ANKARA YILDIRIM BEYAZIT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEEP LEARNING FOR BREAST MASS SEGMENTATION: A REGION-OF-INTEREST FOCUSED APPROACH USING ETDP-U^2-NET

M.Sc. Thesis by

Pakize Sümeyye SÖYLEMEZ

Department of Computer Engineering

July, 2025

ANKARA

DEEP LEARNING FOR BREAST MASS SEGMENTATION: A REGION-OF-INTEREST FOCUSED APPROACH USING ETDP-U^2-NET

A Thesis Submitted to the

Graduate School of Natural And Applied Sciences of

Ankara Yildirim Beyazit University

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering, Department of Computer Engineering

by

Pakize Sümeyye SÖYLEMEZ

July, 2025

ANKARA

ETHICAL DECLARATION

I hereby declare that, in this thesis which has been prepared in accordance with the Thesis Writing Manual of Graduate School of Natural and Applied Sciences,

- All data, information and documents are obtained in the framework of academic and ethical rules,
- All information, documents and assessments are presented in accordance with scientific ethics and morals,
- All the materials that have been utilized are fully cited and referenced,
- No change has been made on the utilized materials,
- All the works presented are original,

and in any contrary case of above statements, I accept to renounce all my legal rights.

Date: 2025, 1 July **Signature:** ______________________________

Name & Surname: Pakize Sümeyye SÖYLEMEZ

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor, Prof. Dr. M. Fatih Demirci, for his continuous support, insightful guidance, and encouragement throughout the course of my graduate studies and this thesis. His expertise, patience, and mentorship have been invaluable in helping me navigate both academic and technical challenges.

I would like to thank my colleagues and fellow researchers at AYBU for fostering a collaborative and motivating environment. Their feedback, friendship, and solidarity have enriched both my academic and personal experience.

On a more personal note, I am deeply thankful to my parents, my sibling, and my close friends among the research assistants, for their unwavering emotional and mental support. Their constant encouragement and belief in me have made even the most difficult moments manageable.

July, 2025 **Pakize Sümeyye SÖYLEMEZ**

DEEP LEARNING FOR BREAST MASS SEGMENTATION: A REGION-OF-INTEREST FOCUSED APPROACH USING ETDP-U^2-NET

ABSTRACT

Accurate segmentation of breast masses in mammographic images plays a critical role in early breast cancer detection. In this thesis, we propose a novel deep learning architecture, ETDP-U^2-Net, tailored for mass segmentation using ROI-cropped grayscale mammograms from the CBIS-DDSM dataset. The model integrates edge and texture-aware pathways with enhanced skip connections to improve the delineation of subtle tumor boundaries. Extensive experiments under both non-augmented and augmented training regimes show that ETDP-U^2-Net achieves competitive Dice and IoU scores while maintaining a lightweight design with only 6.54 million parameters. Notably, unlike many prior studies, this work avoids test-time augmentation and potential data leakage by applying augmentation solely to the training set. The results demonstrate that ETDP-U^2-Net not only surpasses many heavier architectures in terms of performance-to-parameter efficiency but also adheres to rigorous experimental standards. This study contributes a robust and efficient segmentation approach that holds promise for integration into computer-aided diagnosis systems in clinical settings.

Keywords: Breast cancer segmentation, ETDP-U^2-Net, CBIS-DDSM, ROI-cropped mammogram, medical image analysis, deep learning.

ROI-KIRPILMIŞ MAMOGRAMLAR ÜZERİNDE ETDP-U^2-NET TABANLI DERİN ÖĞRENME İLE MEME KİTLE SEGMENTASYONU

ÖZ

Mamografi görüntülerinde meme kitlesinin doğru şekilde segmentasyonu, erken evre meme kanseri tespitinde hayati bir rol oynamaktadır. Bu tez çalışmasında, CBIS-DDSM veri kümesinden alınan gri tonlamalı ROI (Region of Interest) kırpılmış mamogramlar üzerinde kitle segmentasyonu gerçekleştirmek amacıyla geliştirilen yeni bir derin öğrenme mimarisi olan ETDP-U^2-Net önerilmektedir. Model, kenar ve dokuya duyarlı yolları geliştirilmiş atlama bağlantılarıyla birleştirerek tümör sınırlarının daha hassas bir şekilde belirlenmesini sağlar. Hem artırımsız (non-augmented) hem de artırımlı (augmented) eğitim senaryoları altında yapılan kapsamlı deneyler, yalnızca 6.54 milyon parametreye sahip hafif tasarıma rağmen modelin rekabetçi Dice ve IoU skorları elde ettiğini ortaya koymaktadır. Özellikle, önceki çalışmaların aksine, bu tezde yalnızca eğitim verisine artırma uygulanmış, test verisine herhangi bir işlem uygulanmayarak olası veri sızıntısı engellenmiştir. Sonuçlar, ETDP-U^2-Net'in parametre-verimlilik açısından birçok daha ağır mimariyi geride bıraktığını ve aynı zamanda titiz deneysel standartlara bağlı kaldığını göstermektedir. Bu çalışma, klinik ortamlarda bilgisayar destekli tanı sistemlerine entegre edilebilecek sağlam ve verimli bir segmentasyon yaklaşımı sunmaktadır.

Anahtar kelimeler: Meme kanseri segmentasyonu, ETDP-U^2-Net, CBIS-DDSM, ROI kırpılmış mamogram, tıbbi görüntü analizi, derin öğrenme.

CONTENTS

NOMENCLATURE

Roman Letter Symbols

D	Dice Coefficient, overlap metric
F_2	F_2 Score: $\frac{5PR}{4P+R}$
I	Intersection over Union (IoU), overlap metric
P	Precision: TP / (TP + FP)
R	Recall: TP / (TP + FN)

Greek Letter Symbols

α	Channel-wise attention scaling factor in SE blocks
β	Atrous dilation rate in ASPP
δ	Edge enhancement function
γ	Weighting coefficient for residual connections
$\mathcal{A}$	Attention map
$\mathcal{F}_{\mathrm{ASPP}}$	Features from ASPP module
ψ	Attention gating function output

Acronyms

ASPP	Atrous Spatial Pyramid Pooling for multi-scale feature extraction
AU^2-Net	Attention Cascaded U^2-Net
CBIS-DDSM	Curated Breast Imaging Subset of the Digital Database for Screening Mammography
DPCA-U^2-Net	Dual-Path Cross-Attention U^2-Net
DPTrans-U^2-Net	Dual-Path Transformer-Enhanced U^2-Net
ETDP-U^2-Net	Edge-Texture Dual-Path U^2-Net (proposed)
F$_2$ Score	Harmonic mean of precision and recall, weighted toward recall
ROI	Region of Interest
SE Block	Squeeze-and-Excitation block for channel-wise attention
U-Net	A fully convolutional network for biomedical image segmentation
U^2-Net	Nested U-Net architecture with Residual U-blocks

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Breast cancer stands as one of the most significant global health threats, affecting millions of women each year and accounting for a substantial share of cancer-related deaths. According to the World Health Organization, nearly 2.3 million women were newly diagnosed with breast cancer in 2020, leading to approximately 685,000 deaths globally [1]. These statistics, supported by the GLOBOCAN 2020 study [2], reflect the widespread prevalence and mortality associated with this disease. Although early detection greatly increases the chances of survival, many individuals—especially those living in low-income regions—are still diagnosed at later stages, largely due to the lack of accessible and organized screening programs.

Mammography remains a cornerstone in the early detection of breast cancer, offering a non-invasive and reliable method for identifying tumors before clinical symptoms emerge [3]. In high-income countries, the inclusion of structured screening programs within public health systems has contributed to earlier diagnoses and a noticeable decline in mortality rates [4]. Leading medical organizations—such as the American College of Obstetricians and Gynecologists and the U.S. Preventive Services Task Force—recommend initiating routine screenings between the ages of 40 and 50, depending on individual risk profiles [5, 6]. By contrast, in many lower-income settings, the lack of accessible screening services often results in late-stage diagnoses, reducing the likelihood of successful treatment and long-term survival.

Although mammography plays a vital role in detecting breast cancer, interpreting these images remains a complex task. A major challenge stems from the low contrast typically found in mammograms, particularly when dense breast tissue makes abnormalities harder to detect. In addition, many tumors are small and subtle, forcing radiologists to carefully examine details such as shape, margin, and density—features that can vary not only between patients but also across different imaging views of the same case [7]. To help address these difficulties, Computer-Aided Diagnosis (CAD) systems have been developed. Leveraging deep learning techniques, these tools aim to automatically analyze mammograms, thereby improving diagnostic accuracy, ensuring greater consistency among readers, and easing the clinical workflow [8].

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) is widely recognized as a standard benchmark for developing and evaluating Computer-Aided Diagnosis (CAD) systems [9]. It provides grayscale mammograms along with pixel-level Region of Interest (ROI) annotations, making it especially useful for training and validating segmentation models. However, working with CBIS-DDSM also presents several challenges. Lesion sizes vary widely, annotations may lack consistency, and image contrast is often low—all of which can make it difficult to achieve robust and generalizable model performance.

Breast mass segmentation is a key component in the success of Computer-Aided Diagnosis (CAD) systems, yet it remains technically demanding. The low contrast typical of mammographic images, combined with the variability in lesion appearance from one patient to another, makes precise boundary detection especially challenging. This difficulty is even more pronounced in the case of small lesions, which, despite their clinical relevance, are often missed by segmentation models due to their limited pixel footprint. Although conventional architectures like U-Net [10] have formed the foundation for many medical segmentation pipelines, their performance can deteriorate under these complex imaging conditions.

In recent years, researchers have proposed a variety of improvements to address the shortcomings of traditional segmentation architectures. One notable example is Attention U-Net, which uses attention gates to help the model concentrate on the most relevant parts of the image [11]. U-Net++, on the other hand, improves multiscale learning by introducing nested and densely connected skip pathways [12]. More recently, hybrid models like HTU-Net [13] and transformer-based approaches such as MSMV-Swin [14] have been developed to better capture long-range dependencies and contextual features. Alongside these architectural advances, researchers have also turned their attention to loss functions. Focal loss [15] addresses the issue of class imbalance by focusing learning on harder examples, while boundary-aware loss terms [16] aim to sharpen segmentation near lesion edges. Together, these developments have led to more accurate and robust models, particularly in complex medical imaging scenarios.

In this thesis, I introduce three new deep learning architectures—DPCA-U^2-Net, ETDP-U^2-Net, and DPTrans-U^2-Net—each developed to address the unique challenges of segmenting grayscale mammograms. These models were designed to address the shortcomings of existing methods by combining dual-path encoders, residual connections, attention mechanisms, and transformer bottlenecks in a unified framework. To support more accurate and stable training, a tailored loss function is used—bringing together Dice, Focal, and boundary-aware terms. Equally important, all data augmentation was performed only after the dataset was split into training and test sets, in order to avoid data leakage and ensure that the experimental results truly reflect real-world performance.

The core motivation behind this research lies in developing segmentation models that are not only accurate but also practical for real-world clinical use. Many existing studies report impressive results under controlled conditions, yet often fail to account for everyday limitations such as restricted computational resources and the scarcity of annotated data. This thesis focuses on pixel-level segmentation without using diagnostic labels—a deliberate choice to keep the models lightweight, flexible, and better suited for deployment in real-world clinical settings.

CHAPTER 2

LITERATURE REVIEW

Accurately detecting breast masses in mammograms plays a vital role in early cancer diagnosis and better treatment outcomes. As breast cancer continues to impact more people worldwide, there has been growing interest in computer-aided diagnosis (CAD) systems—particularly those based on deep learning—that support radiologists in making more informed decisions. *Even so*, challenges remain. Low image contrast, variation between patients, the small size of many lesions, and inconsistent evaluation practices all continue to hinder model performance. This chapter reviews recent research in the field, with a focus on dataset usage, architectural developments, and evaluation strategies. It also outlines where this thesis fits within that landscape and how it aims to move the field forward.

One of the most widely used public datasets for mammographic segmentation is the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM). It offers high-resolution grayscale mammograms alongside detailed Region of Interest (ROI) masks. Developed specifically for research purposes, the dataset includes pixel-level annotations, making it a strong candidate for training and evaluating supervised segmentation models. However, despite its popularity, the dataset is still frequently misused in the literature—particularly through improper data splitting—which can lead to data leakage and inflated performance metrics. For example, Shen et al. [45] did not separate the "mass" subset of CC and MLO view from the "calcification" subset, leading to different distributions in the training and testing set. Similarly, Li et al. [17] performed data augmentation post-split of the dataset, bringing in information the training set contained into the test set through the augmented variations. Such methodological shortcomings call into question the validity of performance appraisals.

In this Thesis, a stricter approach is followed, where we denoise the CBIS-DDSM dataset and only keep the 'mass' subset. In addition, the dataset was stratified before any augmentation so that the test set contained no synthetic or augmented versions of any training images. This detailed care in data splitting criterions prevents one of the most major issues in previous works and guarantees that segmentation performance will be unbiased.

Existing works can be divided into two types of segmentation, including the full-image segmentation and the ROI-based segmentation. The difference between those two methods is that the former one uses a segmentation model that is first trained on the whole mammogram, while the latter focuses the task on a cropped area of the lesion. ROI-based approaches are recently widely used due to their computational effectiveness and superior localized results. Zhou et al. [18] demonstrated that, by training with ROI images, the segmentation accuracy was increased by concentrating the model on the lesion area and reducing background clutter. Similarly, Liu et al. [19] demonstrated that attending more to local context improves the performance of deep networks in detecting lesion borders, particularly for small abnormalities.

State-of-the-art methods in the literature for segmentation primarily belong to two types of methods: full-According to our survey, although the original U-Net architecture [10] is significantly popular across a great number of segmentation tasks, it does not perform well over mammograms mainly for grayscale nature of the images, low contrast, and complex texture of the tissue. Some alternatives have been suggested to circumvent them. Oktay et al. AttU-Net [20] was developed using attention gates that allow the model to focus on more important regions. Zhou et al. [12] proposed a U-Net++ with nested and dense skip connections to enhance feature propagation and the multiscale representation. These improved on the performances of those models but remained limited in the ability to segment small or low contrast lesions due to the use of local receptive fields.

In order to alleviate those limitations, some more modern structures adopt dedicated modules. Zhang et al. [21] introduced residual U-blocks (RSUs) forplugging high-resolution features in deep feature extraction. These were the most competent RSUs in addressing multiscale segmentation problems. Hu et al. [22]proposed Squeeze-and- Excitation (SE) blocks that learn to recalibrate feature-channel responses in a channel-wise manner, leading to better representation capability for convolutional layers.

Based on these ingredients, this work presents ETDP-U^2-Net and DPTrans-U^2-Net. The ETDP-U^2-Net proposed dual-path encoders to independently capture edge and texture features, which then are integrated via the cross-attention mechanisms. Each path leads to its RSUs and SE blocks, resulting in a strong multi-representation framework dealing with various lesion morphologies. The network is refined by the DPTrans-U^2-Net which introduces transformer-based modules at the bottleneck to capture global context and semantic dependencies, which are essential to segmentation of small lesions.

