.36 ± 0.85
Segformer	29.7	31.71	15.8	68.65 ± 12.11	54.22 ± 12.21	98.83 ± 0.90	66.89 ± 8.26	99.48 ± 0.62
MobileUNETR	14.16	26.88	20.1	64.11 ± 3.79	49.23 ± 4.05	98.97 ± 0.19	57.84 ± 7.26	99.72 ± 0.09
TriNet(3 Stages)	**2.83**	**14.2**	22.3	74.91 ± 3.75	61.37 ± 4.57	99.25 ± 0.18	70.27 ± 4.52	99.77 ± 0.06
TriNet(4 Stages)	2.91	16.16	28.5	**75.68**± 4.43	**62.11**± 5.53	**99.27**± 0.16	**70.39**± 7.57	**99.79**± 0.04

14.2 GMac per 512×512 image while delivering significant performance improvements over substantially larger models.

In addition to parameter count and FLOPs, we evaluated the inference time of the different models, measured in milliseconds per image on a single NVIDIA L40 GPU. As detailed in Tables 4.2, 4.3, and 4.4, lightweight baselines such as U-Net and nnU-Net exhibit fast inference speeds (4.6 ms and 8.1 ms, respectively). The progressive refinement design of TriNet introduces a moderate latency overhead, resulting in 22.3 ms for the 3-stage variant. However, this inference time remains well within the requirements for real-time or near real-time clinical deployment. Crucially, the substantial accuracy gains achieved by TriNet justify this slight increase in computational time on modern hardware accelerators.

Comparing TriNet and its extended 4-stage variant TriNet (+S4) reveals that

Table 4.5 : Quantitative comparisons of SOTA results as reported in literature on the INbreast dataset [2]. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; Best results are shown in boldface.

Model	**Year**	**Resolution**	**DSC↑ (%)**	**SEN↑ (%)**
AUNet [143]	2019	256×256	64.0	66.0
AM-MSP-cGAN [185]	2020	256×256	83.92	85.59
DS U-Net [186]	2021	256×256	79.0	81.0
v19U-Net++ [187]	2021	256×256	80.44	-
ACMSCNet [188]	2021	256×256	84.11	85.91
ARF-Net [144]	2022	256×256	85.06	83.0
TrEnD [29]	2023	256×256	84.68	86.35
SCSGNet [164]	2023	256×256	85.15	85.93
Improved UNet 3+ [34]	2023	320×320	85.89	-
EU-Net [189]	2024	256×256	87.63	-
Cas-DeepLabV3+ [24]	2024	512×512	80.0	-
D-ASP [35]	2024	256×256	74.59	78.16
TriNet (ours)	-	512×512	**88.40**	**88.20**

the default TriNet outperforms TriNet (+S4) despite having fewer parameters and fewer refinement stages. This counterintuitive result can be attributed to the deep supervision weight allocation strategy employed during training. The loss weight distribution $(w_{\text{S2}}, w_{\text{S3}}, w_{\text{Out}}) = (0.4, 0.4, 1.0)$ that proves effective for three-stage training does not translate optimally to the four-stage configuration, where weights are set to $(w_{\text{S2}}, w_{\text{S3}}, w_{\text{S4}}, w_{\text{Out}}) = (0.4, 0.4, 0.4, 1.0)$. In the four-stage setup, the increased number of intermediate supervisions with uniform weighting may dilute the optimization signal at each individual stage, potentially introducing gradient conflicts that hinder the network's ability to learn effective hierarchical refinements. This observation suggests that careful tuning of stage-specific supervision weights is essential when extending architecture depth, and that simply adding more stages with proportional weight distribution does not guarantee performance gains, particularly on datasets with limited sample sizes such as INbreast.

As shown in Table 4.3, TriNet maintains a clear performance advantage on the CBIS-DDSM dataset [3], achieving the highest DSC of 55.94% and IOU of 40.51%.

Notably, all models exhibit substantially lower performance metrics on CBIS-DDSM compared to INbreast, reflecting the increased difficulty of this dataset. This performance degradation can be attributed to the fundamental difference in image acquisition modalities between the two datasets. CBIS-DDSM consists of digitized film mammograms that were originally captured on screen-film systems and subsequently converted to digital format through scanning [3], whereas INbreast contains native full-field digital mammography (FFDM) images. Despite these inherent limitations of the digitized film modality, TriNet's multi-stage refinement architecture continues to outperform competing methods by significant margins, demonstrating its robustness across different mammography acquisition technologies and image characteristics. Results on CBIS-DDSM are reported on the official test set without cross-validation; thus no standard deviation is provided.

On DMID [4] (Table 4.4), the extended 4-stage variant TriNet (+S4) configuration achieves the best overall performance across all metrics, with DSC of 75.68% and IOU of 62.11%, reversing the performance relationship observed on INbreast where the default TriNet (3 stages) was superior. Unlike INbreast's limited sample size, where deeper architectures face optimization challenges due to gradient dilution across multiple supervision points, DMID's larger and more diverse image collection provides sufficient training signal to properly optimize the additional refinement stage. The four-stage architecture can thus fully exploit its enhanced representational capacity without succumbing to overfitting or gradient conflicts. This finding underscores that architectural depth decisions must be informed by dataset characteristics rather than following a one-size-fits-all approach, and that TriNet's hierarchical refinement framework demonstrates adaptive performance across varying data regimes.

Finally, literature–level comparisons on INbreast (Table 4.5) show that TriNet is competitive with or surpasses recent specialized designs while operating at higher

input resolution, reinforcing its effectiveness for whole–image breast tumor segmentation.

Table 4.6 : Ablation study on the INbreast dataset [2]. DSC: Dice Similarity Coefficient; IOU: Intersection over Union; ACC: Accuracy; SEN: Sensitivity; SPEC: Specificity.

Deep supervision	**ABALoss**	**Prior Layer**	**Gray-Scott**	**LRCA**	**DSC↑ (%)**	**IOU↑ (%)**	**ACC↑ (%)**	**SEN↑ (%)**	**SPEC↑ (%)**
✗	✗	✓	✓	✓	83.66± 5.07	72.34± 7.22	99.67± 0.14	82.26± 3.99	99.84± 0.13
✓	✗	✓	✓	✓	83.45± 2.81	72.01± 3.80	99.66± 0.10	85.07± 6.04	99.81± 0.04
✓	✓	✗	✗	✓	85.63± 3.01	75.15± 4.57	99.68± 0.08	84.42± 1.89	99.86± 0.06
✓	✓	✓	✗	✓	87.58± 0.96	78.01± 1.50	99.76± 0.06	85.37± 3.53	99.90± 0.03
✓	✓	✓	✓	✗	85.99± 2.18	75.71± 3.23	99.73± 0.08	83.19± 3.45	99.90± 0.03
✓	✓	✓	✓	✓	**88.40**± 0.68	**79.31**± 1.06	**99.76**± 0.06	**88.20**± 1.54	**99.88**± 0.04

4.4.5 Ablation Study

This study conducts ablation experiments on the key design choices of TriNet using the INbreast dataset to quantify their individual and joint contributions. Unless stated otherwise, the default configuration uses LRCA in the decoder (Sec. 4.3.3), employs prior injection at decoder levels L1/L2, and enables Gray-Scott perturbation on the priors. The model is trained with deep supervision using a stage-wise loss design: S2 is supervised by DiceBCE plus MissPenalty, whereas S3, S4, and the final output head (Out) are supervised by DiceBCE combined with ABALoss. In the ablation setting without deep supervision, only the final output head is supervised, and we use a plain DiceBCE loss at this head without any MissPenalty or ABALoss terms. Table 4.6 summarizes the quantitative comparison results on DSC, IOU, ACC, SEN, and SPEC. To complement these numerical findings, Fig. 4.7 provides a corresponding visual qualitative assessment, illustrating how the presence or



Figure 4.7 : Visual comparison for the ablation study. Rows correspond to different combinations of components evaluated in Table 4.6. The acronym **DAPGL** (inserted above the symbols) indicates the presence (✓) or absence (✗) of **D**eep supervision, **A**BALoss, **P**rior Layer, **G**ray-Scott, and **L**RCA, sequentially. The full TriNet configuration (bottom row) demonstrates the most coherent boundary delineation and fewest false predictions.

absence of each component directly impacts boundary delineation and false positive suppression.

