5-2023

Deep Learning With Attention Mechanisms in Breast Ultrasound Image Segmentation and Classification

Meng Xu
Utah State University

DEEP LEARNING WITH ATTENTION MECHANISMS IN BREAST ULTRASOUND IMAGE SEGMENTATION AND CLASSIFICATION

by

Meng Xu

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Computer Science

Approved:

Xiaojun Qi, Ph.D. Major Professor	Vicki Allan, Ph.D. Committee Member
David Brown, Ph.D. Committee Member	John Edwards, Ph.D. Committee Member
Haitao Wang, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2023

ABSTRACT

Deep Learning with Attention Mechanisms in
Breast Ultrasound Image Segmentation and Classification

by

Meng Xu, Doctor of Philosophy
Utah State University, 2023

Major Professor: Xiaojun Qi, Ph.D.
Department: Computer Science

Breast cancer is a great threat to women's health. Breast ultrasound (BUS) imaging is commonly used in the early detection of breast cancer as a portable, valuable, and widely available diagnosis tool. Automated BUS image segmentation and classification can assist radiologists in making accurate and fast decisions. Deep neural networks have recently been employed to achieve better image segmentation and classification results than conventional approaches. In this dissertation, we introduce three different deep learning architectures, each of which aims to address the drawbacks of their peers and evaluate their performance in terms of segmentation and classification accuracy on two public BUS datasets. The first developed method is called a Multi-Scale Self-Attention Network (MSSA-Net), which can be trained on small datasets to explore relationships between pixels to achieve better segmentation accuracy. Specifically, Our MSSA-Net integrates rich local features and global contextual information at different scales and applies self-attention to multi-scale feature maps. The second developed method is called a Multi-Task Learning Network with Context-Oriented Self-Attention (MTL-COSA) to automatically and simultaneously segment tumors and classify them as benign or malignant. The COSA module incorporates prior medical knowledge to guide the network to learn contextual relationships for better feature representations in BUS images to improve both segmentation and classification performance. The third developed method is called a Regional-Attentive Multi-Task Learning framework (RMTL-Net), which simultaneously segments tumor regions in BUS images and classifies tumors into benign or malignant categories. To improve both segmentation and classification accuracy, we design a Regional Attention (RA) module that employs the segmentation output to automatically guide the classifier to learn important category-sensitive information in the tumor, peritumoral, and background regions and seamlessly fuse them to achieve better classification accuracy. We compare the performance of the three proposed deep learning architectures with state-of-the-art segmentation and classification methods by conducting extensive experiments on two publicly available BUS datasets, including Dataset UDIAT and Dataset BUSI.

(99 pages)

PUBLIC ABSTRACT

Deep Learning with Attention Mechanisms in
Breast Ultrasound Image Segmentation and Classification
Meng Xu

Breast cancer is a great threat to women's health. Breast ultrasound (BUS) imaging is commonly used in the early detection of breast cancer as a portable, valuable, and widely available diagnosis tool. Automated BUS image analysis can assist radiologists in making accurate and fast decisions. Generally, automated BUS image analysis includes BUS image segmentation and classification. BUS image segmentation automatically extracts tumor regions from a BUS image. BUS image classification automatically classifies breast tumors into benign or malignant categories. Multi-task learning accomplishes segmentation and classification simultaneously, which makes it more appealing and practical than an either individual task. Deep neural networks have recently been employed to achieve better image segmentation and classification results than conventional approaches. In addition, attention mechanisms are applied to deep neural networks to make them focus on the important parts of the input to improve the segmentation and classification performance. However, BUS image segmentation and classification are still challenging due to the lack of public training data and the high variability of tumors in shape, size, and location.

In this dissertation, we introduce three different deep learning architectures with attention mechanisms, each of which aims to address the drawbacks of their peers and evaluate their performance in terms of segmentation and classification accuracy on two public BUS datasets. First, we propose a Multi-Scale Self-Attention Network (MSSA-Net) for BUS image segmentation that can be trained on small BUS image datasets. We design a multi-scale attention mechanism to explore relationships between pixels to improve the feature representation and achieve better segmentation accuracy. Second, we propose a Multi-Task Learning Network with Context-Oriented Self-Attention (MTL-COSA) to segment tumors and classify them as benign or malignant automatically and simultaneously. We design a COSA attention mechanism that utilizes segmentation outputs to estimate the tumor boundary, which is treated as prior medical knowledge, to guide the network to learn contextual relationships for better feature representations to improve both segmentation and classification accuracy. Third, we propose a Regional-Attentive Multi-Task Learning framework (RMTL-Net) for simultaneous BUS image segmentation and classification. We design a regional attention mechanism that employs the segmentation output to guide the classifier to learn important category-sensitive information in three regions of BUS images and fuse them to achieve better classification accuracy. We conduct experiments on two public BUS image datasets to show the superiority of the proposed three methods to several state-of-the-art methods for BUS image segmentation, classification, and Multi-task learning.