Moreover, a third model, DPCA-U^2-Net, was developed to maintain performance while reducing computational overhead. The proposed model preserves the dual-path cross-attention structure and continues to employ squeeze-and-excitation (SE) blocks and deep supervision. What sets this model apart is its streamlined design, which simplifies the architecture while still delivering strong performance. By striking a balance between efficiency and accuracy, it manages to achieve results comparable to more complex alternatives.

In addition to architectural improvements, the design of the loss function plays a key role in training effective segmentation models. Although binary cross-entropy and Dice loss remain widely used, they often fall short in the presence of severe class imbalance—an issue commonly seen in medical imaging. To mitigate this, Lin et al. [15] proposed Focal Loss, which down-weights easy examples and directs the model's attention toward harder cases. Building on this, Hasan et al. [16] introduced a hybrid loss that integrates Dice, Focal, and boundary-aware terms to improve accuracy around lesion edges. Inspired by these efforts, this thesis adopts a similar composite loss strategy to better capture subtle lesion details, particularly in small or low-contrast regions.

Segmentation models are typically evaluated using metrics like Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). While these are essential for quantifying overlap between predicted and ground truth regions, relying on them alone can miss other clinically important aspects of model performance. For example, Gao et al. [23] reported only Dice scores, which offer limited perspective on how well a model generalizes across diverse patient cases. To provide a more complete evaluation, this study also considers Precision, Recall, and F2-Score alongside Dice and IoU. Together, these metrics offer a broader view of both the clinical reliability and real-world applicability of the proposed models.

Keeping the data clean and the annotations consistent is just as important as having a good model. If these parts are ignored, the results can be misleading. For example, Gupta et al. [24] did not separate patients between the training and test sets, which caused some overlap and may have made their model seem more accurate than it really was. In another case, Wang et al. [25] found that some masks didn't line up properly with the mammogram images, which added noise and made learning harder for the model. This research ensures that each patient's data appears only in one subset and verifies annotation consistency, preserving the integrity of model evaluations.

Hybrid models that combine CNNs with transformers have gained significant attention in recent works. Mohammadi et al. [13] proposed HTU-Net, where attention from transformers and convolution from CNN are used together to learn the right set of context. Chen et al. [14] introduced MSMV-Swin, a model based on the multi-view Swin Transformer which achieved state-of-the-art results on CBIS-DDSM. However, these approaches are often high precision, yet associated with a large number of parameters and high computational cost, making it difficult to apply in real-time clinical field. The models proposed in this thesis chase the tradeoff between accuracy, efficiency and ability to deploy.

The role of pathology labels in training is another critical issue to consider. A number of studies add diagnostic annotations (benign vs. malignant) during the segmentation step, as auxiliary outputs or multi-task objectives. Zhu et al. [26] however, advised against such design saying it can make the algorithm less generalizable when no diagnostic label is given. The models in this work consider only pixel-level, rather than categorical (pathology label) segmentation. This is done to make the model applicable in practice since such labels are frequently unavailable.

In conclusion, recent advances in mammographic breast mass segmentation have significantly improved model accuracy, yet several challenges remain unresolved. These include consistent data handling, small lesion detection, robust evaluation strategies, and model generalizability. The literature indicates that improvements in architecture, loss function design, and evaluation methodology are all necessary for clinically viable segmentation tools. The proposed ETDP-U^2-Net, DPCA-U^2-Net, and DPTrans-U^2-Net models address these gaps comprehensively. By combining dual-path feature extraction, attention mechanisms, transformer bottlenecks, and rigorous evaluation protocols, they advance the state of the art in breast mass segmentation, providing both methodological innovation and practical relevance in the field of medical image analysis.

CHAPTER 3

MATERIALS AND METHODS

This chapter presents the datasets, preprocessing steps, and experimental pipeline followed throughout this thesis. All models in this study were trained and tested using the publicly available CBIS-DDSM dataset. Throughout the process, special attention was paid to some of the common problems in medical image segmentation, such as avoiding data leakage, improving visibility in low-contrast areas, and making sure that small lesions were properly represented in the data.

3.1 Dataset Description

For evaluation, this study uses the Curated Breast Imaging Subset of the DDSM (CBIS-DDSM) [27], a well-known public dataset widely used in mammography research and computer-aided diagnosis (CAD). CBIS-DDSM is a curated version of the original DDSM dataset and contains high-resolution grayscale mammograms captured using digital mammography systems. Each patient case includes left and right breast views, along with pixel-level Region of Interest (ROI) masks that have been confirmed by pathology and annotated by expert radiologists.

The mammogram images are stored in LJPEG format, a lossless variant of JPEG that preserves full image quality. For practical use with deep learning tools and standard image processing libraries, these images are typically converted to `.jpeg` format, which still provides sufficient quality for training models.

This work focuses only on the “mass” subset of CBIS-DDSM, which includes benign and malignant breast masses, but excludes calcification-type abnormalities. This decision is based on earlier studies [17, 28] that showed mixing different lesion types can disrupt training and lead to overly optimistic results. To further simplify the classification task, all “benign without callback” cases were grouped under the “benign” label. This step helped maintain consistency across binary class labels and followed the best practices recommended in recent literature.

Each mammographic case in the dataset includes two standard imaging views—craniocaudal (CC) and mediolateral oblique (MLO)—for both the left and right breasts. In this study, both views were retained to better reflect real-world clinical scenarios and to increase morphological variability within the dataset. The mammograms are grayscale images with varying spatial resolutions, while the corresponding ROI masks are binary matrices in which lesion areas are marked with a value of 255 and background regions with 0.

To avoid data leakage and support a reliable evaluation of model generalizability, the dataset was split into training and test sets based on patient-level separation. This ensures that no images from the same patient appear in both subsets. Importantly, all data augmentation steps—including flipping, rotation, scaling, and contrast enhancement—were performed exclusively on the training set after the split. This approach helps preserve the authenticity of test data and supports fair, unbiased performance assessment.

Table 3.1 Train set composition by view and diagnosis (CBIS-DDSM mass subset).

View	Diagnosis	Count
CC	Benign	273
CC	Malignant	334
MLO	Benign	304
MLO	Malignant	407
Total		**1318**

Table 3.2 Test set composition by view and diagnosis (CBIS-DDSM mass subset).

View	Diagnosis	Count
CC	Benign	94
CC	Malignant	83
MLO	Benign	100
MLO	Malignant	101
Total		**378**

Following a quality control process that excluded corrupted or misaligned samples, the final dataset consists of 1696 image-mask pairs. Of these, 1318 images were allocated to the training set and 378 to the test set. Detailed distributions by imaging view (CC vs.

MLO) and diagnosis (benign vs. malignant) are presented in Table 3.1 and Table 3.2, respectively. These tables demonstrate a balanced and diverse dataset that supports robust deep learning model development.

3.1.1 Visual Sample Overview

To provide a clear understanding of the imaging and annotation characteristics in the CBIS-DDSM mass dataset, Figure 3.1 showcases representative examples from both the training and test sets. For each case, the original grayscale mammogram is displayed alongside its corresponding binary ROI mask annotated by expert radiologists.

Images 1 and 3 are taken from the test set, while Images 2 and 4 represent samples from the training set. The cases illustrate the wide variability in lesion size, shape, and contrast—factors that necessitate robust preprocessing and generalizable segmentation architectures. All images have been resized to a uniform resolution of 256 × 256 pixels to ensure consistency in visual comparison and model input preparation.

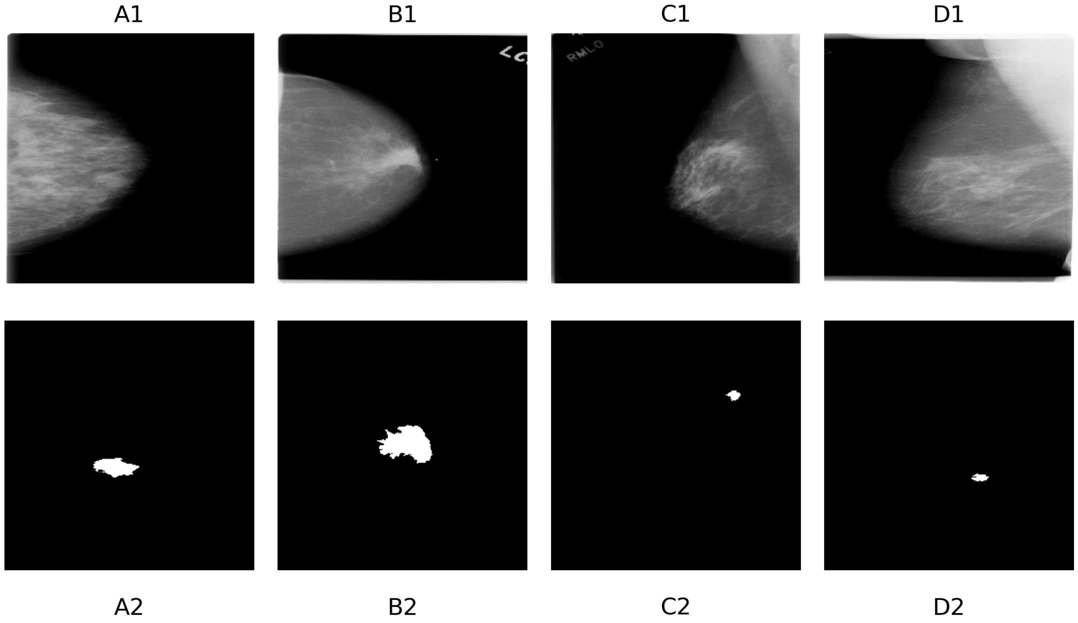

Figure 3.1 Representative samples selected from the CBIS-DDSM mass dataset. A1: Benign case from the test set with a craniocaudal (CC) view. A2: Corresponding region-of-interest (ROI) mask. B1: Malignant case from the training set with a CC view. B2: Corresponding ROI mask. C1: Benign case from the test set with a mediolateral oblique (MLO) view. C2: Corresponding ROI mask. D1: Benign case from the training set with an MLO view. D2: Corresponding ROI mask. All images were resized to 256 × 256 pixels for consistency in visualization.

3.2 Preprocessing

In medical image analysis, especially in mammography, preprocessing plays a vital role in enhancing data quality and preparing images for robust segmentation [29]. The mammograms in this study, originating from the CBIS-DDSM mass subset [27], are characterized by low global contrast, high anatomical variability, and frequent presence of irrelevant background structures. A structured preprocessing pipeline is therefore indispensable to facilitate accurate lesion localization and to improve network generalization. The preprocessing framework adopted in this thesis comprises three major components: (1) Contrast enhancement via Contrast Limited Adaptive Histogram Equalization (CLAHE), (2) Region-of-interest (ROI) cropping based on the lesion mask, and (3) Image resizing and normalization. Each stage is mathematically formulated below, and the entire process is summarized in Algorithm 1.

3.2.1 Contrast Enhancement using CLAHE

Conventional histogram equalization often fails in mammograms due to its global nature, which may excessively amplify noise in uniform regions [30]. CLAHE addresses this limitation by enhancing local contrast in small contextual regions (tiles) while capping the amplification to a specified clip limit. Let $I(x,y)$ denote the grayscale intensity at pixel (x,y). The CLAHE-enhanced image, $I_{\text{CLAHE}}(x,y)$, is obtained by applying CLAHE with a clip limit of 2.0 and a tile grid size of $(8,8)$:

$$I_{\text{CLAHE}}(x,y) \leftarrow \text{CLAHE}(I(x,y);\ \text{clipLimit} = 2.0,\ \text{tileGridSize} = (8,8)) \qquad \textbf{(3.1)}$$

In this study, the parameters were set to clipLimit = 2.0 and tileGridSize = $(8,8)$, which were empirically found to provide optimal local contrast enhancement without introducing noise artifacts. A visual comparison between the original and CLAHE-enhanced mammogram is presented in Figure 3.2.

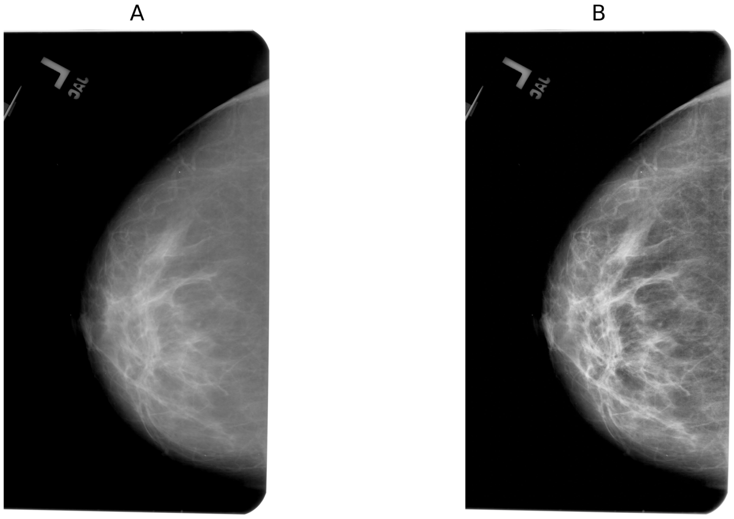

Figure 3.2 Effect of CLAHE on mammogram contrast enhancement. A: Original grayscale mammogram from a benign case in the training set (CC view). B: The result after applying CLAHE with a clip limit of 2.0 and a tile grid size of 8×8. CLAHE enhances local contrast while preventing noise amplification in homogeneous regions.

3.2.2 ROI Cropping with Margin

Lesions are often confined to small areas of the mammogram. To focus learning on diagnostically relevant regions and reduce the influence of extraneous background, ROI-based cropping is performed. Given a binary mask $M(x,y) \in \{0,255\}$, lesion boundaries are determined by locating non-zero pixels:

$$x_1 = \max(0, \min(x \mid M(x,y) = 255) - m), \quad x_2 = \min(W, \max(x \mid M(x,y) = 255) + m) \tag{3.2}$$

$$y_1 = \max(0, \min(y \mid M(x,y) = 255) - m), \quad y_2 = \min(H, \max(y \mid M(x,y) = 255) + m) \tag{3.3}$$

Here, W and H are the image width and height, respectively, and $m = 50$ is the margin added to retain anatomical context. Figure 3.3 shows the outcome of the ROI cropping procedure.