Effect of LRCA. Removing LRCA blocks decreases performance from **88.40** to **85.99** DSC (↓ **2.41** points), and from **79.31** to **75.71** IOU (↓ **3.60**). Sensitivity drops by **5.01** points (88.20 → 83.19), indicating that LRCA's row/column axial attention injects complementary long-range shape cues into the decoder beyond local convolutional evidence.

Effect of prior injection and Gray-Scott perturbation. This study compares three variants: (i) no prior (no perturbation), (ii) prior only (no perturbation), and (iii) prior + Gray-Scott. Introducing priors (L1/L2) without perturbation improves DSC from **85.63** to **87.58** (↑ **1.95**); IOU increases by **2.86** (75.15 → 78.01), with a modest sensitivity gain (+0.95). Adding Gray-Scott on top of the priors yields a further DSC gain of **0.82** (87.58 → 88.40) and IOU +**1.30** (78.01 → 79.31), alongside a larger sensitivity boost (+**2.83**). These results demonstrate that the prior itself improves localization and recall, while Gray-Scott improves robustness to boundary uncertainty.

Effect of loss composition (ABALoss, deep supervision). When deep supervision is disabled (and thus stage-wise MissPenalty/ABALoss are absent), DSC drops from **88.40** to **83.66** (↓ **4.74**); IOU decreases by **6.97** and sensitivity by **5.94** (88.20 → 82.26). Furthermore, removing deep supervision drastically increases the standard deviation of DSC (from ±**0.68** to ±**5.07**) and IOU (from ±**1.06** to ±**7.22**), demonstrating that multi-stage guidance is crucial not only for overall accuracy but also for the stability and robustness of the model across different patient cases. Interestingly, applying deep supervision alone without ABALoss (✓✗✓✓✓) effectively forces the model to discover missing targets (raising sensitivity from **82.26**

to **85.07**), but it introduces severe over-segmentation, leading to a slight drop in DSC (83.66 → 83.45). This perfectly highlights the indispensable role of ABALoss in regularizing these rough predictions into precise boundaries. Ultimately, deep supervision is crucial for multi-stage refinement and recall (MissPenalty primarily boosts sensitivity by reducing misses), while ABALoss sharpens and stabilizes boundaries.

To further substantiate the quantitative findings, Fig. 4.7 presents a visual comparison of the ablation configurations across four representative cases. The acronym **DAPGL** denotes the sequential inclusion of **D**eep supervision, **A**BALoss, **P**rior Layer, **G**ray-Scott, and **L**RCA. As observed in the progressive rows, omitting key components visibly degrades segmentation quality in specific ways:

Deep Supervision and ABALoss: When both Deep supervision and ABALoss are omitted (✗✗✓✓✓), the model suffers from severe false negatives, completely missing the secondary subtle tumor in the first column and the tiny target in the fourth column. Crucially, merely restoring Deep supervision without ABALoss (✓✗✓✓✓) fails to yield meaningful visual improvements; although it forces the network to recover the missing targets, it results in poor boundary alignment with noticeable over-segmentation bleeding into adjacent healthy tissue. This visual evidence highlights the indispensability of ABALoss for precise boundary constraint, perfectly echoing the numerical results in Table 4.6 where adding Deep supervision alone stagnates the Dice score (83.66% vs. 83.45%), while introducing ABALoss is required to achieve substantial performance gains.

Prior Layer and Gray-Scott Perturbation: Regarding spatial guidance, omitting the Prior Layer inherently removes the Gray-Scott perturbation (✓✓✗✗✓) since the noise is applied directly to the prior mask. Without this explicit spatial constraint, the model generates prominent false positive predictions (spurious red blobs) in the background tissues, as clearly seen in the first and second columns. Rein-

troducing the Prior Layer but keeping it deterministic (no Gray-Scott, ✓✓✓✗✓) effectively suppresses these background artifacts and dramatically improves target localization. This visual enhancement strongly correlates with the significant Dice jump (85.63% to 87.58%) reported in the ablation table. Furthermore, comparing this deterministic prior configuration to the full model demonstrates that lacking Gray-Scott perturbation still leaves the boundaries somewhat imprecise and vulnerable to local ambiguities. Adding Gray-Scott noise introduces vital boundary robustness.

LRCA vs. Full Configuration: Finally, the configuration lacking the Lightweight Row-Column Attention (✓✓✓✓✗) illustrates the necessity of global context. Without LRCA, the model's structural awareness is diminished, causing it to completely miss the lower tumor in the first column and struggle with the tiny target in the fourth column. In contrast, the complete TriNet framework (✓✓✓✓✓) successfully detects both large and minimal targets, maintains tight boundary alignment, and effectively minimizes false positives, validating the synergistic design of all proposed modules.

In summary, both the quantitative and qualitative ablation results consistently demonstrate the indispensable nature of all proposed components. Their combined integration yields substantial improvements across all key evaluation metrics while visually resolving critical clinical challenges. Specifically, LRCA supplies essential global shape context; spatial priors (L1/L2) enhance initial target localization and overall recall; Gray-Scott perturbation injects crucial robustness against prior boundary ambiguity; and the tailored loss design (deep supervision combined with ABALoss) is pivotal for enforcing both pixel-level accuracy and morphological stability.