To my parents, who gave me the gift of life and the strength to persevere, and to my husband, Kuan Huang, who completed my soul.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my Ph.D. advisor, Dr. Xiaojun Qi, for her dedicated support and mentorship throughout my Ph.D. journey. Her extensive expertise and patient guidance have been invaluable in my research, career, and life. Dr. Qi has also taught me the necessary skills to be an independent researcher and a strong person. Her endless support and encouragement helped me stay positive during the most challenging times.

I would also like to thank other members of my Ph.D. committee: Dr. Vicki Allan, Dr. David Brown, Dr. John Edwards, and Dr. Haitao Wang, for their insightful comments and consistent encouragement throughout my doctoral studies.

In addition, I would like to express my sincere gratitude to my family. My loving parents have always been my source of strength and bravery from thousands of miles away. My parents-in-law have always been kind and supportive. My husband has always stood by my side and supported me in chasing my dreams. My cat Coco has companioned me through days and nights and provided comfort and stress relief. Thank you all for your love, care, and support.

Lastly, I would like to thank my friends who have helped and supported me during the darkest times of my Ph.D. journey: Qiuxiao Chen, Junnan Geng, Haixuan Guo, Peiyu Li, and Mohammadreza Javanmardi. I also want to thank my friends who have made my days at Utah State University pleasant: Cheng Chen, Xiaobei Chen, Jiyao Li, Xiankun Yan, Yiming Zhao, Fei Xu, Guoqin Ding, Xiaoyin Zhang, Shida Zhong, and Amir Hossein Farzaneh.

Meng Xu

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

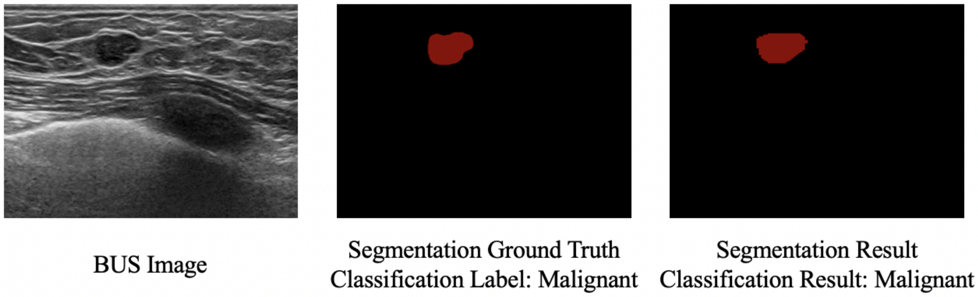
BUS	Breast UltraSound
CAD	computer-aided diagnosis
MTL	multi-task learning
MSE	mean squared error
BCE	binary cross-entropy
CCE	categorical cross-entropy
SGD	stochastic gradient descent
Adam	ADAptive Moment estimation
GD	gradient descent
CNN	convolutional neural network
RoI	region of interest
SVM	support vector machine
KNN	K-nearest neighbors
MSSA	multi-scale self-attention
COSA	context-oriented self-attention
RMTL	regional-attentive multi-task learning
RA	regional attention (related to RMTL)

CHAPTER 1

INTRODUCTION

Breast cancer is a significant threat to women's health and is the most commonly diagnosed cancer and the leading cause of cancer mortality among women worldwide in 2020. It accounts for 1 in 4 cancer cases and 1 in 6 cancer deaths in women [1]. A forecast indicates that breast cancer will result in more than 3 million new cases and 1 million deaths by 2040 [2]. Breast cancer mortality rates are much higher in low- and middle-income countries than in high-income countries due to the delayed detection and treatment [3, 4]. Early diagnosis and appropriate treatments can significantly increase survival rates. Mammography and breast ultrasound (BUS) are two popular screening modalities for early breast cancer detection. BUS has been commonly used in the early diagnosis of breast cancer in women of all ages, especially in low- and middle-income countries, because it is portable, widely available, low-cost, and highly sensitive [5, 6].