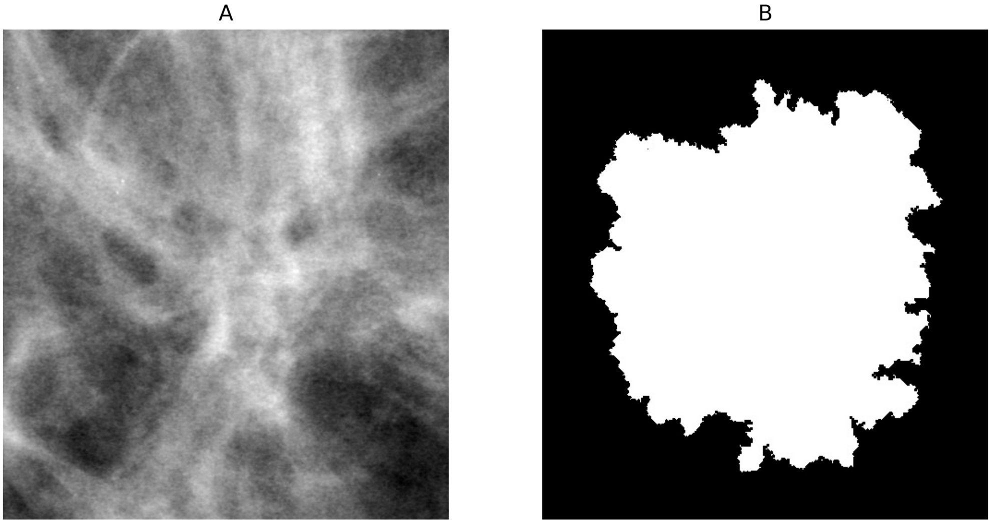

Figure 3.3 ROI-based cropping performed on a sample mammogram. A: Cropped grayscale image containing the lesion area with a margin of 50 pixels. B: Corresponding region-of-interest (ROI) mask showing the precise annotated lesion boundary within the cropped region.

3.2.3 Resizing and Normalization

After cropping, the image and the mask are resized to 256×256 to have the same input size in the dataset. Resize is performed with bilinear interpolation both for images and nearest- neighbor interpolation for binary masks. The grayscale values in the image are normalized to $[0, 1]$ as follows:

$$I_{\text{norm}}(x, y) = \frac{I(x, y)}{255} \tag{3.4}$$

$$M_{\text{bin}}(x, y) = \begin{cases} 1, & \text{if } M(x, y) > 127 \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

The output of this stage is a float32 normalized image and a binary mask that can be provided to the segmentation model.

3.2.4 Preprocessing Algorithm Summary

The entire pipeline is formally described in Algorithm 1, which clearly defines each step from CLAHE enhancement to binary mask creation. This pseudocode can be implemented using standard libraries such as OpenCV and NumPy, ensuring reproducibility.

Algorithm 1 Pseudo-code for Mammogram Image Preprocessing Pipeline

```
Require: Original grayscale mammogram $I$, corresponding binary mask $M$, margin $m$, resize size $s$
Ensure: Preprocessed image $I_{out}$ and binary mask $M_{out}$
Step 1: Apply CLAHE for Contrast Enhancement
Divide $I$ into non-overlapping tiles of size $8 \times 8$
Clip histogram with clip limit = 2.0
Apply histogram equalization locally
    $I_{\text{clahe}} \leftarrow \text{CLAHE}(I, \text{clipLimit} = 2.0, \text{tileGridSize} = 8 \times 8)$
Step 2: ROI-Based Cropping
Extract coordinates of lesion from mask:
    $X_{\min}, X_{\max} \leftarrow \min(x), \max(x) \mid M(x,y) = 255$
    $Y_{\min}, Y_{\max} \leftarrow \min(y), \max(y) \mid M(x,y) = 255$
Apply margin $m$ and crop both image and mask:
    $I_{\text{crop}} \leftarrow I_{\text{clahe}}[Y_{\min} - m : Y_{\max} + m, X_{\min} - m : X_{\max} + m]$
    $M_{\text{crop}} \leftarrow M[Y_{\min} - m : Y_{\max} + m, X_{\min} - m : X_{\max} + m]$
Step 3: Resize to Fixed Dimensions
Resize $I_{\text{crop}}$ and $M_{\text{crop}}$ to $s \times s$ (e.g., $256 \times 256$):
    $I_{\text{resized}} \leftarrow \text{resize}(I_{\text{crop}}, s)$
    $M_{\text{resized}} \leftarrow \text{resize}(M_{\text{crop}}, s)$
Step 4: Normalize and Binarize
    $I_{out}(x,y) \leftarrow \frac{I_{\text{resized}}(x,y)}{255}$
    $M_{out}(x,y) \leftarrow \begin{cases} 1 & \text{if } M_{\text{resized}}(x,y) > 127 \\ 0 & \text{otherwise} \end{cases}$
return $I_{out}, M_{out}$
```

This pipeline addresses major challenges in mammographic image preprocessing. Its modular design allows easy substitution or extension (e.g., gamma correction or log transform as alternatives to CLAHE), which can be explored in future work for comparative analysis [31].

3.3 Baseline Architectures

3.3.1 From CNN to U-Net

Deep learning for medical image segmentation is based on the concept of Convolutional Neural Networks (CNN) that have revolutionized computer vision by their ability for hierarchical feature learning. The original design of LeNet-5 and subsequent AlexNet [32] showed that convolutional layers could learn spatial hierarchies of features via backpropagration. CNNs are a composition of layers, such as convolutional layers, pooling layers and fully connected layers, that process the raw input image into a compact and discriminative representation.

In semantic segmentation challenges on medical image data, conventional CNNs are not well suitable due to the poor spatial resolution caused by multiple poolings. The reason for the limited receptive field is fully connected layers, and fully convolutional layers (FCN) [33] are introduced without fully connected layers in order to preserve spatial information. FCNs used the upsampling layers, such as the transposed convolutions, to reconstruct the pixels with higher spatial resolution.

Nevertheless, FCNs continued to encounter difficulties in boundary-localization accuracy for complex biomedical data. This problem resulted in the U-Net architecture to be proposed by Ronneberger et al. [10], achieved wide recognition in the biomedical imaging community because of its encoder-decoder design and skip connections.

3.3.2 Vanilla U-Net

The vanilla U-Net architecture is one of the most commonly used convolutional neural networks, specially designed for semantic segmentation in medical images [10]. It has a symmetric encoder–decoder architecture with skip connections between layers at the corresponding position along the encoding and decoding paths. These relations allow the model to sufficiently integrate high-level semantic representations and low-level spatial details, and are essential for accurate boundary localization for biomedical images.

The **encoder** is composed of four downampling blocks. Within each block two consecutive 3 × 3 kernel convolutional layers are stacked, followed by a ReLU, and a 2 × 2 max pooling with stride 2. The spatial resolution of the input is reduced by half and the number of feature channels is doubled as the input is transmitted though each block. This is hierarchical feature extraction that has the effect of abstracting context from the input image.

At the deepest level, the **bottleneck** layer consists of two convolutional layers with 1024 filters, both using 3×3 kernels and ReLU activations. This part of the network processes highly compressed feature representations and acts as a bridge between the encoder and decoder.

The **decoder** mirrors the encoder structure, employing upsampling followed by convolution. Each decoder block begins with an upsampling operation—implemented using nearest-neighbor interpolation to double the spatial resolution—followed by the concatenation of feature maps from the corresponding encoder layer via skip connections. This is followed by two 3×3 convolutional layers with ReLU activations. These operations progressively reconstruct the segmentation mask at full resolution.

The output layer consists of a 1×1 convolution with a sigmoid activation to generate a binary segmentation mask.

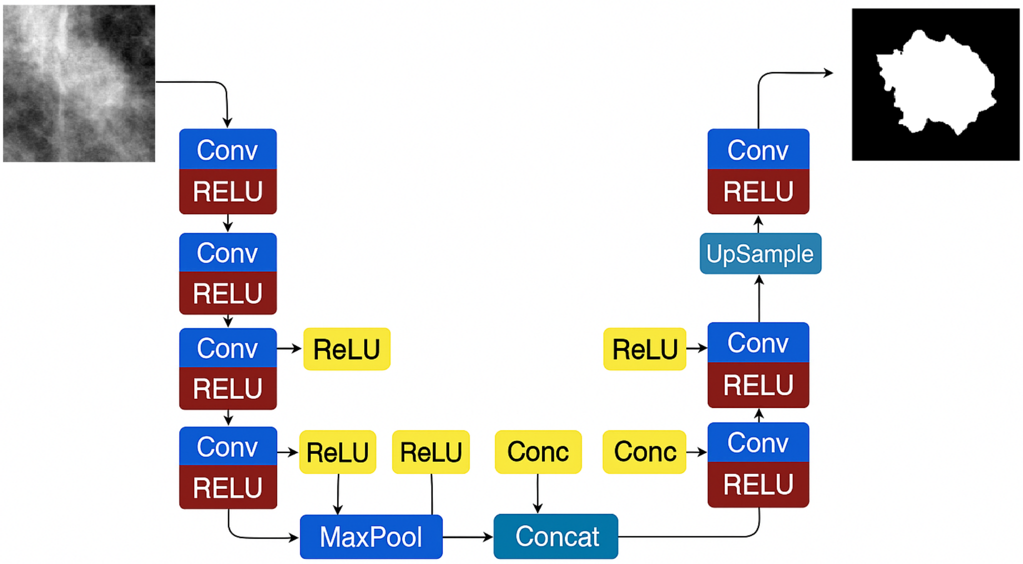

Figure 3.4 Vanilla U-Net architecture used as a baseline in this study. The model accepts a 256 × 256 grayscale mammogram image as input and outputs a binary ROI mask of the same spatial resolution.

The overall structure of the Vanilla U-Net employed in this research is displayed in Figure 3.4. The Vanilla U-Net takes as an input of a 256 × 256 grayscale mammogram image, and the expected output is a binary region-of-interest (ROI) mask in the same resolution. During training, the model is trained to predict this binary mask from input images form input images also, enabling the delineation of suspicious/abnormal tissue regions in the breast.

Formally the operations in the network can be represented as:

$$E_i = \mathrm{MaxPool}(\mathrm{ReLU}(\mathrm{Conv}_{3\times3}(\mathrm{ReLU}(\mathrm{Conv}_{3\times3}(x_{i-1}))))) \tag{3.6}$$

$$D_j = \mathrm{ReLU}(\mathrm{Conv}_{3\times3}(\mathrm{ReLU}(\mathrm{Conv}_{3\times3}([\mathrm{Up}(D_{j+1}), E_{i-j}])))) \tag{3.7}$$

$$f_{out} = \sigma(\mathrm{Conv}_{1\times1}(D_1)) \qquad \textbf{(3.8)}$$

Here x_{i-1} is the input to next encoder block and E_i denotes encoded feature map, D_j denotes decoded feature map, Up($\cdot$) stands for the upsampling and σ is the sigmoid activation function for binary mask output.

3.3.3 Attention U-Net

The vanilla U-Net architecture [10] plays a pioneering role in the field of modern biomedical image segmentation, presenting a symmetry encoder-decoder architecture that can maintain spatial context by the utilization of skip connections. However, the model globally passes all encoder features to the decoder, including useless background information. This limitation may introduce inaccuracy in segmentation, in particular for mammographic images, where lesions are frequently embedded in complicated anatomical background.

To circumvent this, Oktay et al. proposed the Attention U-Net [11] that adds attention gates to U-Net framework (AGs). These components serve as active filters for the skip connections, allowing the model to suppress unimportant regions while enhancing salient anatomy such as tumour boundaries. In Figure 3.5, all attention gates are used just before the concatenation process in the decoder to ensures that we use only the most suitable encoder features for reconstruction.

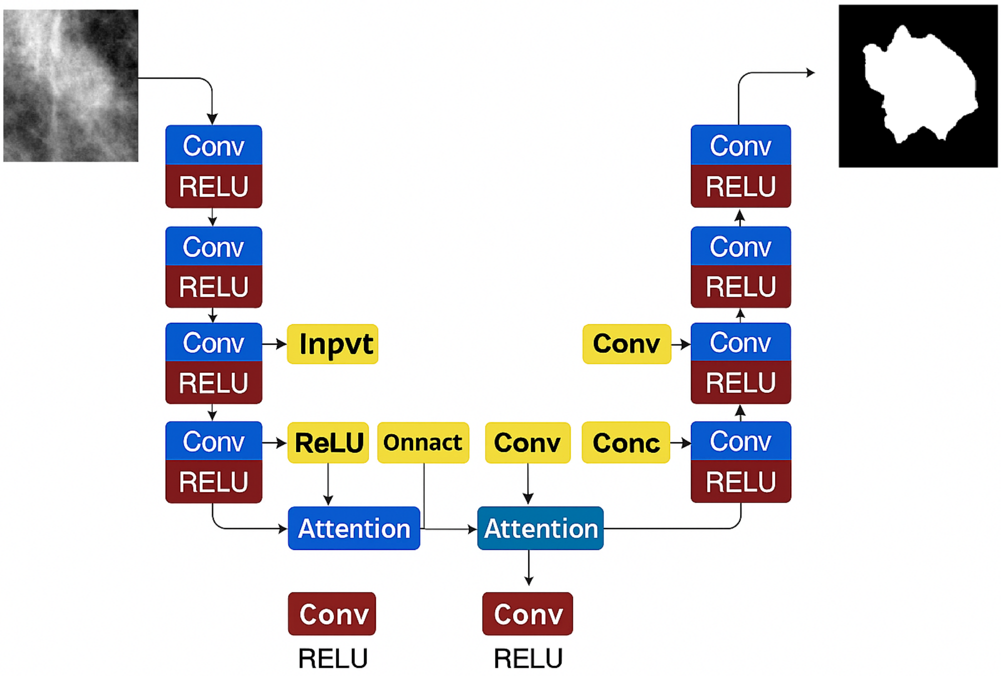

Figure 3.5 Attention U-Net architecture. Attention gates (AGs) filter encoder features before being concatenated with upsampled decoder features, enabling the network to emphasize lesion-relevant regions.

The encoder path consists of four stages and every stage is composed of two 3×3 convolutional layers with ReLU activations and then 2×2 max pooling. These stages gradually decrease the spatial resolution and increase the number of feature channels of the map to capture the texture information of breast tissue in a hierarchical manner.

At the bottleneck (the bottom most layer of the architecture), there are two 3 × 3 convolution with 1024 filters that capture more abstract features by repeating convolutions and synthesis global context information from the down sampled input.

The architecture of the decoder is the same as that of the encoder, and uses upsampling layers to recover the spatial resolution. At every decoder layer, similar attention gate is applied to the corresponding output of the encoder before they are concatenated. This gating mechanism uses decoder features g to compute the attention coefficients α to gate encoder features x as contextual guidance. The mathematically definition of attention gating is:

$$\alpha = \sigma\left(\psi^T\left(\mathrm{ReLU}(W_x x + W_g g + b)\right)\right) \tag{3.9}$$

where W_x and W_g are feature projection 1×1 convolutions, ψ^T is another 1×1 convolution to the joint feature map, and $\sigma$is the sigmoid activation.

The resulting attention map α is applied element-wise to the encoder features, effectively suppressing background noise and enhancing lesion-related activations. These refined features are then concatenated with upsampled decoder features and processed through two additional 3×3 convolutions to generate a context-rich feature representation.