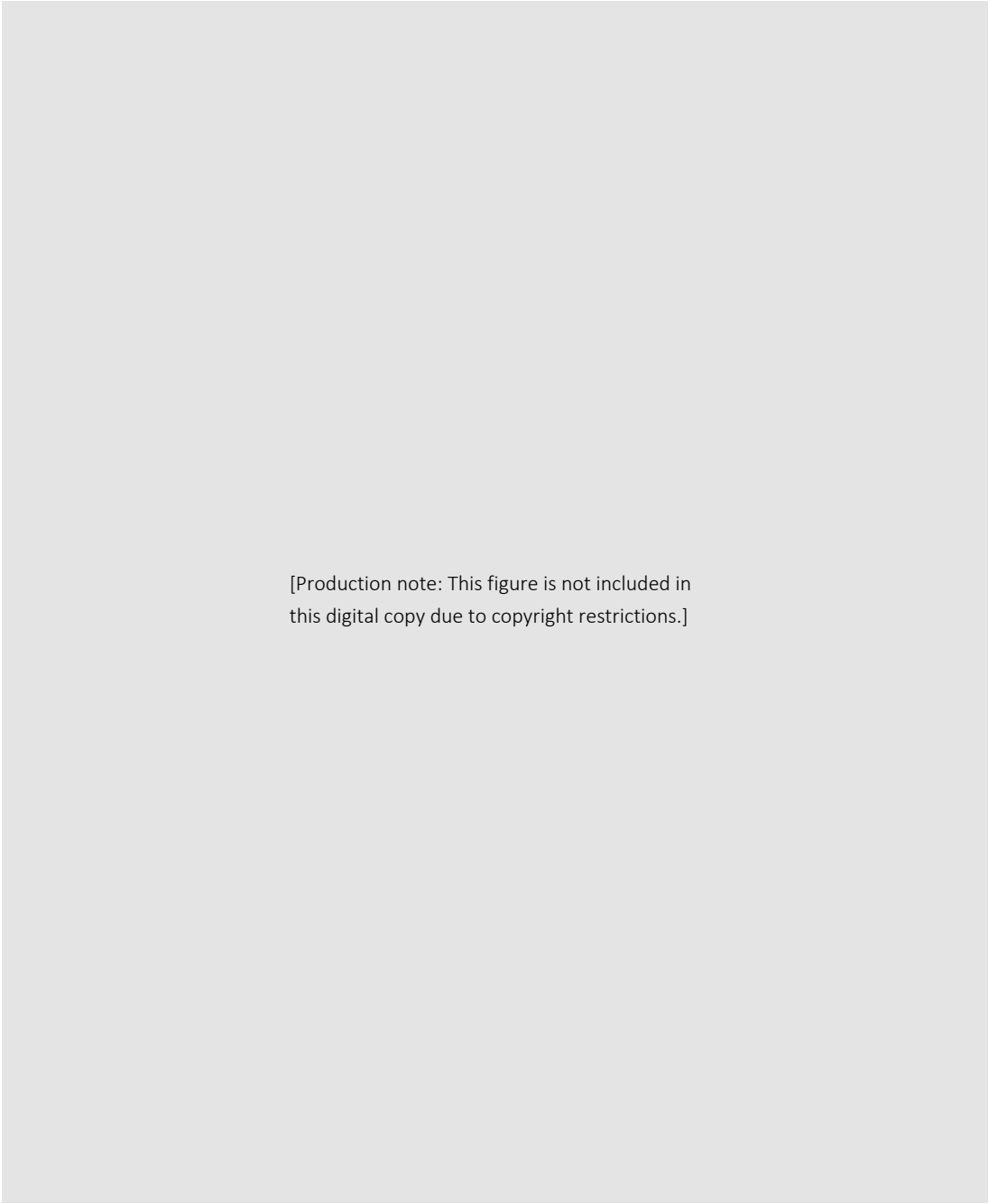


Figure 4.8 : Comparison of breast tumor segmentation results on the INbreast dataset [2]. Rows correspond to different models (including Original and Ground Truth), and columns correspond to different images. **TriNet** demonstrates strong boundary delineation.

Visual Comparisons

Fig. 4.8 presents a qualitative comparison of breast tumor segmentation results across multiple state-of-the-art models on representative cases from the INbreast

dataset [2]. The selected cases encompass a spectrum of clinical challenges commonly encountered in mammographic analysis, including tumors with indistinct boundaries, images containing tumor-mimicking regions, and subtle small-scale abnormalities. In the visualization, ground truth annotations are rendered as semi-transparent white regions, while model predictions are overlaid in semi-transparent orange, enabling direct assessment of segmentation accuracy and boundary delineation quality.

The first column exemplifies a particularly challenging scenario featuring two large tumors with highly ambiguous boundaries between tumor tissue and surrounding normal parenchyma. This case represents a common difficulty in breast imaging where infiltrative tumor margins gradually blend into adjacent fibroglandular tissue without sharp demarcation. The baseline U-Net exhibits fragmented predictions with incomplete coverage of the tumor extent, particularly struggling to capture the diffuse infiltrative patterns. Similarly, the widely adopted nnU-Net baseline, although slightly more robust, still struggles to capture the diffuse infiltrative patterns comprehensively. MobileUNETR demonstrates modest improvements in continuity but still fails to fully delineate the irregular tumor boundaries. DeepLabV3 and SegFormer, despite their sophisticated feature extraction mechanisms, produce over-conservative predictions that substantially underestimate the true tumor extent. In contrast, TriNet achieves comprehensive coverage of both tumors while maintaining strong correspondence with the ground truth boundaries, demonstrating its superior capability to model complex morphological patterns and ambiguous tissue transitions.

The second and third columns present medium-sized tumors in cases where the surrounding breast tissue contains non-tumor regions that exhibit visual characteristics remarkably similar to a tumor. These tumor-mimicking regions represent a critical challenge for automated segmentation systems, as they can easily trigger

spurious activations. Notably, nearly all baseline methods fail on these cases, incorrectly classifying the interfering regions as tumor tissue and producing substantial false positive predictions. In stark contrast, TriNet successfully distinguishes the true tumor boundaries from the confounding regions, maintaining accurate segmentation without spurious activations. This discriminative capability can be attributed primarily to the proposed prior integration mechanism, which leverages the coarse mask from S2 as spatial guidance. This prior knowledge provides explicit spatial constraints that guide the network to focus on regions with high tumor likelihood while suppressing false activations in confounding areas. The Lightweight Row-Column Attention (LRCA) modules further enhance this capability by capturing long-range spatial dependencies and holistic shape patterns, enabling the model to reason about contextual coherence beyond local texture similarities. The ablation studies corroborate these observations: without prior knowledge integration, TriNet exhibits misclassifications similar to baseline methods on these interfering regions, while removal of LRCA leads to minor but noticeable false positive activations, confirming that the synergy between spatial priors and global attention mechanisms is essential for robust discrimination in challenging cases with tumor-mimicking features.

The fourth and fifth columns showcase extremely challenging cases involving small, subtle tumors that are difficult to detect even through careful visual inspection. These cases represent scenarios where early-stage detection is clinically critical yet technically demanding due to limited spatial extent and low contrast against surrounding tissue. The fourth column presents a particularly revealing case where all baseline methods fail to identify the single small tumor present in the image. TriNet, however, successfully detects and segments this subtle abnormality, producing a compact and well-localized prediction that closely corresponds to the ground truth annotation. This remarkable sensitivity to small tumors can be directly at-

tributed to the deep supervision strategy. The ablation studies provide compelling evidence for this design choice: when deep supervision is disabled and only the final output receives supervisory signals, TriNet fails to detect the tumor in the fourth column. The fifth column, featuring another small tumor with slightly higher contrast, further demonstrates TriNet's robust performance on minimal targets.

Across all presented cases, TriNet consistently demonstrates superior boundary delineation compared to competing approaches. This qualitative assessment corroborates the quantitative metrics presented in Table 4.2, confirming that TriNet's architectural innovations translate into tangible improvements in practical segmentation scenarios.

4.5 Conclusion

This chapter presents **TriNet**, a progressive refinement network specifically designed to address three fundamental challenges in mammographic tumor segmentation: boundary precision in low-contrast regions, adaptability to extreme size variations, and robust utilization of imperfect priors. Unlike conventional multi-stage approaches that concatenate full encoder-decoder networks—thereby approximately doubling parameters and inference time—TriNet employs a single MobileViT encoder (S1) paired with sequential decoder stages (S2, S3, and optionally S4), where all decoders are instantiated from a unified primitive that fuses encoder skips with top-down features.

To complement local convolutional processing with global shape reasoning, this study introduces **Lightweight Row-Column Attention (LRCA)**, a width-preserving attention unit that factorizes two-dimensional self-attention into row-wise and column-wise operations. This design reduces computational complexity from $\mathcal{O}(L^4 d)$ to $\mathcal{O}(L^3 d)$ while supplying holistic context at selected intermediate decoder levels (L1/L2)—particularly valuable for capturing long-range dependencies in mammo-

graphic tumor shapes.

To explicitly model prior uncertainty rather than treating intermediate predictions as deterministic, this study perturbs the previous stage's mask with **Gray–Scott reaction–diffusion** noise and selectively injects it at intermediate resolutions during refinement. This strategy substantially improves robustness to imperfect priors and mitigates error propagation across stages. Training proceeds in a stage-specific manner: S2 employs DiceBCE with MissPenalty to prioritize recall, while subsequent stages (S3/S4) and the fused output adopt DiceBCE with the proposed **ABALoss** to consolidate boundary precision across diverse tumor scales through adaptive morphological gradients.

Experimental validation across three diverse mammography datasets reveals TriNet's architectural advantages and adaptive performance characteristics. On INbreast, the three-stage TriNet achieves **88.40%** DSC and **79.31%** IoU, surpassing both established clinical baselines like **nnU-Net** and recent efficient transformers like MobileUNETR (by 7.27% and 10.57% respectively). Notably, TriNet achieves this while using 5× fewer parameters (2.83M vs. 14.16M) and requiring just **22.3 ms of inference time per image**. Furthermore, comprehensive visual ablation studies confirm that this synergistic design effectively eliminates spurious activations in tumor-mimicking regions and recovers subtle microscopic targets that conventional architectures fail to detect. The performance gap widens further against larger models like DeepLabV3 (58.63M parameters), which achieves only 78.50% DSC despite having 20× more parameters.