Computer-aided-diagnosis (CAD) systems are proposed to help radiologists interpret BUS images, make a more accurate diagnosis, and reduce their workload [7, 8]. In general, a CAD system for breast cancer detection includes automated segmentation and classification as two primary steps for further processing. BUS image segmentation automatically extracts tumor regions from a BUS image. Accurate segmentation can assist radiologists in identifying and locating breast tumors precisely. In addition, it can aid in visualizing and tracking changes in breast tumors over time, which enables radiologists to easily monitor the progress of breast cancer and the efficacy of treatments. BUS image classification automatically classifies breast tumors into benign or malignant categories. Multi-task learning (MTL) simultaneously accomplishes BUS image segmentation and classification, which makes it more appealing and practical than either individual segmentation or classification. Figure 1.1 shows an example of a BUS image, its segmentation ground truth and classification label, and segmentation and classification results.



BUS Image

Segmentation Ground Truth
Classification Label: Malignant

Segmentation Result
Classification Result: Malignant

Fig. 1.1: An example of a BUS image, its segmentation ground truth and classification label, and its segmentation and classification results.

Given a BUS image, a BUS image segmentation/classification/MTL CAD system consists of two major components: (1) a feature extraction module to represent breast tumors in numerical features and (2) a feature segmentation module to draw the tumor contour, a feature classification module to predict a benign or malignant tumor, or both. Figure 1.2 shows a high-level diagram of a BUS image segmentation CAD system, a BUS image classification CAD system, and a BUS image MTL CAD system. First, an input BUS image is fed into a feature extraction module to extract features that are most relevant to the specific task for later steps. Next, a trainable segmentation/classification/MTL module uses a machine learning algorithm to segment the tumor region from the input image, categorize the input image as benign or malignant, or do both based on extracted features.

Automated analysis of BUS images can help radiologists make efficient diagnoses of breast cancer. However, it is still challenging due to the lack of public training data and the high variability of tumors in shape, size, and location [9, 10]. Supervised learning CAD methods for image segmentation and classification require a sufficient number of labeled training data. The quality of the manual annotation process is a vital factor in determining the performance of the developed CAD methods. But acquiring labeled data is time-consuming and labor-intensive, especially for medical images. For example, BUS images need to be manually labeled by experienced radiologists. For each BUS image, radiologists need to assess whether there is a tumor, classify the tumor into benign or malignant categories, identify the tumor regions, and draw the tumor contours using specialized software tools. The high cost of manual annotation and the need to protect patient privacy lead to

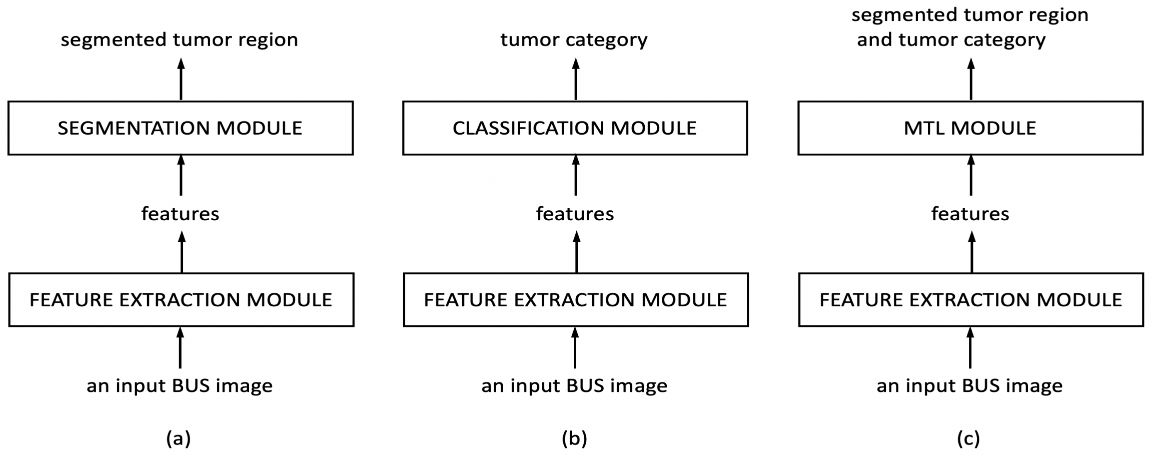


Fig. 1.2: A high-level representation of (a) a BUS image segmentation CAD system, (b) a BUS image classification CAD system, and (c) a BUS image MTL CAD system. The input of three types of CAD systems is a BUS image, and the output is a segmented tumor region and/or tumor category.