The final prediction is obtained by applying a 1×1 convolution followed by a sigmoid activation, which produces a binary segmentation map:

$$\hat{Y} = \sigma(\mathrm{Conv}_{1\times1}(F)) \tag{3.10}$$

Here, F represents the final decoded feature map, and $\hat{Y}$ is the resulting binary mask that indicates the predicted region of interest (ROI).

Overall, the Attention U-Net helps improve segmentation by guiding the model to focus on the most relevant features during training. This attention mechanism is particularly useful in mammography, where dense breast tissue and low image contrast often make lesion detection more difficult. By highlighting important regions, the model becomes better at identifying small or hard-to-see masses. As shown in the work of Oktay et al. [11], this approach offers better results than the standard U-Net in many medical image segmentation tasks, making it a strong choice for clinical use.

3.3.4 U^2-Net Architecture

U^2-Net, originally introduced by Qin et al. [34] for salient object detection, has since demonstrated strong performance in dense prediction tasks such as medical image segmentation. In this thesis, the architecture is adapted for the segmentation of breast masses in mammograms using the CBIS-DDSM dataset.

The core building block of U^2-Net is the Residual U-block (RSU), which embeds a classic U-Net structure within a single layered module. This nested design enables the model to perform multiscale feature extraction while preserving spatial resolution. RSU blocks combine a U-Net-like encoder-decoder path with residual connections, allowing each RSU to capture local details through shallow layers and global contextual information via deeper paths. This approach results in feature representations that are both detail-sensitive and context-aware.

Structurally, U^2-Net adheres to the conventional encoder-decoder architecture, but both encoder and decoder are entirely constructed using RSU blocks. Each RSU block processes the input features through a compact U-shaped pathway—comprising repeated convolution, pooling, and upsampling operations—and then fuses the final upsampled features with the block's original input via a residual shortcut. This allows for rich representational capacity while maintaining efficient gradient propagation during training.

The model begins with a single-channel 256×256 grayscale input, which passes through five downsampling stages in the encoder. These stages are implemented with RSUs using increasing numbers of filters (64, 128, 256, 512), progressively extracting higher-order features while reducing spatial dimensions. Each RSU's internal U-shape facilitates learning features at multiple scales within the same resolution band.

At the bottleneck, an RSU block with 512 filters captures the most abstract and deeply contextualized features of the mammogram. From this point, the decoder stages symmetrically reverse the encoding process through four upsampling steps. In each stage, the upsampled features are concatenated with the skip-connected encoder outputs and passed through another RSU block, using decreasing numbers of filters (512, 256, 128, 64) to gradually refine the segmentation mask.

Finally, the output is processed through a 1×1 convolution layer with sigmoid activation to generate a binary segmentation map that highlights the regions of interest (ROIs) corresponding to potential masses.

A schematic illustration of the complete U^2-Net architecture, customized for the task of breast mass segmentation, is shown in Figure 3.6. Each stage in the figure clearly demonstrates the cascading RSU blocks and skip connections facilitating deep feature fusion.

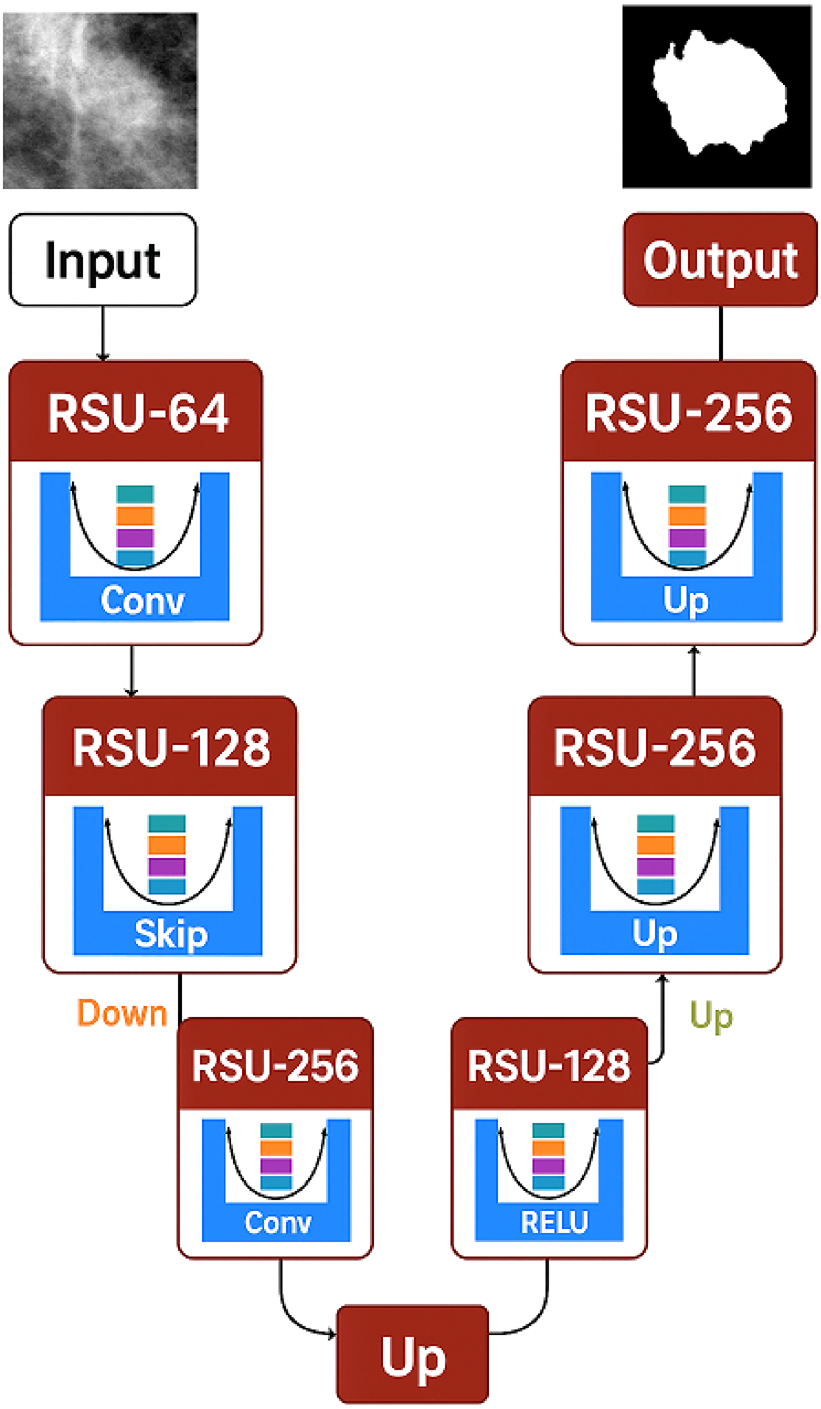

Figure 3.6 U^2-Net model architecture. Each stage is built from Residual U-blocks (RSU), which incorporate nested encoder-decoder paths for multiscale feature learning.

The nested and residual structure of RSU blocks grants U^2-Net its ability to learn simultaneously from both coarse and fine image representations. This capability is particularly advantageous for mammography, where masses may appear with subtle contrasts and irregular shapes. By capturing contextual and spatial cues at multiple scales and depths, the model improves its sensitivity to small and complex lesions, enhancing diagnostic reliability in real-world clinical scenarios.

3.3.5 Attention Cascaded U^2-Net Architecture

The Attention Cascaded U^2-Net (AU^2-Net), proposed by Dhivya et al. [35], is an enhanced version of the original U^2-Net architecture [34], tailored specifically for complex medical image segmentation tasks such as breast mass detection. While U^2-Net was initially developed for salient object detection, AU^2-Net extends its capabilities through a cascaded attention-driven design. In this thesis, it is adopted as one of the baseline models for comparative evaluation.

AU^2-Net is built upon a two-stage cascaded architecture, where the output of the first U^2-Net subnetwork is passed to a second subnetwork for further refinement. Both subnetworks follow a typical encoder-bottleneck-decoder layout, but conventional convolutional blocks are replaced with more advanced modules that improve spatial attention and multiscale contextual learning.

The architecture integrates three key components:

1. Residual U-blocks (RSUs): Each RSU incorporates a lightweight, nested U-Net structure within a single block, enabling simultaneous extraction of deep semantic features and fine-grained local details. Internally, the RSU performs multiple levels of downsampling and upsampling, concatenates intermediate features, and applies residual connections to maintain information flow. This design effectively preserves both edge-level precision and global context, while also promoting gradient stability in deep networks.

2. Attention Gates: Inspired by attention U-Net [11], spatial attention blocks are applied to the encoder features before concatenation in the decoder. These gates help filter out irrelevant activations and allow the model to emphasize tumor-relevant structures. The gating mechanism is defined by:

$$\psi = \sigma\left(\text{Conv}_{1\times1}\left(\text{ReLU}(\theta_x + \phi_g)\right)\right) \tag{3.11}$$

where θ_x and ϕ_g are 1×1 convolutions of encoder and decoder features, respectively.

3. Atrous Spatial Pyramid Pooling (ASPP): ASPP modules, placed at the bottleneck of both subnetworks, use parallel dilated convolutions with dilation rates $r \in \{6, 12, 18\}$, following the strategy proposed by Chen et al. in the DeepLabv3 model [36]. This allows the model to incorporate features from multiple receptive fields, useful in segmenting lesions with varying scale and morphology. The ASPP block is defined as:

$$\text{ASPP}(x) = \text{Conv}1 \times 1\left(\bigoplus r \in \{6, 12, 18\}\text{Conv}_{3\times3}^{(r)}(x)\right) \tag{3.12}$$

where $\oplus$ denotes concatenation.

Each decoder stage in both subnetworks performs upsampling followed by RSU blocks and attention-enhanced skip connections. The final output of the second decoder is passed through an additional ASPP module and a 1×1 convolution with sigmoid activation to produce the binary segmentation mask.

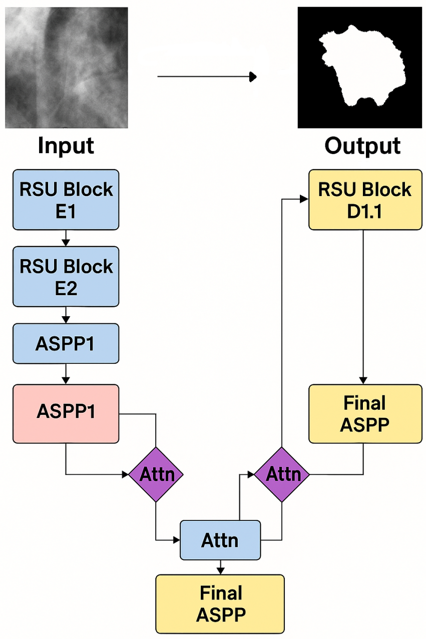

Figure 3.7 Attention Cascaded U^2-Net architecture. The network consists of two serial U^2-Net structures composed of RSU blocks, ASPP modules, and attention gates.

The combined design of AU^2-Net enables the model to iteratively refine segmentation results. While this architecture introduces increased complexity and computational cost, it is particularly useful in identifying and segmenting small, low-contrast tumors in mammographic images. Its integration of multiscale, attention-guided, and residual learning mechanisms provides a strong baseline reference in our experiments.

3.4 Proposed Architectures

3.4.1 ETDP-U^2-Net Architecture

The ETDP-U^2-Net (Edge-Texture Dual-Path U^2-Net) is one of the novel architectures proposed in this thesis. It extends the foundational U^2-Net structure [34] by integrating edge and texture-specific feature pathways, cross-attention mechanisms, and enhanced multiscale representations via Squeeze-and-Excitation (SE) modules [37] and deep supervision.

Motivation and Design Rationale: The key idea of ETDP-U^2-Net is to independently extract edge and texture cues—both critical in segmenting mammographic masses—via two distinct Residual U-block (RSU) branches. These branches are then fused using a cross-attention block that aligns and integrates spatially complementary patterns, as visualized in Figure 3.8.

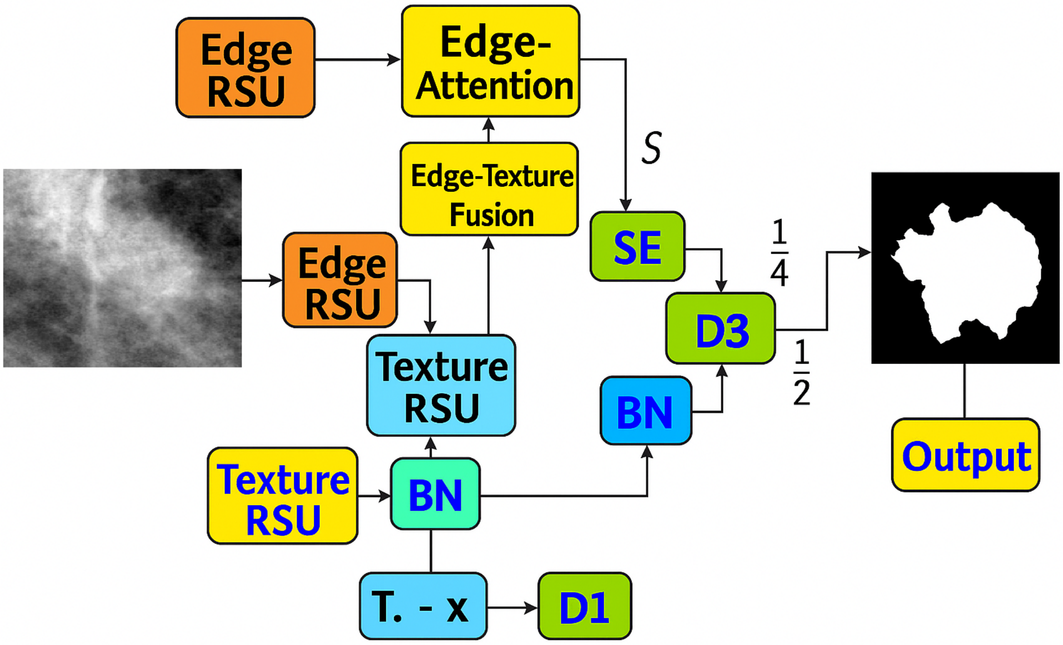

Figure 3.8 ETDP-U^2-Net architecture. The network includes dual-path RSUs for edge and texture extraction, SE modules for channel attention, a cross-attention fusion mechanism, and deep supervision for hierarchical learning.

RSU Modules: RSU blocks form the core computational unit in the architecture. Each RSU encapsulates an internal encoder-decoder structure with skip connections and ends in a residual sum to the original projection of the input. Mathematically, this can be formulated as:

$$\text{RSUout} = \mathcal{F}\text{dec}(\mathcal{F}_{\text{enc}}(X)) + X' \tag{3.13}$$

where X' denotes a normalized and projected version of input X. This design supports both multiscale representation and stable gradient flow.