Performance analysis on CBIS-DDSM, where digitized film characteristics challenge all methods, demonstrates TriNet's robustness with **55.94%** DSC—a 4.29% improvement over DeepLabV3 and 7.22% over Segformer. The substantial performance degradation across all models (30–35% DSC drop from INbreast) reflects

the fundamental differences between native digital and digitized film modalities, yet TriNet maintains its relative advantage through effective hierarchical refinement.

On the larger DMID dataset, an interesting architectural pattern emerges: the extended 4-stage variant TriNet (+S4) (**75.68%** DSC) outperforms the default TriNet (3 stages) (74.91% DSC), reversing the relationship observed on INbreast. This dataset-dependent performance inversion suggests that deeper refinement stages benefit from larger training sets, while smaller datasets like INbreast may experience gradient dilution with excessive intermediate supervisions.

In conclusion, TriNet advances mammographic tumor segmentation by systematically addressing boundary precision, scale adaptability, and prior uncertainty through an architecturally efficient framework. By avoiding encoder redundancy, integrating lightweight global context, and explicitly modeling prior imperfection, TriNet provides a practical and clinically viable solution for enhancing breast cancer screening workflows. The framework's efficiency-accuracy balance positions it as a strong foundation for real-world deployment, where computational constraints and diagnostic precision must be simultaneously satisfied.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis addressed fundamental challenges in computational breast cancer imaging by developing specialized deep learning frameworks for two complementary but distinct clinical scenarios: volumetric tumor detection in three-dimensional Automated Breast Ultrasound System (ABUS) data and precise tumor segmentation in two-dimensional mammography images. The overarching objective was to enhance diagnostic capabilities across both established and emerging breast imaging modalities while maintaining the computational efficiency and clinical practicality essential for real-world deployment in diverse healthcare environments.

5.1.1 Addressing the Research Problems

As outlined in Chapter 1, current breast cancer screening modalities face significant limitations, including reduced sensitivity in dense breast tissue, operator-dependent variability, high false-positive rates, and computational constraints that hinder the adoption of advanced imaging technologies. Traditional computer-aided diagnosis systems relying on handcrafted features demonstrated poor generalization across imaging protocols, while contemporary deep learning approaches exhibited pronounced asymmetries in research focus. Specifically, the overwhelming concentration on two-dimensional mammography segmentation left three-dimensional modalities like ABUS severely underexplored despite their clinical advantages, including superior spatial information and elimination of tissue superposition artifacts.

This thesis systematically addressed these gaps through a dual-track research strategy targeting the distinct computational and clinical requirements of volumetric detection and planar segmentation.

5.1.2 Principal Contributions and Outcomes

ABUSDet: Efficient 3D Tumor Detection for ABUS. For three-dimensional ABUS volumes, this study proposed **ABUSDet**, a novel 2.5D detection framework that fundamentally reconceptualizes volumetric medical data processing for object detection tasks. Unlike approaches that either sacrifice depth information through pure two-dimensional processing or incur prohibitive computational costs through full three-dimensional convolutions, ABUSDet introduces an intelligent architectural middle ground. The framework partitions each ABUS volume into overlapping multi-slice blocks, where 28 adjacent slices are stacked as input channels and compressed to three channels via a Channel Squeeze module, enabling integration with established two-dimensional detection architectures (Faster R-CNN) and facilitating transfer learning from natural image domains. Following two-dimensional candidate generation, three-dimensional tumor coordinates are reconstructed and clustered using DBSCAN to produce final detections.

A critical innovation in ABUSDet is the **deliberate training mechanism**, inspired by cognitive learning theory, which dynamically prioritizes under-trained batches during optimization based on a batch-level positive-hit criterion. This strategy directly addresses the severe class imbalance inherent in screening applications, where malignancies represent less than one percent of examined tissue. Experimental validation demonstrated that deliberate training substantially improved detection performance, increasing sensitivity from 78.33% to 83.33% and elevating the area under the Free-Response Receiver Operating Characteristic (FROC) curve from 2.65 to 3.06 under controlled ablation conditions.

On the complete test set, ABUSDet achieved **77.94% sensitivity at 3.0 false positives per volume**, significantly surpassing both pure two-dimensional methods (66.63%) and a three-dimensional baseline (14.58%), while maintaining practical clinical throughput with an inference time of approximately **11.0 seconds per volume** at a stride equals ten. These results validate that volumetric context and clinical efficiency need not be mutually exclusive, providing a viable pathway for integrating ABUS into routine screening workflows at specialized medical centers.

TriNet: Precise and Efficient Mammographic Segmentation. For two-dimensional mammography images, this study introduced **TriNet**, a multi-stage iterative refinement network that treats segmentation as progressive boundary optimization while maintaining computational efficiency. Unlike conventional multi-stage approaches that concatenate multiple full encoder-decoder networks and thereby approximately double parameters and inference time, TriNet employs a **single shared MobileViT encoder** (Stage 1) paired with a configurable sequence of decoder stages (Stages 2 and 3, with an optional Stage 4 evaluated experimentally). All decoder stages are instantiated from a unified primitive that consistently fuses encoder skip connections with top-down features, concentrating computational capacity where detail recovery actually occurs.

To complement local convolutional processing with holistic shape reasoning, this study developed **Lightweight Row-Column Attention (LRCA)**, a width-preserving attention mechanism that factorizes two-dimensional self-attention into parallel row-wise and column-wise operations. This architectural innovation reduces computational complexity from $\mathcal{O}(L^4 d)$ to $\mathcal{O}(L^3 d)$ while supplying global spatial dependencies at every decoder level, proving particularly valuable for capturing the long-range shape consistency inherent in mammographic tumor boundaries under tight computational budgets.

To explicitly model uncertainty in intermediate predictions rather than treating them as deterministic, TriNet incorporates **uncertainty-aware prior integration** through Gray-Scott reaction-diffusion perturbations. The previous stage's mask is stochastically perturbed to emulate realistic boundary variability and selectively injected at intermediate decoder resolutions (L_1/L_2) where spatial detail is forming yet features remain stable. This design substantially improves robustness to imperfect priors and mitigates error propagation across refinement stages.

Training employs a **stage-specific supervision strategy** aligned with each stage's architectural role: Stage 2 prioritizes recall through Dice Binary Cross-Entropy (DiceBCE) combined with Miss Penalty loss, while subsequent stages (Stage 3 and optionally Stage 4) and the fused output adopt DiceBCE combined with the proposed **Adaptive Boundary-Alignment Loss (ABALoss)**. ABALoss enforces boundary fidelity across extreme tumor size variations by computing scale-adaptive morphological gradients on probabilistic predictions, enabling thin edge constraints for small tumors and thicker contours for large masses.

Comprehensive experimental validation across three diverse mammography datasets demonstrated TriNet's superior performance and clinical applicability. On the INbreast benchmark, TriNet achieves state-of-the-art performance with **88.40% Dice score**, **79.31% IoU**, **99.76% accuracy**, **88.20% sensitivity**, and **99.88% specificity**. The architecture maintains robust performance on the challenging CBIS-DDSM digitized film dataset with **55.94% DSC** and **40.51% IoU**, outperforming DeepLabV3 (51.65% DSC) and other baselines despite the inherent modality limitations. On the large-scale DMID dataset, the four-stage configuration achieves **75.68% DSC** and **62.11% IoU**, demonstrating the architecture's ability to adaptively leverage additional refinement stages when sufficient training data is available.