the shortage of high-quality, publicly accessible BUS image datasets for research purposes. There are only two commonly used high-quality public BUS image datasets, dataset UDIAT [11] and dataset BUSI [12], including only 943 images in total. A few other public BUS datasets either lack pixel-wise segmentation ground truth of partial or all images or provide cropped images containing breast tumors at the center with a limited amount of surrounding background. Therefore, these datasets are not included in our study. The limited training data makes it more challenging to train a robust CAD method for BUS image segmentation or classification. In addition, tumors in BUS images exhibit significant variation in shape, size, and location, making segmentation a daunting task. For example, Figure 1.3 shows six examples of BUS images from each of the two datasets that contain benign and malignant tumors, respectively. Red lines delineate tumor regions in various shapes, sizes, and locations.

In this dissertation, we focus on developing novel deep learning-based CAD methods for the automated analysis of BUS images. We propose three methods, including a method for BUS image segmentation and two MTL methods for simultaneous BUS image segmentation and classification. The remainder of this dissertation is organized as follows: Chapter 2 provides the background of deep learning and attention mechanism and introduces related

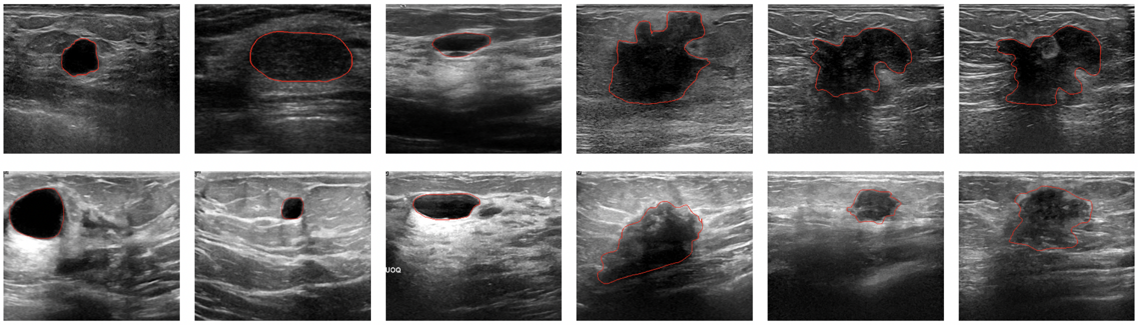


Fig. 1.3: Illustration of BUS images containing benign tumors (the first three columns) and malignant tumors (the last three columns). The first row shows images from dataset UDIAT and the second row shows images from dataset BUSI.

works of BUS image segmentation, classification, and MTL performing segmentation and classification tasks at the same time. Chapters 3, 4, and 5 introduce our proposed MSSA-Net, MTL-COSA, and RMTL-Net, respectively. Chapter 6 presents datasets, evaluation metrics, segmentation and classification results of the proposed three methods on two public BUS datasets, and their comparison with the state-of-the-art methods. Chapter 7 presents the comparison between the three proposed methods and discusses the advantages, potential usefulness, limitations, and future work of the proposed methods. Chapter 8 draws a conclusion. To reduce the notations and increase the readability, notations in each chapter are only applicable within the chapter itself.

CHAPTER 2
RELATED WORKS

2.1 Deep Learning

Deep learning is a subfield of machine learning that uses artificial neural networks to solve problems in various fields, such as computer vision, natural language processing, speech recognition, and robotics. Compared to traditional machine learning algorithms, deep learning algorithms tend to have better performance on complex tasks, more efficient feature engineering, and more flexibility and scalability [13–15]. For example, deep learning models can better fit complex non-linear patterns, which makes them work better on complex tasks (*e.g.*, image recognition, scene understanding, object tracking, etc.) in the real world. In addition, deep learning models automatically extract the most relevant features from the input data without or with little human involvement. They can further handle a wide range of data types and large volumes of data. Deep learning algorithms can be categorized into three main types: supervised learning, unsupervised learning, and semi-supervised learning [16]. Supervised learning requires all training data to be labeled to train a model, while unsupervised learning does not require training data to be labeled to train a model. Semi-supervised learning uses some labeled training data and some unlabeled training data to train a model.

In this dissertation, we focus on developing supervised deep learning-based architectures for Breast UltraSound (BUS) image segmentation and classification. In this section, we introduce the relevant mathematical background of deep learning used in our work. Specifically, we provide a comprehensive overview of the fundamental concepts in supervised learning, loss function, optimization, and Convolutional Neural Networks (CNNs) to build a BUS image segmentation and classification system.