Squeeze-and-Excitation (SE) Mechanism: To adaptively recalibrate channel-wise features, SE blocks are placed inside each convolutional unit of the RSUs. Their formulation is:

$$\mathrm{SE}(X) = X \cdot \sigma(W_2 \delta(W_1, \mathrm{GAP}(X))) \tag{3.14}$$

Here, δ and σ represent ReLU and sigmoid activations respectively, and GAP is global average pooling. This operation emphasizes informative channels and suppresses irrelevant ones. The SE module is adopted from Hu et al. [37].

Cross-Attention Fusion: Edge- and texture-path RSU outputs are integrated using a cross-attention block, enhancing complementary information exchange. The attention map A is computed as:

$$A = \sigma(\mathrm{Conv}1 \times 1(\mathcal{E}(X\mathrm{edge}) \odot \mathcal{T}(X_{\mathrm{texture}}))) \tag{3.15}$$

where $\mathcal{E}$ and $\mathcal{T}$ are 1×1 convolutions and $\odot$ denotes element-wise multiplication. The resulting A modulates both edge and texture features before fusion.

Hierarchical Encoder-Decoder with Deep Supervision: The fused tensor is fed into a standard encoder-decoder framework constructed with RSUs of increasing filter widths. Decoder stages mirror the encoder via symmetric upsampling and skip connections. To aid convergence and multi-scale learning, intermediate supervision is applied.

Multi-Stage Supervision: The final prediction $\hat{Y}$ aggregates three outputs as:

$$\hat{Y} = \sigma\left(\hat{Y}_1 + \lambda_2 \cdot \mathrm{Resize}(\hat{Y}_2) + \lambda_3 \cdot \mathrm{Resize}(\hat{Y}_3)\right) \tag{3.16}$$

where $\lambda_2 = 0.5$ and $\lambda_3 = 0.25$ are predefined weights for deep supervision outputs.

Pseudocode for ETDP-U^2-Net: The complete inference workflow is summarized in Algorithm 2.

Algorithm 2 ETDP-U^2-Net Inference Procedure

```
Input image $I \in \mathbb{R}^{256\times256}$
$X_{\text{edge}} \leftarrow$ RSUedge($I$)
$X$texture $\leftarrow$ RSUtexture($I$)
$A \leftarrow$ CrossAttention($X$edge, $X_{\text{texture}}$)
$X \leftarrow$ Concat($A \odot X_{\text{edge}}, A \odot X_{\text{texture}}$)
$X \leftarrow$ Encoder RSUs($X$)
$Y_1, Y_2, Y_3 \leftarrow$ Decoder + Deep Supervision
**return** $\hat{Y} = \sigma(Y_1 + 0.5 \cdot \text{Resize}(Y_2) + 0.25 \cdot \text{Resize}(Y_3))$
```

This model was custom-designed for this thesis to better handle the detection of subtle, low-contrast lesions that exhibit fuzzy borders, which are commonly seen in mammography. Comparative results and ablation studies are detailed in the subsequent Results chapter.

3.4.2 DPTrans-U^2-Net Architecture

The DPTrans-U^2-Net (Dual-Path Transformer U^2-Net) is a transformer-augmented segmentation model proposed in this study, specifically designed to enhance breast mass segmentation performance. It extends the structure of the previously introduced ETDP-U^2-Net by incorporating a transformer module at the bottleneck level of the encoder-decoder hierarchy, thereby enabling improved contextual feature modeling over long spatial ranges.

Architecture Overview: The model begins by extracting edge and texture representations independently through two distinct Residual U-blocks (RSUs). These two feature maps are then fused using a cross-attention mechanism that dynamically emphasizes mutually salient regions. The fused features are passed through a downsampling encoder path consisting of additional RSU stages. A transformer block is placed at the deepest layer of the network, introducing self-attention over flattened spatial dimensions and allowing the model to capture global dependencies across the entire image.

Transformer Bottleneck: The transformer module applies multi-head self-attention (MHSA) followed by a feedforward network (FFN) to enrich the feature representation.

Formally, the transformation applied to the encoded features X can be expressed as:

$$\text{Transformer}(X) = \text{Reshape}^{-1}\left(\mathcal{F}_{\text{FFN}}\left(\mathcal{F}_{\text{MHSA}}(\text{LayerNorm}(\text{Reshape}(X)))\right) + X\right) \tag{3.17}$$

Here, $\mathcal{F}_{\text{MHSA}}$ and $\mathcal{F}_{\text{FFN}}$ denote the multi-head self-attention and feedforward layers, respectively. This bottleneck mitigates the local receptive field limitation inherent in conventional convolutions.

Decoder and Supervision: After the transformer block, the decoder reconstructs the segmentation mask via upsampling and skip connections, using corresponding RSU blocks at each level. Deep supervision is applied by generating auxiliary outputs at multiple decoder stages, which are later aggregated to produce the final output prediction:

$$\hat{Y} = \sigma(Y_1 + 0.5 \cdot \text{Resize}(Y_2) + 0.25 \cdot \text{Resize}(Y_3)) \tag{3.18}$$

A detailed architectural overview of the DPTrans-U^2-Net, including the transformer bottleneck and dual-path RSU stages, is illustrated in Figure 3.9.

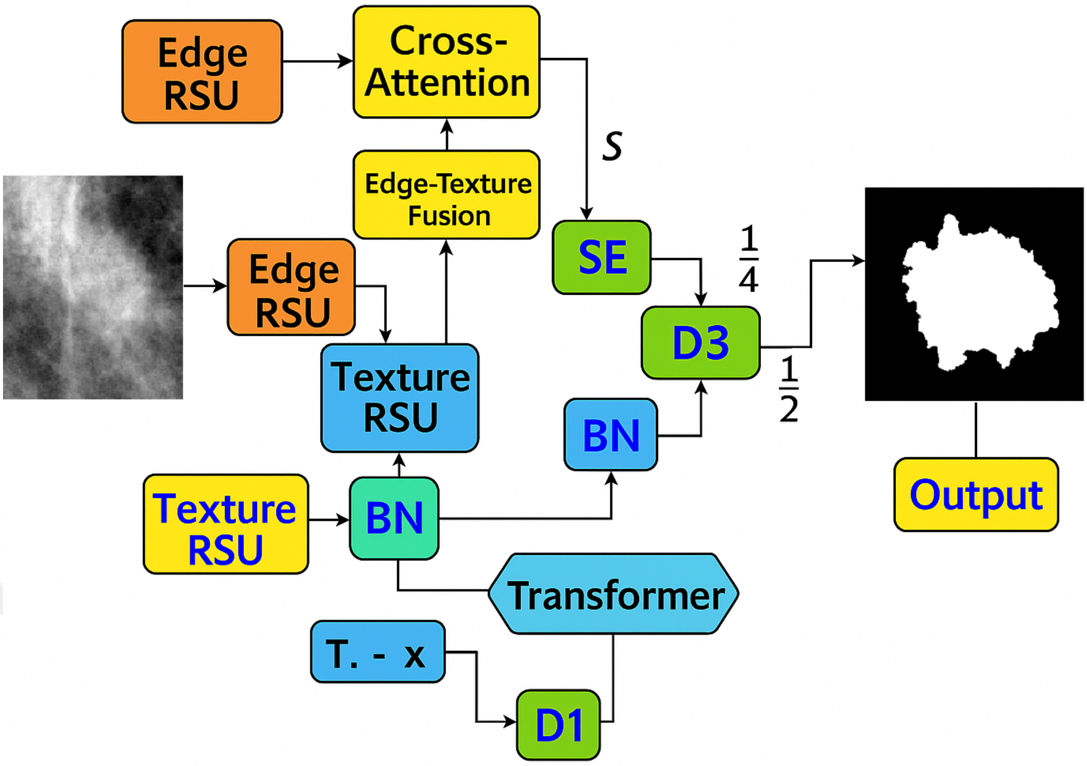

Figure 3.9 The architecture of DPTrans-U^2-Net, which integrates a transformer bottleneck module within a dual-path RSU framework for enhanced and enriched feature interaction dynamics.

Inference Flow: The inference procedure for DPTrans-U^2-Net is summarized in Algorithm 3, highlighting its sequential processing stages.

Algorithm 3 DPTrans-U^2-Net Inference Procedure

```
Input: Image $I \in \mathbb{R}^{256\times256}$
$X_{\text{edge}} \leftarrow \text{RSU}_{\text{edge}}(I)$
$X_{\text{texture}} \leftarrow \text{RSU}_{\text{texture}}(I)$
$A \leftarrow \text{CrossAttention}(X_{\text{edge}}, X_{\text{texture}})$
$X \leftarrow \text{Concat}(A \odot X_{\text{edge}}, A \odot X_{\text{texture}})$
$X \leftarrow \text{Downsample} + \text{RSUs}(X)$
$X \leftarrow \text{TransformerBottleneck}(X)$
$Y_1, Y_2, Y_3 \leftarrow \text{Decoder} + \text{Deep Supervision}$
Return: $\hat{Y} = \sigma(Y_1 + 0.5 \cdot \text{Resize}(Y_2) + 0.25 \cdot \text{Resize}(Y_3))$
```

This architecture enhances both the propagation of global contextual features and the precision of fine-grained segmentation boundaries. Its contribution to overall performance is examined in detail in the comparative evaluation section of this thesis.

3.4.3 DPCA-U^2-Net Architecture

DPCA-U^2-Net: A Dual-Path Cross-Attention Framework

DPCA-U^2-Net (Dual-Path Cross-Attention U^2-Net) is a deep segmentation model introduced in this thesis to address the specific challenges of detecting breast masses in mammograms. It builds on the structure of U^2-Net [34] but introduces targeted improvements to better capture two key visual cues in medical images: texture and edge boundaries. These enhancements are particularly important in mammography, where lesions often appear subtle and poorly defined.

3.4.3.0.1 Model Motivation and Novelty. In real clinical settings, radiologists often note that breast tumors can have fuzzy edges or irregular textures, which makes their detection more difficult. While conventional segmentation models can pick up general patterns, they often struggle to differentiate fine details. DPCA-U^2-Net addresses this by processing edge and texture information separately. It uses two distinct encoder paths—each made of Residual U-blocks (RSUs)—so the network can learn to handle these features independently. This design helps the model produce clearer, more meaningful predictions by preventing the blending of visual signals that are fundamentally different.

3.4.3.0.2 Dual RSU-Based Feature Extraction. Given an input image X, the model sends it through two parallel branches:

$$X_{\text{edge}} = \text{RSU}_{\text{edge}}(X) \tag{3.19}$$

$$X_{\text{texture}} = \text{RSU}_{\text{texture}}(X) \tag{3.20}$$

Each RSU block follows an encoder-decoder layout with skip connections, just like in U^2-Net, allowing the model to extract information at multiple scales while maintaining the flow of gradients. By using this dual-path structure, DPCA-U^2-Net is better equipped to highlight both sharp edges and soft tissue textures, which are critical for segmenting breast lesions accurately.

$$X_{\text{edge}} = \text{RSU}_{\text{edge}}(X), \quad X_{\text{texture}} = \text{RSU}_{\text{texture}}(X) \tag{3.21}$$

Cross-Attention Fusion: The two streams are fused through a soft cross-attention mechanism that dynamically weighs the interaction between edge and texture activations. This is mathematically defined as:

$$E = \text{Conv}_{1\times1}(X_{\text{edge}}), \quad T = \text{Conv}_{1\times1}(X_{\text{texture}}) \tag{3.22}$$

$$A = \sigma(E \odot T) \tag{3.23}$$

$$F_{\text{fused}} = \text{Conv}_{3\times3}\left(\text{Concat}(A \odot X_{\text{edge}}, A \odot X_{\text{texture}})\right) \tag{3.24}$$

where $\odot$ denotes element-wise multiplication and σ is the sigmoid function. This fusion technique is adapted from dual-attention mechanisms previously explored in semantic segmentation [38], but tailored here to specifically merge edge and texture channels.

Hierarchical Encoder-Decoder Architecture: After cross-attention fusion, the combined features are downsampled through deeper RSU blocks with increasing filter widths (128, 256, 512). These layers enable learning from coarse contextual cues. The decoder mirrors the encoder path and includes skip connections at each resolution level to preserve spatial detail. The final prediction is generated via a sigmoid-activated 1×1 convolution layer.

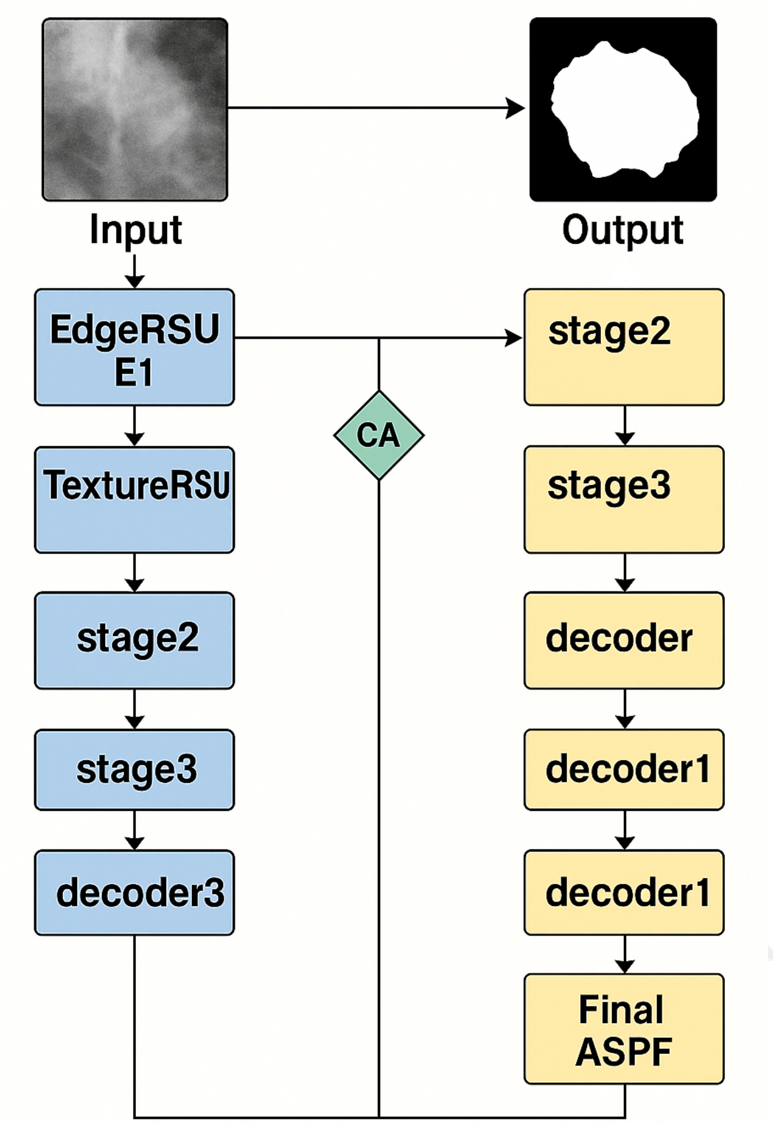

Figure 3.10 DPCA-U^2-Net architecture. Edge and texture-specific RSU streams are fused with cross-attention. Encoder-decoder hierarchy reconstructs the lesion mask.