This consistent superiority across native digital (INbreast, DMID) and digi-

tized film (CBIS-DDSM) modalities, achieved with only **2.83–2.91M parameters**, represents a 13–20× reduction compared to conventional architectures. Such exceptional parameter efficiency enables deployment in resource-constrained clinical environments without sacrificing segmentation quality, addressing the practical reality that many healthcare facilities, particularly in developing regions, operate under significant computational and economic constraints. The architecture's adaptive performance—with three-stage configurations excelling on smaller datasets and four-stage variants optimizing for larger data regimes—further enhances its clinical versatility across diverse deployment scenarios.

5.1.3 Clinical Significance and Deployment Considerations

Together, ABUSDet and TriNet provide a coherent, clinically aligned computational framework that couples efficient three-dimensional detection with precise two-dimensional segmentation. This dual-track approach directly addresses the pronounced research asymmetry identified in Chapter 1, where the overwhelming focus on two-dimensional mammography left three-dimensional modalities underexplored despite their clinical relevance.

The frameworks were designed with explicit attention to real-world deployment constraints: computational efficiency enabling integration into existing clinical workflows, architectural modularity facilitating adaptation to diverse imaging protocols, and uncertainty modeling supporting clinical decision-making transparency. ABUSDet provides a practical solution for specialized medical centers equipped with ABUS systems, while TriNet offers state-of-the-art segmentation performance accessible to the ubiquitous mammography platforms deployed globally, including in resource-limited settings where breast cancer burden is rising most rapidly.

Both contributions emphasize deployability through their respective design philosophies. ABUSDet balances volumetric context capture with inference throughput

compatible with screening time constraints, achieving detection performance that meaningfully reduces radiologist review burden. TriNet refines tumor boundaries without encoder duplication and incorporates uncertainty-aware priors to enhance clinical trust, delivering boundary precision critical for treatment planning while respecting computational realities.

In summary, this thesis advances computational breast cancer imaging by systematically addressing detection efficiency for emerging three-dimensional modalities and segmentation precision for established two-dimensional platforms. By developing specialized architectures that respect clinical requirements, computational constraints, and workflow integration needs, this work provides practical pathways toward enhanced diagnostic accuracy and broader accessibility of advanced breast cancer screening technologies.

5.2 Future Work

While this thesis has made substantial contributions to breast tumor detection and segmentation, three critical research directions warrant further exploration to maximize clinical impact and methodological advancement.

5.2.1 Advancing Architectural Frameworks for Enhanced Performance

The architectural innovations presented in this thesis establish strong foundations for future enhancements. For ABUSDet, the transition from 2.5D to hybrid volumetric architectures represents a natural evolution, where selective three-dimensional processing could be dynamically allocated to suspicious regions through volumetric attention mechanisms while maintaining computational efficiency. Similarly, TriNet's progressive refinement paradigm could be extended to three-dimensional modalities such as Digital Breast Tomosynthesis (DBT) and volumetric ABUS data, where slice-wise processing with cross-slice prior propagation could leverage the same

architectural principles. Furthermore, as revealed by our stratified performance analysis, the detection of microscopic lesions (e.g., < 15 mm) remains challenging due to subtle sonographic contrasts and acoustic shadowing. Future iterations should incorporate targeted amplification mechanisms or multi-scale feature pyramids specifically optimized for early-stage, low-contrast tumor detection. Additionally, incorporating explicit uncertainty quantification through calibrated confidence estimates would enhance clinical trust by providing radiologists with reliable uncertainty maps that identify regions requiring particular scrutiny, thereby supporting more transparent clinical decision-making.

5.2.2 Leveraging Unlabeled Data through Advanced Learning Paradigms

The scarcity of expertly annotated medical images remains a fundamental bottleneck in developing robust diagnostic systems. Future research should prioritize self-supervised and semi-supervised learning strategies that can exploit the vast quantities of unlabeled clinical data available in hospital archives. For ABUS volumes, contrastive learning frameworks adapted to volumetric medical imaging could enable models to learn robust anatomical representations before fine-tuning on limited labeled datasets. For mammography segmentation, consistency regularization methods and pseudo-labeling strategies could substantially reduce the pixel-wise annotation burden. Furthermore, federated learning frameworks would enable collaborative model training across multiple institutions without sharing raw patient data, addressing both data scarcity and privacy concerns while developing more generalizable algorithms that benefit from diverse multi-center data distributions.

5.2.3 Clinical Integration and Real-World Deployment

The ultimate measure of success for computational breast imaging systems lies in their demonstrated impact on clinical care. Future work must prioritize seamless integration with existing healthcare infrastructure, including standardized in-

terfaces for Picture Archiving and Communication Systems (PACS) and comprehensive multi-center external validation across diverse scanner manufacturers and patient populations. Domain adaptation techniques should be investigated to ensure robust performance across institutional variations without requiring site-specific retraining. Additionally, the development of breast imaging-specific vision-language models that can generate structured, clinically meaningful reports from detection and segmentation outputs would enhance interpretability and facilitate clinical adoption. Such models could provide natural language descriptions of tumor characteristics grounded in visual evidence, enabling more effective communication between AI systems and clinical practitioners.

5.2.4 Synthesis

These future research directions share a unified vision: advancing computational breast cancer imaging systems that are not only technically sophisticated but also clinically practical and widely deployable. By pursuing targeted improvements in architectural design, learning efficiency, and clinical integration, the field can progress toward AI systems that deliver meaningful improvements in breast cancer screening outcomes across diverse healthcare settings, from resource-rich academic medical centers to underserved communities where diagnostic technology access remains limited. Success in these areas requires sustained collaboration between computer scientists, radiologists, and healthcare systems researchers, united by the commitment to developing technologies that are scientifically innovative, clinically effective, and equitably accessible.

Bibliography

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, February 2012.

[3] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, no. 1, p. 170177.

[4] P. Oza, U. Oza, R. Oza, P. Sharma, S. Patel, P. Kumar, and B. Gohel, "Digital mammography dataset for breast cancer diagnosis research (dmid) with breast mass segmentation analysis," *Biomedical Engineering Letters*, vol. 14, no. 2, pp. 317–330.

[5] M. Buda, A. Saha, R. Walsh, S. Ghate, N. Li, A. Świecicki, J. Y. Lo, and M. A. Mazurowski, "A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images," *JAMA Network Open*, vol. 4, no. 8, p. e2119100.

[6] A. Pawłowska, A. Ćwierz Pieńkowska, A. Domalik, D. Jaguś, P. Kasprzak, R. Matkowski, Fura, A. Nowicki, and N. Żołek, "Curated benchmark dataset

for ultrasound based breast lesion analysis," *Scientific Data*, vol. 11, no. 1, p. 148.

[7] A. Saha, M. R. Harowicz, L. J. Grimm, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski, "A machine learning approach to radiogenomics of breast cancer: A study of 922 subjects and 529 dce-mri features," *British Journal of Cancer*, vol. 119, no. 4, pp. 508–516.

[8] R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, "Cancer statistics, 2025," *CA: A Cancer Journal for Clinicians*, vol. 75, no. 1, pp. 10–45, January 2025.

[9] R. Strandberg, M. Illipse, K. Czene, P. Hall, and K. Humphreys, "Influence of mammographic density and compressed breast thickness on true mammographic sensitivity: A cohort study," *Scientific Reports*, vol. 13, no. 1, p. 14194, August 2023.

[10] D. L. Monticciolo, M. S. Newell, L. Moy, C. S. Lee, and S. V. Destounis, "Breast cancer screening for women at higher-than-average risk: Updated recommendations from the acr," *Journal of the American College of Radiology*, vol. 20, no. 9, pp. 902–914, September 2023.

[11] W. Ren, M. Chen, Y. Qiao, and F. Zhao, "Global guidelines for breast cancer screening: A systematic review," *The Breast*, vol. 64, pp. 85–99, August 2022.

[12] W. K. Moon, Y.-W. Shen, M. S. Bae, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Computer-aided tumor detection based on multi-scale blob detection algorithm in automated breast ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1191–1200, 2013.