2.1.1 Supervised Learning

In supervised learning, the model is trained on a labeled dataset, where each input is paired with an output. We call the output a ”label” or ”ground truth,” the true answer to the problem. A supervised learning model learns a pattern (function) that maps from the input to the output during training. The trained model then makes predictions for new input based on the pattern it has learned from the labeled dataset. Following [17], given a training dataset of input-output pairs $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, a supervised learning problem can be formulated as

$$f : X \to Y \tag{2.1}$$

where $X = \{x_1, x_2, ..., x_N\}$ is the input space and $Y = \{y_1, y_2, ..., y_N\}$ is the output space. Each input-output pair was generated by an unknown function $y_i = f(x_i)$, where $i \in \{1, ..., N\}$ and N is the number of input (i.e., training data). The goal of a supervised learning problem is to find a function h that approximates the function f. The function h is called a hypothesis that is drawn from a hypothesis space $\mathcal{H}$ of possible functions. We call the learned function h a trained model of the training data. For each input x_i in the input space, the model makes a prediction $\hat{y_i} = h(x_i)$. We cannot expect an exact match between h and f, but we hope they are as close enough so that the model can make an accurate prediction $\hat{y_i} \approx y_i$ for any input. More formally, we find a best-fit function h in the space $\mathcal{H}$ of possible functions by minimizing a loss function $\mathcal{L}(\hat{y_i}, y_i)$ over all samples in the training dataset:

$$h^\star = \underset{h \in \mathcal{H}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, h(x_i)) \tag{2.2}$$

where $\mathcal{L}(y_i, h(x_i))$ measures the difference between the predicted value $\hat{y_i} = h(x_i)$ and the actual value y_i. The function $h^\star$ is the best-fit function we are looking for.

For BUS image segmentation and classification, the input is a set of N BUS images $X = \{x_1, x_2, ..., x_N\}$. The label of BUS image classification is $Y_{cls} = \{y_1, y_2, ..., y_N\}$, where each y_i is either benign or malignant. For BUS image segmentation, each pixel of a BUS image is a training sample. In other words, segmentation is a pixel-wise classification.

The ground truth of BUS image segmentation is a binary image of the same size as the input, where 0 represents a background pixel and 1 represents a tumor pixel, *i.e.*, $Y_{seg} = \{y_{1,j}, y_{2,j}, ..., y_{N,j}\}$, where $j \in \{1, ..., M\}$ and M is the total number of pixels in a BUS image. Training a robust deep learning-based, supervised learning model needs a large amount of training data. For BUS image segmentation, our two small-size datasets are sufficient because each pixel is a training sample. However, they are insufficient for training a classification model. To solve the problem, we propose to do Multi-Task Learning (MTL) of BUS image segmentation and classification. Training an MTL model for multiple tasks can help reduce the amount of data required for each individual task and can lead to better performance on all tasks. MTL helps to improve the efficiency of learning by feature sharing between tasks. In an MTL model consisting of a segmentation network and a classification network, the classification network takes advantage of shared features learned from the segmentation network and therefore achieves better classification results than a single-task classification network. For an MTL model, the label set includes segmentation ground truths and classification labels, *i.e.*, $Y_{MTL} = \{(y_1, y_{1,j}), (y_2, y_{2,j}), ..., (y_i, y_{i,j}), ..., (y_N, y_{N,j})\}$, where $i \in \{1, ..., N\}$ and $j \in \{1, ..., M\}$.

2.1.2 Loss Function

In deep learning, the loss function quantifies the difference between the predicted values and the true values, and the goal of training a deep learning model is to minimize this difference or loss. A good loss function is important because it accurately evaluates the model's performance on a task during training and testing. In addition, the loss function is used to optimize the trainable parameters of the model. A good loss function trains a model that can make accurate predictions on new input data. In this section, we briefly introduce three commonly used loss functions in deep learning for computer vision tasks, including Mean squared error (MSE), binary cross-entropy (BCE), and categorical cross-entropy (CCE).

The MSE loss is commonly used for image regression tasks. It measures the average squared difference between the predicted values and true values. Given a set of N training samples, the MSE loss is defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.3}$$

where y_i is the true label and $\hat{y}_i$ is the predicted value of the i^{th} training sample. They are both continuous values in a regression problem. The MSE loss is simple and easy to interpret. It penalizes large errors more heavily than small errors, which is often desirable in regression tasks. However, the MSE loss is sensitive to outliers, which makes it heavily influenced by extreme values in the training data.