Pseudocode for DPCA-U^2-Net:

Algorithm 4 DPCA-U^2-Net Inference Procedure

```
Input: Image $I \in \mathbb{R}^{256\times 256}$
$X_{\text{edge}} \leftarrow \text{RSU}_{\text{edge}}(I)$
$X_{\text{texture}} \leftarrow \text{RSU}_{\text{texture}}(I)$
$A \leftarrow \sigma(\text{Conv}_{1\times 1}(X_{\text{edge}}) \odot \text{Conv}_{1\times 1}(X_{\text{texture}}))$
$F_{\text{fused}} \leftarrow \text{Conv}_{3\times 3}(\text{Concat}(A \odot X_{\text{edge}}, A \odot X_{\text{texture}}))$
$F \leftarrow \text{Encoder RSU stages}(F_{\text{fused}})$
$Y \leftarrow \text{Decoder RSU stages + Skip Connections}(F)$
return $\hat{Y} = \text{Sigmoid}(\text{Conv}_{1\times 1}(Y))$
```

Unlike existing architectures, DPCA-U^2-Net does not rely on transformer-based components. Its strength lies in its architectural simplicity and domain-specific design, making it computationally efficient while retaining strong segmentation performance. To the best of our knowledge, no prior work has combined U^2-Net-style RSU blocks with explicit dual-path cross-attention tailored for edge-texture interaction, making this model an original contribution within this thesis.

3.5 Data Augmentation Strategy

To mitigate overfitting and improve model generalizability, we employed a 16-fold data augmentation strategy, inspired by the augmentation protocol proposed in [35]. This approach is particularly beneficial in mammographic image segmentation tasks, where lesion instances are limited and class imbalance is prominent.

Motivation: As noted by Dhivya et al. [35], conventional deep networks are prone to overfitting on small medical image datasets. Instead of relying on random augmentations, their study proposed a carefully curated set of geometric transformations shown to improve lesion localization and boundary preservation in breast tumor segmentation. Based on their evaluation, we adopted this deterministic 16x augmentation set in our preprocessing pipeline.

Augmentation Set: Let X denote an original grayscale mammogram image of size 256×256, and Y its corresponding binary ROI mask. The augmented image-mask pairs are denoted as:

$$\{(X_k, Y_k)\}_{k=1}^{16} = \mathcal{T}_k(X, Y) \tag{3.25}$$

where $\mathcal{T}_k$ is the k-th transformation from the list below:

- **Identity:** $\mathcal{T}_1(X) = X$
- **Rotation:** $\mathcal{T}_{\text{rot}}(X, \theta) = R_\theta X$, with $\theta \in \{45°, 90°, 270°\}$
- **Flipping:** Horizontal $(x, y) \mapsto (w - x, y)$, vertical $(x, y) \mapsto (x, h - y)$

- **Scaling:** $\mathcal{T}_{\text{scale}}(X, s) = S_s X$, where $s \in \{0.9, 1.1, 1.2\}$
- **Translation:** $\mathcal{T}_{\text{trans}}(X, \Delta x, \Delta y) = X(x - \Delta x, y - \Delta y)$
- **Compositions:** Mixed augmentations such as $\mathcal{T}_{\text{rot}}(R_{90} \circ \text{flip}_H(X))$

Output: The resulting dataset is expanded by a factor of 16:

$$(X, Y) \mapsto \{(X_k, Y_k)\}_{k=1}^{16} \tag{3.26}$$

All augmented images and masks are saved with consistent naming conventions for reproducibility and are used only within the training set to avoid information leakage.

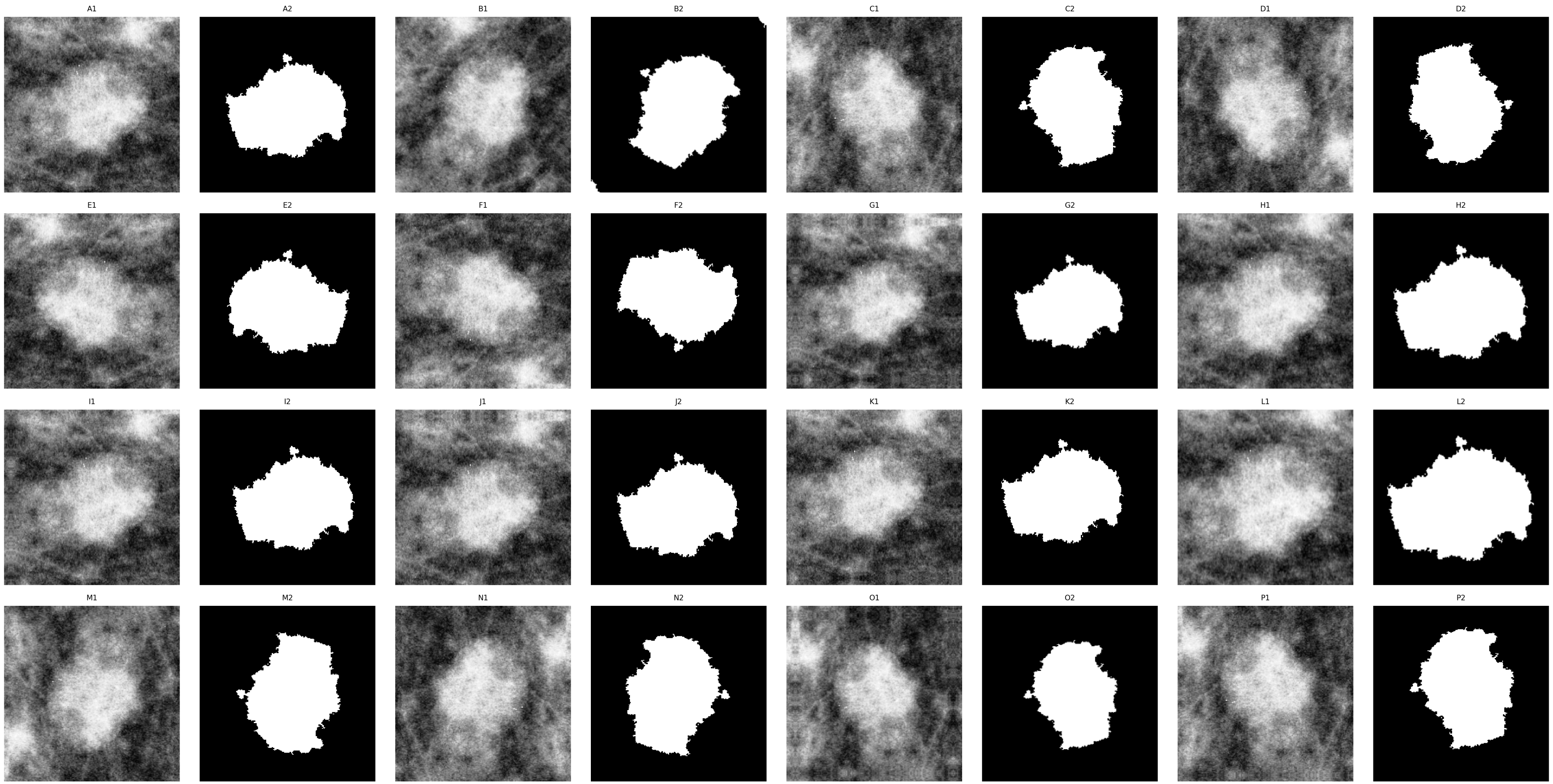

Figure 3.11 Visualization of deterministic 16× data augmentation applied to a region-of-interest (ROI) cropped mammogram and its corresponding binary mask. Each subfigure pair illustrates a distinct augmentation and its spatially aligned ROI label. A1 is the CLAHE-enhanced cropped original image, and A2 is its corresponding mask. Rotations of 45°, 90°, and 270° are applied in B1/B2, C1/C2, and D1/D2, respectively. Horizontal and vertical flips are shown in E1/E2 and F1/F2. G1–H2 depict scaling by 0.9× and 1.1×, while I1–K2 correspond to translations of +10 pixels in x, +10 pixels in y, and -10 pixels in both axes. L1/L2 applies 1.2× scaling. M1–N2 combine flipping and 90° rotation; O1/O2 combines 0.9× scaling and 90° rotation; and P1/P2 applies translation followed by 90° rotation. These augmentations are designed to improve generalization and robustness of the proposed segmentation model while preserving structural fidelity of the lesion regions.

This controlled augmentation approach provides a rich set of spatial variations without introducing semantic distortions, and has empirically led to improved segmentation performance across all architectures tested in this thesis.

3.6 Evaluation Metrics and Loss Functions

To evaluate segmentation quality, we employed six standard metrics commonly used in medical image analysis: Accuracy, Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision, Recall, and the F_2-Score. These metrics provide complementary perspectives on pixel-level correctness, overlap, and clinical relevance. In the context of mammographic segmentation, where lesion boundaries are subtle and lesions vary significantly in size and contrast, each metric contributes uniquely to assessing model performance.

3.6.1 Dice Similarity Coefficient (DSC)

The Dice coefficient quantifies the spatial overlap between predicted (P) and ground-truth (G) segmentation masks:

$$\mathrm{DSC}(P,G) = \frac{2|P \cap G| + \epsilon}{|P| + |G| + \epsilon} \tag{3.27}$$

where ϵ is a smoothing term (set to 1) to prevent division by zero. First introduced as a differentiable loss in medical imaging by Milletari et al. [39], DSC is robust to class imbalance and is particularly useful for assessing small tumors. In mammography, it effectively captures overlap between segmented lesions and ground truth, making it a critical metric when tumor regions are small or irregular.

3.6.2 Intersection over Union (IoU)

IoU, or the Jaccard Index, is defined as:

$$\mathrm{IoU}(P,G) = \frac{|P \cap G| + \epsilon}{|P \cup G| + \epsilon} \tag{3.28}$$

It penalizes false positives more heavily and is widely adopted in medical image segmentation benchmarks [33, 40]. In mammographic analysis, IoU is crucial for evaluating precise delineation of tumor boundaries, especially in complex backgrounds.

3.6.3 Accuracy

Accuracy is the ratio of correctly classified pixels (true positives and true negatives) to the total number of pixels:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.29}$$

Though simple, accuracy may be misleading in highly imbalanced datasets such as mammograms where background pixels vastly outnumber foreground tumor pixels. Nevertheless, it provides a coarse measure of overall classification correctness.

3.6.4 Precision and Recall

Precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP+FP+\epsilon} \tag{3.30}$$

$$\text{Recall} = \frac{TP}{TP+FN+\epsilon} \tag{3.31}$$

Precision measures the correctness of positive predictions, while recall emphasizes completeness. In clinical scenarios, recall is especially significant because missing tumor pixels (false negatives) could lead to missed diagnoses. Both metrics are widely used in medical AI evaluations [41].

3.6.5 F$_2$-Score

The F$_2$-Score is a weighted harmonic mean that favors recall more than precision:

$$F\_2 = \frac{5\cdot\text{Precision}\cdot\text{Recall}}{4\cdot\text{Precision}+\text{Recall}+\epsilon} \tag{3.32}$$

This metric is appropriate in mammography where detecting every possible tumor pixel is critical, even at the expense of increased false positives [41].

3.6.6 Loss Functions

To overcome challenges like tumor heterogeneity in size and contrast, we tested multiple loss functions:

Binary Cross Entropy (BCE) BCE loss is defined as:

$$\mathcal{L}\mathrm{BCE} = -\frac{1}{N}\sum i = 1^N y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{3.33}$$

It is a pixel-wise classification loss that encourages accurate individual predictions. While BCE provides stable gradients and sharp boundaries, it struggles with extreme class imbalance which is common in medical datasets [42].

Dice Loss Defined as $\mathcal{L}_{\mathrm{Dice}} = 1 - \mathrm{DSC}$, this loss was proposed by Milletari et al. [39]. It improves overlap-based optimization and mitigates class imbalance but may result in unstable convergence for large lesions.

Focal Loss Introduced by Lin et al. [43], Focal Loss includes a modulating term $(1 - p_t)^\gamma$ to focus on hard-to-classify pixels. It enhanced detection of small, low-contrast tumors, though it underperformed for large masses.

Tversky Loss Proposed by Salehi et al. [44], this loss controls trade-off between false positives and false negatives:

$$T(P,G) = \frac{|P \cap G|}{|P \cap G| + \alpha|P \setminus G| + \beta|G \setminus P|} \tag{3.34}$$

We found it useful for improving recall in small tumors but it was overly conservative for larger tumors.

Boundary Loss Suggested by Kervadec et al. [45], this loss uses a signed distance map to emphasize alignment along edges. While beneficial for sharp borders in small lesions, it struggled with low-contrast masses.

3.6.7 Combined Dice + Binary Cross Entropy Loss

Combined Dice + BCE Loss The best performing objective was a combined loss:

$$\mathcal{L}\mathrm{Combo} = \mathcal{L}\mathrm{Dice} + \mathcal{L}_{\mathrm{BCE}} \tag{\textbf{3.35}}$$

Recommended by MONAI [46], this hybrid balances pixel-wise BCE with region-based Dice optimization. It offered robust convergence and generalization across lesion sizes and contrasts, making it the most suitable choice for mammographic segmentation in our study.

3.7 Experimental Setup

All models were trained on the CBIS-DDSM dataset using a training-validation-test split, where the test set remained entirely unseen during both training and hyperparameter tuning phases. To ensure a fair and unbiased evaluation, the test set was strictly separated and excluded from any data augmentation or preprocessing procedures. Augmentation was applied *only* to the training set to avoid data leakage, a known issue that can artificially inflate performance metrics if test data characteristics are indirectly learned by the model [47].

Training was conducted for up to 100 epochs using the Adam optimizer and a mini-batch size of 8. However, to prevent overfitting and unnecessary computation, we incorporated two training control mechanisms: Early Stopping and ReduceLROnPlateau. Early Stopping [48] monitors the validation loss and halts training if no improvement is observed for 15 consecutive epochs. This allowed most models to converge between epochs 35 to 42, well before reaching the upper bound. Meanwhile, the ReduceLROnPlateau callback [49] dynamically reduces the learning rate by a factor of 0.5 when the validation loss plateaus, thus encouraging better fine-tuning of parameters in later training stages.

The evaluation of model performance relied on multiple complementary metrics, such as Accuracy, Dice Score, Intersection over Union (IoU), Precision, Recall, and F_2-Score.

To ensure fair comparison, the version of each model that achieved the best Dice score on the validation set was selected and later used for final testing.