[13] C.-M. Lo, R.-T. Chen, Y.-C. Chang, Y.-W. Yang, M.-J. Hung, C.-S. Huang, and R.-F. Chang, "Multi-dimensional tumor detection in automated whole

breast ultrasound using topographic watershed," *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1503–1511, 2014.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[16] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[17] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 574–584.

[18] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research, vol. 172. PMLR, 2022, pp. 1336–1348.

[19] Y. Zhou, H. Chen, Y. Li, S. Wang, L. Cheng, and J. Li, "3d multi-view tumor detection in automated whole breast ultrasound using deep convolutional neural network," *Expert Systems with Applications*, vol. 168, p. 114410, 2021.

[20] T.-C. Chiang, Y.-S. Huang, R.-T. Chen, C.-S. Huang, and R.-F. Chang, "Tumor detection in automated breast ultrasound using 3-d cnn and prioritized

candidate aggregation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 240–249, 2019.

[21] N. Wang, C. Bian, Y. Wang, M. Xu, C. Qin, X. Yang, T. Wang, A. Li, D. Shen, and D. Ni, "Densely deep supervised networks with threshold loss for cancer detection in automated breast ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 641–648.

[22] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, and A. S. Elmaghraby, "Connected-unets: a deep learning architecture for breast mass segmentation," *NPJ Breast Cancer*, vol. 7, no. 1, p. 151, 2021.

[23] M. Alkhaleefah, T.-H. Tan, C.-H. Chang, T.-C. Wang, S.-C. Ma, L. Chang, and Y.-L. Chang, "Connected-segnets: A deep learning model for breast tumor segmentation from x-ray images," *Cancers*, vol. 14, no. 16, p. 4030, August 2022.

[24] U. Oza, B. Gohel, P. Kumar, and P. Oza, "Presegmenter cascaded framework for mammogram mass segmentation," *International Journal of Biomedical Imaging*, vol. 2024, no. 1, p. 9422083.

[25] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[26] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference*

on acoustics, speech and signal processing (ICASSP). Ieee, 2020, pp. 1055–1059.

[27] J. Chen *et al.*, "Transunet: Rethinking the u-net architecture design for medical image segmentation," *Medical Image Analysis*, 2024.

[28] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24*. Springer, 2021, pp. 14–24.

[29] D. Liu, B. Wu, C. Li, Z. Sun, and N. Zhang, "Trend: A transformer-based encoder-decoder model with adaptive patch embedding for mass segmentation in mammograms," *Medical Physics*, vol. 50, no. 5, pp. 2884–2899.

[30] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[31] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[32] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. Pluim, U. Bauer, and B. H. Menze, "cldice-a novel topology-preserving loss function for tubular structure segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 560–16 569.

[33] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-aware transformers for skin lesion segmentation," in *International conference on med-*

ical image computing and computer-assisted intervention. Springer, 2021, pp. 206–216.

[34] S. Jin, H. Dai, J. Peng, Y. He, M. Zhu, X. Wu, and F. Peng, "A breast mass image segmentation method based on improved unet 3+ network," in *IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, pp. 1–7.

[35] P. Aliniya, M. Nicolescu, M. Nicolescu, and G. Bebis, "Improved loss function for mass segmentation in mammography images using density and mass size," *Journal of Imaging*, vol. 10, no. 1, p. 20.

[36] K. Anders Ericsson, "Deliberate practice and acquisition of expert performance: a general overview," *Academic emergency medicine*, vol. 15, no. 11, pp. 988–994, 2008.

[37] J. C. Santos, M. S. Santos, and P. H. Abreu, "Enhancing mammography: A comprehensive review of computer methods for improving image quality," *Progress in Biomedical Engineering*, vol. 6, no. 4, p. 042002, October 2024.

[38] F. Pesapane, C. Trentin, F. Ferrari, G. Signorelli, P. Tantrige, M. Montesano, C. Cicala, R. Virgoli, S. D'Acquisto, L. Nicosia, D. Origgi, and E. Cassano, "Deep learning performance for detection and classification of microcalcifications on mammography," *European Radiology Experimental*, vol. 7, no. 1, p. 69, November 2023.

[39] J. Pi, Y. Qi, M. Lou, X. Li, Y. Wang, C. Xu, and Y. Ma, "Fs-unet: Mass segmentation in mammograms using an encoder-decoder architecture with feature strengthening." *Computers in biology and medicine*, vol. 137, p. 104800, October 2021.

[40] S. Kulkarni and R. Rabidas, "Fully convolutional network for automated detection and diagnosis of mammographic masses," *Multimedia Tools and Applications*, vol. 82, no. 29, pp. 44 819–44 840, December 2023.

[41] S. Mohammadi and M. Ahmadi Livani, "Enhanced breast mass segmentation in mammograms using a hybrid transformer unet model," *Computers in Biology and Medicine*, vol. 184, p. 109432, January 2025.

[42] E. F. Conant, E. F. Beaber, B. L. Sprague, S. D. Herschorn, D. L. Weaver, T. Onega, A. N. A. Tosteson, A. M. McCarthy, S. P. Poplack, J. S. Haas, K. Armstrong, M. D. Schnall, and W. E. Barlow, "Breast cancer screening using tomosynthesis in combination with digital mammography compared to digital mammography alone: A cohort study within the prospr consortium," *Breast Cancer Research and Treatment*, vol. 156, no. 1, pp. 109–116, February 2016.

[43] S. Ciatto, N. Houssami, D. Bernardi, F. Caumo, M. Pellegrini, S. Brunelli, P. Tuttobene, P. Bricolo, C. Fantò, M. Valentini, S. Montemezzi, and P. Macaskill, "Integration of 3d digital mammography with tomosynthesis for population breast-cancer screening (storm): a prospective comparison study," *The Lancet Oncology*, vol. 14, no. 7, pp. 583–589, 2013.

[44] P. Skaane, A. I. Bandos, R. Gullien, E. B. Eben, U. Ekseth, U. Haakenaasen, M. Izadi, I. N. Jebsen, G. Jahr, M. Krager, L. T. Niklason, S. Hofvind, and D. Gur, "Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program," *Radiology*, vol. 267, no. 1, pp. 47–56, 2013.

[45] A. Sommer, S. Weigel, H.-W. Hense, J. Gerß, V. Weyer-Elberich, L. Kerschke, E. Nekolla, H. Lenzen, and W. Heindel, "Radiation exposure and screening

yield by digital breast tomosynthesis compared to mammography: results of the tosyma trial breast density related," *European Radiology*, vol. 35, no. 1, pp. 166–176, 2025.

[46] T. Hovda, S. Holen, K. Lång, J. L. Albertsen, H. Bjørndal, S. H. Brandal, K. K. Sahlberg, P. Skaane, P. Suhrke, and S. Hofvind, "Interval and consecutive round breast cancer after digital breast tomosynthesis and synthetic 2d mammography versus standard 2d digital mammography in breastscreen norway," *Radiology*, vol. 294, no. 2, pp. 256–264, 2020.

[47] W. A. Berg, E. A. Rafferty, S. M. Friedewald, C. B. Hruska, and H. Rahbar, "Screening algorithms in dense breasts: Ajr expert panel narrative review," *American Journal of Roentgenology*, vol. 216, no. 2, pp. 275–294, February 2021.

[48] L. Yang, S. Wang, L. Zhang, C. Sheng, F. Song, P. Wang, and Y. Huang, "Performance of ultrasonography screening for breast cancer: A systematic review and meta-analysis," *BMC Cancer*, vol. 20, no. 1, p. 499, December 2020.

[49] Z. Zhang, Y. Li, W. Wu, H. Chen, L. Cheng, and S. Wang, "Tumor detection using deep learning method in automated breast ultrasound," *Biomedical Signal Processing and Control*, vol. 68, p. 102677, 2021.