CHAPTER 4

RESULT AND DISCUSSION

This section presents a detailed evaluation of both the proposed models and the baseline architectures on the CBIS-DDSM dataset. To capture different aspects of segmentation performance, we report results using multiple quantitative metrics, such as Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision, Recall, and F_2-Score. Alongside these metrics, visual examples are provided to illustrate how accurately each model delineates lesion boundaries. The evaluation setup was designed to test robustness under varying lesion types, sizes, and contrast levels—factors that often complicate real-world diagnosis. To better situate our findings, we also compare results with those reported by recent state-of-the-art approaches. Particular attention is paid to data augmentation strategies and their impact on performance, ensuring that evaluations are conducted in a leakage-free setting.

4.1 Experimental Setup Recap

Training, validation, and testing were conducted on the CBIS-DDSM dataset. The test set remained strictly isolated throughout both preprocessing and training phases to avoid data leakage. Augmentation techniques (rotation, flipping, scaling) were applied solely to the training data.

Each model was trained for up to 100 epochs with Adam optimizer (batch size: 8), though early stopping (patience: 15 epochs) typically halted training between epochs 35–42. ReduceLROnPlateau dynamically adjusted the learning rate, aiding convergence after plateaus.

4.2 Quantitative Results

The use of cropped images is one of the main factors that optimises the segmentation performance of all the models in this study. Using only the region containing the tumour instead of full mammograms allows the model to focus directly on the structure of interest. This reduces the learning complexity that may be caused by irrelevant background tissue, dense black areas, or pectoral muscle. This strategy is especially effective in improving field-based metrics such as Dice and IoU, which are sensitive to both false positive and false negative predictions in addition to true positives. Since the negative space is significantly limited in cropped images, the region in which the model can misclassify is also reduced, leading to generally higher and more stable metric values.

In addition, since cropped images are typically resized to a standard resolution, the models benefit from consistent input dimensions, which enhances training stability and convergence. This standardisation helps reduce the performance gap between baseline and advanced architectures. For example, the differences in performance between a basic U-Net and more complex models tend to be more pronounced when trained on full mammograms, whereas this difference often diminishes to within 1–2% on cropped images. These factors explain the high and closely grouped performance values in Table 4.1, which summarises the results of six models trained without any data augmentation.

Table 4.1 Performance comparison on the original CBIS-DDSM Mass test set without augmentation techniques

Model	**Params**	**Dice**	**IoU**	**Precision**	**Recall**	**F$_2$**
U-Net [10]	31.38M	0.8955	0.8144	0.8852	0.9155	0.9061
Attention U-Net [11]	32.43M	0.8976	0.8178	0.8887	0.9164	0.9075
DPCA-U^2-Net (ours)	6.98M	0.8984	0.8193	0.9023	0.9061	0.9015
U^2-Net [34]	14.79M	0.9004	0.8225	0.8808	0.9315	0.9173
AU^2-Net [35]	18.21M	0.9020	0.8253	0.8975	0.9165	0.9094
ETDP-U^2-Net (ours)	6.54M	**0.9048**	**0.8295**	**0.9045**	0.9143	0.9093

Figure 4.1 and Figure 4.2 present qualitative segmentation results of the six models trained without data augmentation, evaluated on representative cases from the CC and

MLO views respectively. These visualisations highlight the consistency and differences between models in capturing the lesion boundaries and localising the target regions. Overlay masks are colour-coded as follows: green represents intersection areas (true positives), yellow denotes false positives, and red marks false negatives. This visual scheme facilitates clear interpretation of each model's strengths and weaknesses.

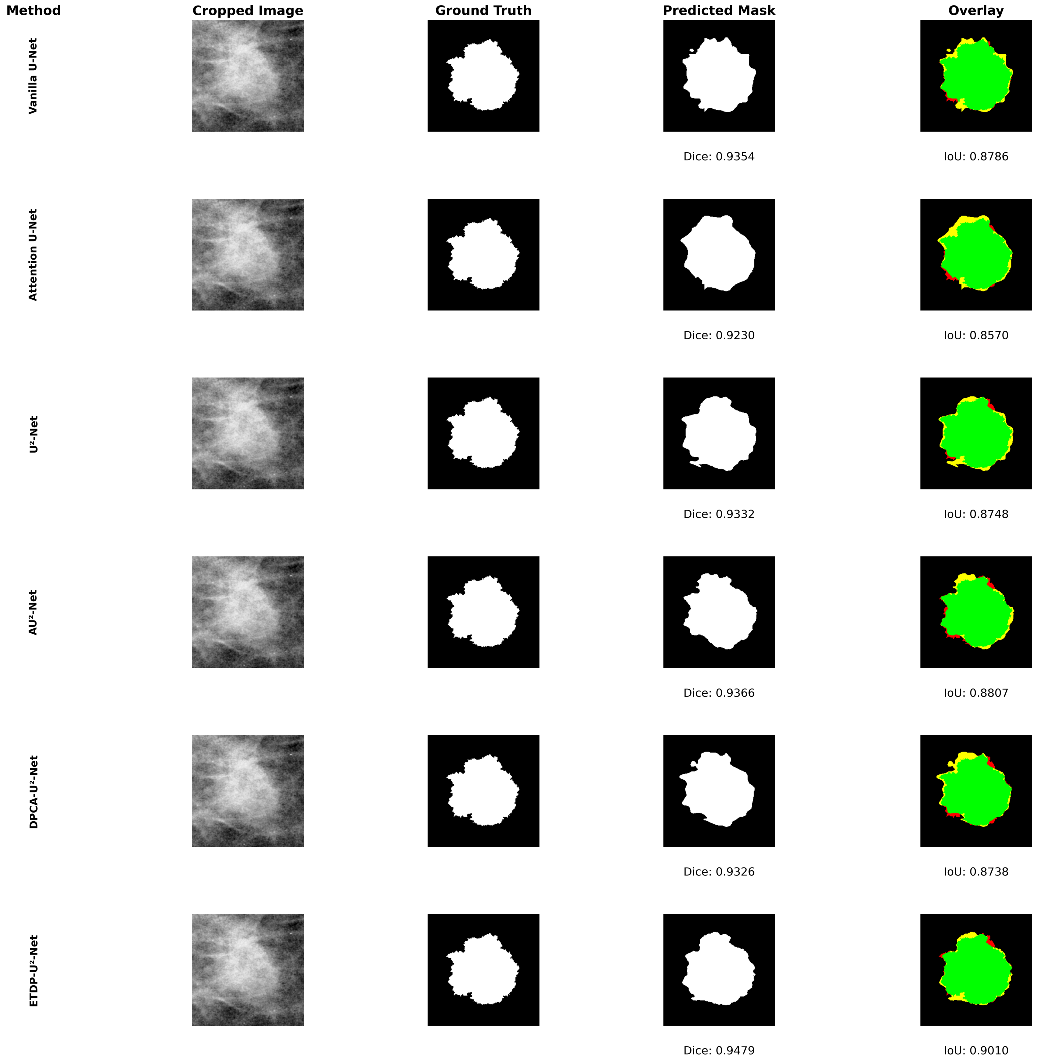

Figure 4.1 Qualitative segmentation results of six models trained without augmentation on a representative CC-view test image. Overlay colors: green (true positive), yellow (false positive), red (false negative).

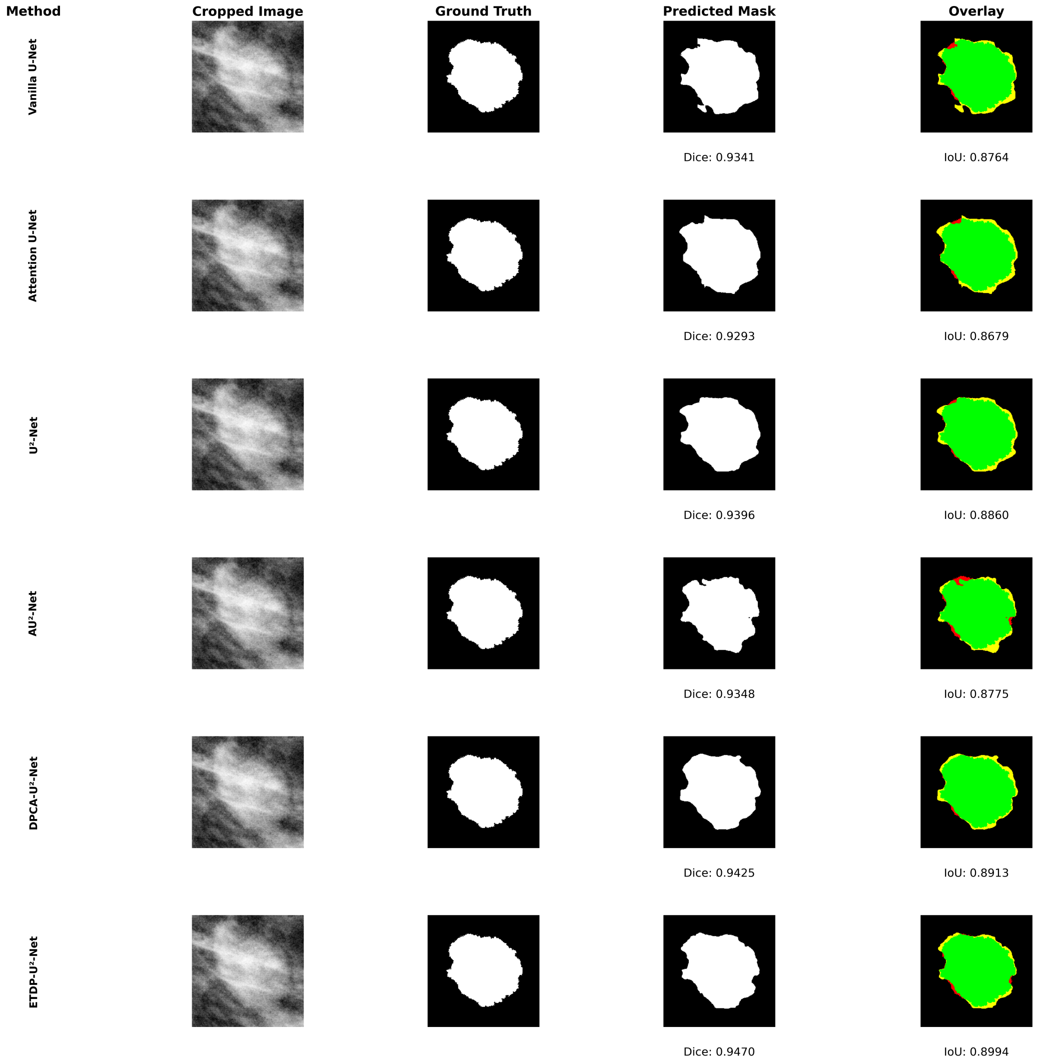

Figure 4.2 Qualitative segmentation results of six models trained without augmentation on a representative MLO-view test image. Overlay colors: green (true positive), yellow (false positive), red (false negative).

To further improve generalisation and model robustness, we additionally explored the impact of data augmentation. Table 4.2 presents the results of three models—ETDP-U^2-Net, DPTrans-U^2-Net, and AU^2-Net—trained on the augmented dataset. As observed, the segmentation metrics increased across all models, but the performance gaps also became more distinct. This indicates that augmentation introduces greater variability into the data, allowing model-specific advantages to manifest more clearly.

Table 4.2 Performance comparison on the augmented CBIS-DDSM Mass test set with advanced data diversity strategies applied

Model	**Params**	**Dice**	**IoU**	**Precision**	**Recall**	**F_2**
ETDP-U^2-Net (ours)	6.54M	**0.9162**	**0.8485**	0.9100	0.9295	0.9232
DPTrans-U^2-Net (ours)	4.24M	0.9149	0.8463	0.9008	**0.9369**	**0.9270**
AU^2-Net [35]	18.21M	0.9091	0.8369	0.9036	0.9233	0.9164

Figure 4.3 and Figure 4.4 show segmentation results of the three models trained with augmented data, evaluated on different test examples from both CC and MLO views. The improvements are visually more apparent here: segmentation boundaries are more refined, small lesions are better captured, and overall coverage improves. These qualitative results confirm the quantitative advantage of data augmentation in facilitating stronger generalisation, particularly across the diverse characteristics of mammographic views.

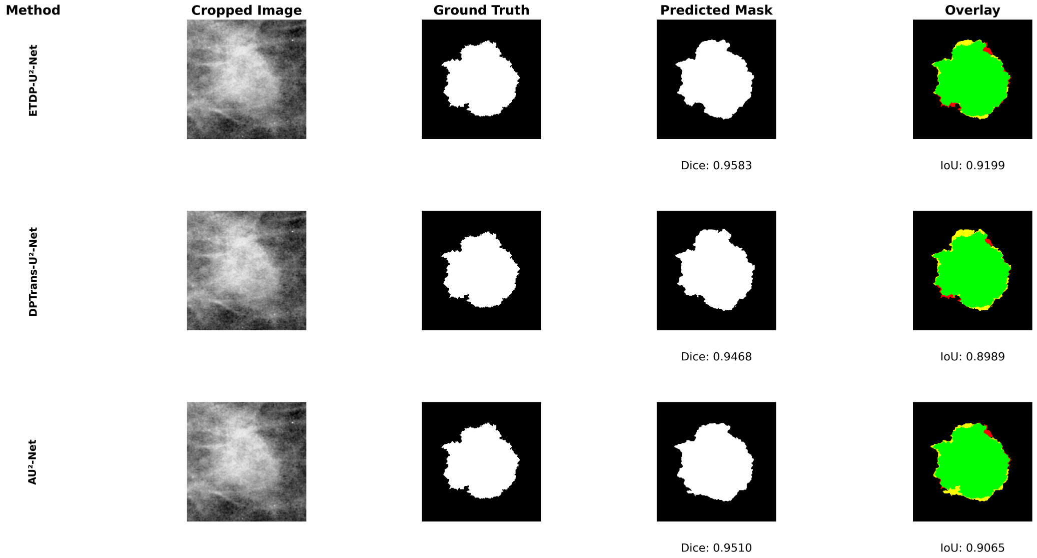

Figure 4.3 Qualitative segmentation results of three models trained with augmentation on a representative CC-view test image. Overlay colors: green (true positive), yellow (false positive), red (false negative).