[50] Y. Guo, N. Li, C. Song, J. Yang, Y. Quan, and H. Zhang, "Artificial intelligence-based automated breast ultrasound radiomics for breast tumor diagnosis and treatment: A narrative review," *Frontiers in Oncology*, vol. 15, p. 1578991, May 2025.

[51] K. Oh, S. E. Lee, and E.-K. Kim, "3-d breast nodule detection on automated breast ultrasound using faster region-based convolutional neural networks and

u-net," *Scientific Reports*, vol. 13, p. 22625, 2023.

[52] X. Song, X. Lu, G. Fang, X. He, X. Fan, L. Cai, W. Jia, and Z. Wang, "Abusdet: A novel 2.5 d deep learning model for automated breast ultrasound tumor detection," *Applied Intelligence*, vol. 53, no. 21, pp. 26 255–26 269, 2023.

[53] X. Tao, Y. Cao, Y. Jiang, X. Wu, D. Yan, W. Xue, S. Zhuang, X. Yang, R. Huang, J. Zhang, and D. Ni, "Enhancing lesion detection in automated breast ultrasound using unsupervised multi-view contrastive learning with 3d detr," *Medical Image Analysis*, vol. 101, p. 103466, 2025.

[54] Y. Yan, P.-H. Conze, G. Quellec, M. Lamard, B. Cochener, and G. Coatrieux, "Two-stage multi-scale breast mass segmentation for full mammogram analysis without user intervention," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 746–757, April 2021.

[55] N. Aristokli, I. Polycarpou, S. Themistocleous, D. Sophocleous, and I. Mamais, "Comparison of the diagnostic performance of magnetic resonance imaging (mri), ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient's history: A review," *Radiography*, vol. 28, no. 3, pp. 848–856, 2022.

[56] H. Satake, S. Ishigaki, R. Ito, and S. Naganawa, "Radiomics in breast mri: Current progress toward clinical application in the era of artificial intelligence," *La radiologia medica*, vol. 127, no. 1, pp. 39–56.

[57] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[58] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," in *IEEE*

TPAMI, vol. 24, no. 7, 2002, pp. 971–987.

[59] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[60] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[61] B. Sahiner, H.-P. Chan, N. Petrick, J. Wei, M. A. Helvie, D. D. Adler, M. M. Goodsitt, and L. Hadjiiski, "Computerized characterization of masses detected on mammograms: Shape, edge sharpness, and texture," *Medical Physics*, vol. 25, no. 4, pp. 516–526, 1998.

[62] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," in *Proc. NIPS*, 2002, pp. 1307–1314.

[63] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE TPAMI*, vol. 27, no. 8, pp. 1226–1238, 2005.

[64] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[65] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 850–855.

[66] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE TPAMI*, vol. 6, no. 6, pp. 721–741, 1984.

[67] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proc. CVPR*, 2000, pp. 731–737.

[68] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, 2004.

[69] A. E. Burgess, F. L. Jacobson, and P. F. Judy, "On the froc methodology for performance evaluation of detection tasks," *Academic Radiology*, vol. 8, no. 4, pp. S154–S160, 2001.

[70] D. P. Chakraborty, "Analysis of location specific observer performance data: Validating the methodology," in *Proc. SPIE Medical Imaging*, vol. 7966, 2011, p. 79660K.

[71] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[72] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on hamilton–jacobi formulations," *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988.

[73] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

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

[75] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[76] S. Beucher and F. Meyer, "The morphological approach to segmentation: the watershed transformation," in *Mathematical Morphology in Image Processing*, E. R. Dougherty, Ed. New York: Marcel Dekker, 1993, pp. 433–481.

[77] R. Adams and L. Bischof, "Seeded region growing," *IEEE TPAMI*, vol. 16, no. 6, pp. 641–647, 1994.

[78] K. Dembrower, F. Wärnberg, M. Lárusdóttir *et al.*, "External validation and generalization of ai for screening mammography in a national population-based cohort," *The Lancet Digital Health*, vol. 5, no. 7, pp. e385–e394, 2023.

[79] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[80] S. E. Hickman, R. Woitek, E. P. V. Le, Y. R. Im, C. Mouritsen Luxhøj, A. I. Aviles-Rivero, G. C. Baxter, J. W. MacKay, and F. J. Gilbert, "Machine learning for workflow applications in screening mammography: systematic review and meta-analysis," *Radiology*, vol. 302, no. 1, pp. 88–104, 2022.

[81] I. Sechopoulos and R. M. Mann, "Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: A review," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e200430, 2021.

[82] T. Kyono *et al.*, "Prospective validation of an artificial intelligence system for mammography screening workflow," *NPJ Digital Medicine*, vol. 3, no. 1, p. 66, 2020.

[83] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[84] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[85] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[86] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[87] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sanchez, R. Mann, G. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical Image Analysis*, vol. 35, pp. 303–312, 2017.

[88] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Scientific reports*, vol. 8, no. 1, p. 4165, 2018.

[89] R. Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[90] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[91] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. ICCV*, 2017, pp. 2961–2969.

[92] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[93] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016.

[94] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[95] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018.

[96] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[97] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[98] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[99] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 109–119.

[100] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 000–16 009.

[101] X. Luo, J. Chen, T. Song, Y. Chen, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation through dual-task consistency," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 8801–8809.

[102] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, p. 102517, 2022.

[103] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "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 (CVPR)*, 2022, pp. 20 730–20 740.

[104] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *BrainLesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer, 2022, pp. 272–284.

[105] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3d tumor segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 162–172.

[106] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Unetr++: delving into efficient and accurate 3d medical image segmentation,"

IEEE Transactions on Medical Imaging, vol. 43, no. 9, pp. 3377–3390, 2024.

[107] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, and Y. Zheng, "mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 107–117.

[108] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[109] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 13438. Springer, 2022, pp. 35–45.

[110] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "Medsegdiff-v2: Diffusion-based medical image segmentation with transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 6, 2024, pp. 6030–6038.

[111] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[112] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine learning for healthcare conference*. PMLR, 2022, pp. 2–25.

[113] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle *et al.*, "Making the most of text semantics to improve biomedical vision–language processing," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[114] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 965–10 975.

[115] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7086–7096.

[116] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[117] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[118] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, H. Sun, J. He, S. Zhang, M. Zhu, and Y. Qiao, "Sam-med2d," 2023.

[119] H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, J. Chen, Y. Su, Z. Huang, Y. Shen *et al.*, "Sam-med3d: towards general-purpose segmentation models for volumetric medical images," in *European Conference on Computer Vision*. Springer, 2024, pp. 51–67.

[120] Y. Li, M. Hu, and X. Yang, "Polyp-sam: Transfer sam for polyp segmentation," in *Medical imaging 2024: computer-aided diagnosis*, vol. 12927. SPIE,

2024, pp. 749–754.

[121] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First Conference on Language Modeling*, 2024.

[122] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

[123] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2024, pp. 578–588.

[124] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[125] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[126] A. Migowski, "Early detection of breast cancer and the interpretation of results of survival studies/a deteccao precoce do cancer de mama e a interpretacao dos resultados de estudos de sobrevida," *Ciencia saude coletiva*, vol. 20, no. 4, pp. 1309–1310, 2015.

[127] C. Duggan, D. Trapani, A. M. Ilbawi, E. Fidarova, M. Laversanne, G. Curigliano, F. Bray, and B. O. Anderson, "National health system characteristics, breast cancer stage at diagnosis, and breast cancer mortality: a population-based analysis," *The Lancet Oncology*, vol. 22, no. 11, pp. 1632–1642, 2021.

[128] K. Zamora, E. Allen, and B. Hermecz, "Contrast mammography in clinical practice: Current uses and potential diagnostic dilemmas," *Clinical Imaging*, vol. 71, pp. 126–135, 2021.