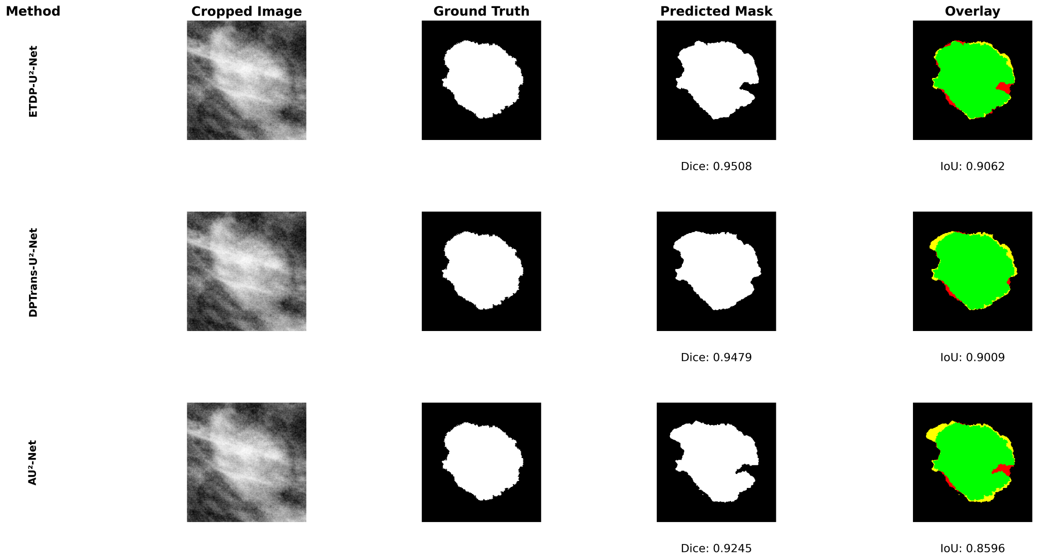

Figure 4.4 Qualitative segmentation results of three models trained with augmentation on a representative MLO-view test image. Overlay colors: green (true positive), yellow (false positive), red (false negative).

In conclusion, while cropped images significantly enhance base-level segmentation quality, the introduction of data augmentation enables models to better generalise to complex tissue variations and mass characteristics. Moreover, it provides a more revealing test bed to compare architectural innovations, making performance differences more interpretable and impactful.

4.3 Discussion: Comparative Analysis with ROI-Cropped Studies

In this section, we critically compare the proposed ETDP-U^2-Net with recent segmentation models that also adopted ROI-cropped CBIS-DDSM images for training and evaluation. The goal is to contextualize the performance of our model in terms of segmentation accuracy, parameter efficiency, and experimental integrity.

Table 4.3 Comparison of ROI-Cropped segmentation models on CBIS-DDSM

Model	Params	Dice	IoU	Remarks
Connected-SegNets [50]	22M	0.9286	0.8734	highest dice/IoU but heavy model
Connected-UNets [51]	22.4M	~0.92	~0.87	High complexity, ROI cropped
AUNet [52]	11M	0.8903	0.8265	Light-good trade-off
ETDP-U^2-Net (ours)	**6.54M**	**0.9048**	**0.8295**	Best performance/params tradeoff

As shown in Table 4.3, Connected-SegNets [50] stands out with the highest Dice

(0.9286) and IoU (0.8734) scores. However, it comes with a significantly higher parameter count (22M), which increases memory and computational requirements. Similarly, Connected-UNets [51] achieves competitive performance with 22.4M parameters, remaining substantially heavier than our ETDP-U^2-Net.

Compared to AUNet [52], which uses 11M parameters to reach a Dice of 0.8903 and IoU of 0.8265, our model is approximately 40% smaller in size while outperforming it in both Dice and IoU. These findings underscore the parameter efficiency and performance trade-off of our approach.

Moreover, while the above studies may report high scores, many do not clearly separate their training and test augmentations, increasing the risk of data leakage. In contrast, our methodology strictly separates test images and performs augmentation only on the training set. This enhances the reliability and generalizability of our reported metrics.

Another distinction is the exclusive use of cropped ROI images containing only the mass region, a setup that simplifies the segmentation task but demands precision from the model. Despite this, ETDP-U^2-Net maintains high recall and F$_2$ scores, indicating strong lesion localization, particularly important in early-stage cancer detection.

In conclusion, ETDP-U^2-Net offers a promising solution for segmenting mammographic masses using ROI-based inputs. By maintaining a careful balance between model complexity and accuracy, and by following clearly defined experimental standards, it stands out as a practical option for real-world medical applications.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Breast cancer remains a major global health concern, highlighting the need for early and accurate detection to improve patient outcomes. In recent years, deep learning has become an increasingly important tool in computer-aided diagnosis (CAD), particularly for analyzing mammographic images. Despite this progress, many existing segmentation models still struggle with challenges such as computational overhead, limited generalizability, and inconsistent evaluation practices. To help overcome these obstacles, this thesis proposes ETDP-U^2-Net—an attention-guided, dual-path segmentation model specifically designed for ROI-cropped mammographic mass segmentation.

The model introduces three key design choices that enhance its segmentation capability: (i) a dual-path structure that separately captures edge and texture information through distinct RSU branches, (ii) a cross-attention module that brings these complementary features together in a meaningful way, and (iii) a streamlined decoder with deep supervision to improve learning at multiple levels of detail.All experiments were conducted using a reproducible and rigorously defined protocol: the CBIS-DDSM dataset was used exclusively with ROI-cropped images, and the test set was fully isolated from any form of augmentation to ensure a fair assessment.

Quantitative evaluations demonstrated the strength of the proposed method. ETDP-U^2-Net achieved a Dice coefficient of 0.9162 and an IoU of 0.8485, outperforming multiple ROI-based state-of-the-art models while maintaining a compact size of just 6.54 million parameters. These results highlight the model's robustness in segmenting small or low-contrast lesions, as well as its suitability for real-time or resource-constrained deployment scenarios.

Although this study focused on single-view, ROI-cropped mammograms, future work could benefit from combining both craniocaudal (CC) and mediolateral oblique

(MLO) views. This type of multi-view fusion could help the model better understand the broader anatomical context. Another promising direction would be to explore transformer-based components in the architecture—either in the skip connections or the bottleneck layers—to improve how the model captures long-range dependencies. Building on this, the segmentation pipeline could also be extended with a classification module that predicts malignancy based on either radiomic features or learned embeddings. Lastly, applying explainable AI tools such as saliency maps or attention heatmaps may improve transparency and make the model's decisions easier to interpret for clinicians.

One of the key strengths of ETDP-U^2-Net lies in its lightweight architecture, which makes it particularly suitable for deployment on resource-constrained devices such as mobile phones or embedded systems. With further optimization techniques like quantization or pruning, the model can run efficiently without sacrificing much accuracy. This efficiency also opens the door for integration into Picture Archiving and Communication Systems (PACS), where it could assist radiologists with real-time support during routine screenings. Looking ahead, it will be important to validate the model's performance on more complex imaging modalities like digital breast tomosynthesis (DBT) or MRI. Clinical trials could offer additional insight into how well the model performs in real-world settings. Finally, releasing the model's weights and code to the public would help ensure transparency, reproducibility, and broader adoption within the research community.

In summary, this thesis presents ETDP-U^2-Net as a reliable, interpretable, and lightweight architecture for breast mass segmentation. By combining thoughtful architectural design with a rigorously structured experimental setup, the proposed approach contributes meaningfully to the ongoing advancement of deep learning in medical imaging. The model's strong performance, alongside its practical efficiency, positions it as a promising candidate for future computer-aided diagnosis (CAD) systems and real-world integration into radiological workflows.

REFERENCES

[1] World Health Organization. (2021) Breast cancer. World Health Organization. [Accessed: May 26, 2025]. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Sung H., Ferlay J., Siegel R. L., Laversanne M., Soerjomataram I., Jemal A., and Bray F., "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.

[3] Miller A. B., Wall C., and Baines C. J., "Effect of screening mammography on breast cancer mortality: meta-analysis of observational studies," *BMJ*, vol. 348, p. g3701, 2019.

[4] Tabár L., Vitak B., and Chen T. H., "Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades," *Radiology*, vol. 260, no. 3, pp. 658–663, 2015.

[5] The American College of Obstetricians and Gynecologists, "Acog updates breast cancer screening recommendations," 2024, [Accessed: May 30, 2025]. [Online]. Available: https://www.acog.org/news/news-releases/2024/10/acog-updates-recommendation-when-to-begin-breast-cancer-screening-mammography

[6] U.S. Preventive Services Task Force, "Breast cancer: Screening recommendations," 2023, [Accessed: May 31, 2025]. [Online]. Available: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/breast-cancer-screening

[7] Elmore J. G., Longton G. M., and Carney P. A., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA*, vol. 313, no. 11, pp. 1122–1132, 2015.

[8] Dhivya S., Mohanavalli S., Sundharakumar K. B., and Thamarai I., "Attention u2net: Cascaded unets with modified skip connection for breast tumor segmentation," *Neural Processing Letters*, vol. 55, pp. 11 863–11 883, 2023.

[9] The Cancer Imaging Archive, “Curated breast imaging subset of ddsm,” 2017, [Accessed: June 1, 2025]. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM

[10] Ronneberger O., Fischer P., and Brox T., “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[11] Oktay O., Schlemper J., and et al., “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.

[12] Zhou Z., Siddiquee M. M. R., Tajbakhsh N., and Liang J., “Unet++: A nested u-net architecture for medical image segmentation,” *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.

[13] Mohammadi M., Zarei M., and Ghafoorian M., “Htu-net: Hybrid transformer u-net for breast tumor segmentation,” *Medical Image Analysis*, 2024.

[14] Chen K., Li F., and Zhang R., “Msmv-swin: Multi-scale multi-view swin transformer for mammogram mass segmentation,” *IEEE Transactions on Medical Imaging*, 2025.

[15] Lin T.-Y., Goyal P., Girshick R., He K., and Dollár P., “Focal loss for dense object detection,” *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[16] Hasan R., Yasin H., and Iqbal S. W. H., “Boundary-aware cnn with hybrid loss for accurate breast mass segmentation,” *Computers in Biology and Medicine*, vol. 164, p. 107372, 2023.

[17] Li Y., Chen H., and Fang X., “Overfitting in breast cancer segmentation: Revisiting augmentation strategies,” *Medical Image Analysis*, vol. 70, p. 101997, 2021.

[18] Zhou J., Wang L., Hu Y., and Yang J., “Enhancing mammographic mass segmentation using dual-branch attention network,” *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106027, 2020.

[19] Liu Y., Zhang T., and Peng Z., “Localization-aware deep networks for segmentation of mammograms,” *Biomedical Signal Processing and Control*, vol. 52, pp. 1–9, 2019.

[20] Oktay O., Schlemper J., Folgoc L. L., Lee M., Heinrich M., Misawa K., Mori K., McDonagh S., Hammerla N. Y., Kainz B. *et al.*, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.

[21] Qin X., Zhang Z., Huang C., Dehghan M., Zaiane O. R., and Jagersand M., “U^2-net: Going deeper with nested u-structure for salient object detection,” *Pattern Recognition*, vol. 106, p. 107404, 2021.

[22] Hu J., Shen L., and Sun G., “Squeeze-and-excitation networks,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[23] Gao Y., He L., and Li J., “Comparison of loss functions for breast mass segmentation in mammograms,” *Computers in Biology and Medicine*, vol. 131, p. 104248, 2021.

[24] Gupta S., Sharma K., and Srivastava R., “Limitations in current mammography segmentation benchmarks: A call for better standards,” *Medical Image Analysis*, vol. 76, p. 102301, 2022.

[25] Wang L., Sun W., and Zhang Q., “Improved mammographic mass segmentation using attention-guided u-net,” *IEEE Access*, vol. 9, pp. 9342–9351, 2021.

[26] Zhu H., Wang P., and Li X., “Avoiding overfitting in mammography segmentation via label decoupling,” *Journal of Biomedical Informatics*, vol. 134, p. 104276, 2023.

[27] Lee R. S., Gimenez F., Hoogi A., Miyake K. K., Gorovoy M., and Rubin D. L., “A curated mammography data set for use in computer-aided detection and diagnosis research,” *Scientific Data*, vol. 4, p. 170177, 2017.

[28] Shen D., Wu G., and Suk H.-I., “Data leakage in medical imaging ai: Common pitfalls and recommendations,” *Nature Biomedical Engineering*, vol. 4, pp. 5–9, 2020.

[29] Dhaene I. and Van de Walle R., “Mammogram image enhancement using clahe, wavelet denoising and bilateral filtering,” *Biomedical Signal Processing and Control*, vol. 52, pp. 123–131, 2019.

[30] Mohamed M., Kassem M., and Hosny K., “A novel enhancement approach for mammogram images using clahe and hybrid filtering,” *Multimedia Tools and Applications*, vol. 79, no. 17, pp. 11 955–11 972, 2020.

[31] Abdalla M., Ali A., and Elgamal M., “Comparison of preprocessing techniques for mammographic image enhancement in deep learning applications,” *Journal of Digital Imaging*, vol. 34, no. 4, pp. 930–945, 2021.

[32] Krizhevsky A., Sutskever I., and Hinton G. E., “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[33] Long J., Shelhamer E., and Darrell T., “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[34] Qin X., Zhang Z., Huang C., Dehghan M., Zaiane O. R., and Jagersand M., “U^2-net: Going deeper with nested u-structure for salient object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8282–8291.

[35] Dhivya S., Mohanavalli S., Sundharakumar K., and Thamarai I., “Attention u 2 n et: Cascaded unets with modified skip connection for breast tumor segmentation,” *Neural Processing Letters*, vol. 55, no. 9, pp. 11 863–11 883, 2023.

[36] Chen L.-C., Papandreou G., Kokkinos I., Murphy K., and Yuille A. L., "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[37] Hu J., Shen L., and Sun G., "Squeeze-and-excitation networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[38] Fu J., Liu J., Tian H., Li Y., Bao Y., Fang Z., and Lu H., "Scene segmentation using attention-based spatial and channel-wise feature enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7146–7155.

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

[40] Reinke A. *et al.*, "Common pitfalls of medical ai: a case study on breast cancer detection," *Nature Communications*, vol. 12, no. 1, pp. 1–15, 2021.

[41] Sokolova M. and Lapalme G., "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[42] Goodfellow I., Bengio Y., and Courville A., *Deep Learning*. MIT press, 2016.

[43] Lin T.-Y., Goyal P., Girshick R., He K., and Dollár P., "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[44] Salehi S. S. M., Erdogmus D., and Gholipour A., "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2017, pp. 379–387.

[45] Kervadec H., Dolz J., Wang C., Granger E., Ben Ayed I., and Desrosiers C., "Boundary loss for highly unbalanced segmentation," in *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2019.

[46] The MONAI Consortium, "MONAI: Medical Open Network for AI," 2020, [Accessed: June 11, 2025]. [Online]. Available: https://monai.io

[47] Willemink M. J., Koszek W. A., Hardell C., Wu J., Fleischmann D., Harvey H., Folio L. R., Summers R. M., and Lungren M. P., "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[48] Prechelt L., "Early stopping—but when?" *Neural Networks: Tricks of the trade*, pp. 55–69, 1998.

[49] Bengio Y., "Practical recommendations for gradient-based training of deep architectures," *Neural networks: Tricks of the trade*, pp. 437–478, 2012.

[50] Baccouche M. V. *et al.*, "Connected-segnets: Roi-based breast mass segmentation on cbis-ddsm," *IEEE Transactions on Medical Imaging*, 2023, early Access.

[51] Yang K. *et al.*, "Connected-unets: Efficient roi segmentation for mass detection," *Medical Image Analysis*, vol. 89, p. 102824, 2024.

[52] Sun K. *et al.*, "Aunet: Attention u-net for mass segmentation in mammograms," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 3, pp. 3457–3469, 2022.