[129] E. Devolli-Disha, S. Manxhuka-Kërliu, H. Ymeri, and A. Kutllovci, "Comparative accuracy of mammography and ultrasound in women with breast symptoms according to age and breast density," *Bosnian journal of basic medical sciences*, vol. 9, no. 2, p. 131, 2009.

[130] C. Kuhl, "Mri of breast tumors," *European radiology*, vol. 10, no. 1, pp. 46–58, 2000.

[131] R. M. Mann, C. Balleyguier, P. A. Baltzer, U. Bick, C. Colin, E. Cornford, A. Evans, E. Fallenberg, G. Forrai, M. H. Fuchsjäger *et al.*, "Breast mri: Eusobi recommendations for women's information," *European radiology*, vol. 25, no. 12, pp. 3669–3678, 2015.

[132] M. Zanotel, I. Bednarova, V. Londero, A. Linda, M. Lorenzon, R. Girometti, and C. Zuiani, "Automated breast ultrasound: basic principles and emerging clinical applications," *La radiologia medica*, vol. 123, no. 1, pp. 1–12, 2018.

[133] T. Tan, B. Platel, R. Mus, L. Tabár, R. M. Mann, and N. Karssemeijer, "Computer-aided detection of cancer in automated 3-d breast ultrasound," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1698–1706, 2013.

[134] Y. Wang, N. Wang, M. Xu, J. Yu, C. Qin, X. Luo, X. Yang, T. Wang, A. Li, and D. Ni, "Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 866–876, 2020.

[135] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp.

7132–7141.

[136] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data knowledge engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[137] G. Ayana, K. Dese, and S.-w. Choe, "Transfer learning in breast cancer diagnoses via ultrasound imaging," *Cancers*, vol. 13, no. 4, p. 738, 2021.

[138] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[139] D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (froc) data," *Medical physics*, vol. 16, no. 4, pp. 561–568, 1989.

[140] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "Auc: a misleading measure of the performance of predictive distribution models," *Global ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.

[141] X. Hou, Y. Bai, Y. Xie, and Y. Li, "Mass segmentation for whole mammograms via attentive multi-task learning framework," *Physics in Medicine Biology*, vol. 66, no. 10, p. 105015, 2021.

[142] J. Chowdary, P. Yogarajah, P. Chaurasia, and V. Guruviah, "A multi-task learning framework for automated segmentation and classification of breast tumors from ultrasound images," *Ultrasonic imaging*, vol. 44, no. 1, pp. 3–12, 2022.

[143] H. Sun, C. Li, B. Liu, Z. Liu, M. Wang, H. Zheng, D. D. Feng, and S. Wang, "Aunet: attention-guided dense-upsampling networks for breast mass segmen-

tation in whole mammograms," *Physics in Medicine Biology*, vol. 65, no. 5, p. 055005, 2020.

[144] C. Xu, Y. Qi, Y. Wang, M. Lou, J. Pi, and Y. Ma, "Arf-net: An adaptive receptive field network for breast mass segmentation in whole mammograms and ultrasound images," *Biomedical Signal Processing and Control*, vol. 71, p. 103178.

[145] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo, "Multi-compound transformer for accurate biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 326–336.

[146] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[147] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[148] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[149] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[150] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[151] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020, pp. 558–564.

[152] J. Zhuang, "Laddernet: Multi-path networks based on u-net for medical image segmentation," *arXiv preprint arXiv:1810.07810*, 2018.

[153] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[154] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[155] Y. Yang, X. Guo, Y. Pan, P. Shi, H. Lv, and T. Ma, "Uncertainty quantification in medical image segmentation with multi-decoder u-net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*, ser. Lecture Notes in Computer Science. Springer, 2022, vol. 12963, pp. 570–577.

[156] M. Berman, T. Remez, A. Vedaldi, I. Kokkinos, N. Berman, T. Remez, and S. Bagon, "Sharing decoders: Network fission for multitask pixel prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[157] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[158] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.

[159] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[160] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021.

[161] Y. Gao *et al.*, "Utnet: A hybrid transformer architecture for medical image segmentation," in *MICCAI*, ser. LNCS, 2021.

[162] J. M. J. Valanarasu and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *MICCAI*, ser. LNCS, 2021.

[163] Z. Huang *et al.*, "Missformer: An effective medical image segmentation transformer," *IEEE Transactions on Medical Imaging*, 2022.

[164] Q. Li, J. Xu, R. Yuan, Y. Zhang, and R. Feng, "Scsgnet: Spatial-correlated and shape-guided network for breast mass segmentation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5.

[165] W. Wang, C. Chen, M. Ding, J. Li, H. Yu, and S. Zha. TransBTS: Multimodal Brain Tumor Segmentation Using Transformer.

[166] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 272–284.

[167] C. Zhu, X. Chai, Y. Xiao, X. Liu, R. Zhang, Z. Yang, and Z. Wang, "Swin-Net: A Swin-Transformer-Based Network Combing with Multi-Scale Features for Segmentation of Breast Tumor Ultrasound Images," *Diagnostics*, vol. 14, no. 3, p. 269.

[168] H. Zhang, J. Lian, Z. Yi, R. Wu, X. Lu, P. Ma, and Y. Ma, "Hau-net: Hybrid cnn-transformer for breast ultrasound image segmentation," *Biomedical Signal Processing and Control*, vol. 87, p. 105427, 2024.

[169] H. Yang and D. Yang, "Cswin-pnet: A cnn-swin transformer combined pyramid network for breast lesion segmentation in ultrasound images," *Expert Systems with Applications*, vol. 213, p. 119024, 2023.

[170] A. Iqbal and M. Sharif, "Bts-st: Swin transformer network for segmentation and classification of multimodality breast cancer images," *Knowledge-Based Systems*, vol. 267, p. 110393, 2023.

[171] Y. Zhong and Y. Piao, "Gated axial transformer network for mass segmentation in mammographic," in *Fourth International Conference on Computer Science and Communication Technology (ICCSCT 2023)*, vol. 12918. SPIE, 2023, pp. 141–149.

[172] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," *The Tenth International Conference on Learning Representations*, 2022.

[173] ——, "Separable self-attention for mobile vision transformers," *arXiv preprint arXiv:2206.02680*, 2022.

[174] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 14 420–14 430.

[175] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *European conference on computer vision*. Springer, 2022, pp. 68–85.

[176] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "Fastvit: A fast hybrid vision transformer using structural reparameterization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 5785–5795.

[177] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. A. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[178] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[179] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[180] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*,

vol. 31, 2018.

[181] J. E. Pearson, "Complex patterns in a simple system," *Science*, vol. 261, no. 5118, pp. 189–192, 1993.

[182] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[183] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[184] S. Perera, Y. Erzurumlu, D. Gulati, and A. Yilmaz, "Mobileunetr: a lightweight end-to-end hybrid vision transformer for efficient medical image segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 281–299.

[185] Y. Wang, S. Wang, J. Chen, and C. Wu, "Whole mammographic mass segmentation using attention mechanism and multiscale pooling adversarial network," *Journal of Medical Imaging*, vol. 7, no. 05.

[186] R. R. N, V. R, E. N, and N. Ramesh, "Deeply supervised for mass segmentation in digital mammograms," *International Journal of Imaging Systems and Technology*, vol. 31, no. 1, pp. 59–71.

[187] Y. Yan, P.-H. Conze, G. Quellec, M. Lamard, B. Cochener, and G. Coatrieux, "Two-stage multi-scale mass segmentation from full mammograms," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1628–1631.

[188] W. Zhao, M. Lou, Y. Qi, Y. Wang, C. Xu, X. Deng, and Y. Ma, "Adaptive channel and multiscale spatial context network for breast mass segmentation in full-field mammograms," *Applied Intelligence*, vol. 51, no. 12, pp. 8810–8827.

[189] G. J. Chowdary and P. Yoagarajah, "EU-Net: Enhanced U-shaped Network for Breast Mass Segmentation," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–11.