The BCE loss is also known as log loss. It is commonly used for binary classification problems. It measures the difference between the predicted values and true values of a binary classification task. The BCE loss is defined as:

$$BCE = -\frac{1}{N} \sum_{i=1}^{N} y_i * \log \hat{y}_i + (1 - y_i) * \log (1 - \hat{y}_i) \tag{2.4}$$

where y_i is either 0 or 1, representing two categories, and $\hat{y}_i$ is a value in the range [0,1], representing the probability of the i^{th} training sample belonging to a category.

The CCE loss is a popular choice for multi-class classification problems. It is also known as Softmax loss. Assuming there are C classes, the CCE loss is defined as:

$$CCE = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{i,j} * \log \hat{y}_{i,j} \tag{2.5}$$

where $y_{i,j}$ and $\hat{y}_{i,j}$ are the true label and predicted probability of the i^{th} image that belongs to the j^{th} class, respectively. Note that y_i is a one-hot encoded vector representing the actual categories, and $y_{i,j}$ is a vector of probability scores in range [0, 1] representing the predicted categories. For example, $y_i = [1, 0, 0, 0]$ means the training sample belongs to the first category in a four-category classification problem. And $\hat{y}_i = [0.2, 0.2, 0.4, 0.2]$ is a vector of probabilities of the training sample belonging to each category.

The cross-entropy loss is easy to implement and computationally efficient, which makes it suitable for large-scale data. In addition, the log operation keeps gradients from varying too widely, which makes it suitable for gradient-based optimization methods like stochastic gradient descent (SGD). BUS image segmentation is a pixel-wise binary classification problem, and BUS image classification is a binary classification problem. Considering the advantages of cross-entropy loss, we adopt the BCE loss as the loss function for both the segmentation and classification tasks of our three proposed methods.

2.1.3 Optimization

Most deep learning algorithms involve some kind of optimization, which refers to the task of either minimizing or maximizing a loss function. In deep learning, an optimizer is an algorithm that updates the parameters of a model in order to minimize the loss function during training. The goal of the optimizer is to find the optimal values of the model parameters that result in the lowest loss on the training dataset. Mini-batch SGD and Adaptive Moment Estimation (Adam) are two commonly used optimizers in computer vision.

Following [18], suppose we have a function $y = h(x)$ where both x and y are real numbers. Denote the derivative of this function as $h'(x)$. To reduce $h(x)$, we can move x in small steps to the direction of $-h'(x)$. We call this technique Gradient Descent (GD). The GD optimizer iteratively updates the parameters of a deep learning model until the loss function converges to a minimum or until some other stopping criterion is met. When $h'(x) = 0$, the derivative gives no information about which direction to move. Points where $h'(x) = 0$ are called critical points. A local minimum is a point where $h(x)$ is smaller than all neighboring points in a small range of x values. A global minimum is a point where $h(x)$ is the smallest value for all possible x values. In deep learning, a loss function can have multiple local minima that are not optimal. It is very difficult to find a global minimum for all problems, especially when the loss function takes multidimensional inputs. Therefore, we expect to find a $h(x)$ value that is small enough but not necessarily global minimal for all possible x values in real-world deep learning problems. In real-world problems like BUS image segmentation and classification, the loss function takes multi-dimensional inputs. In this case, we use a partial derivative $\frac{\partial}{\partial x_i}h(x)$ to measure how h changes as x_i increases at point x. A critical point is a point where the gradient $\frac{\partial}{\partial x_i}h(x) = 0$ for all possible x_i.

The GD optimizer updates the model parameters using the gradient of the loss function computed over the entire training dataset in each iteration. It converges smoothly. However, the computational cost of gradients in one iteration is expensive, especially for large datasets. To solve this problem, the mini-batch SGD algorithm is proposed to update the model parameters using the gradient computed over a small subset of the training data. This leads to more frequent updates and a faster convergence speed, but the loss function is not as well minimized as in the case of GD. In SGD, the updates of model parameters may not always go in the optimal direction. But in most cases, the learned model parameters are good enough. There are several variants of the SGD optimizer that improves its performance and convergence, including SGD with momentum [19] and weight decay [20].

GD-based optimizers update the model parameters by taking a step in the direction of the negative gradient of the loss function with respect to each parameter in an iterative manner during training. We use a hyperparameter named learning rate to control the size of the step that the optimizer takes when updating the model parameters. For the mini-batch SGD optimizer, the learning rate is fixed and must be chosen by the user. The Adam optimizer [21] is an extension of SGD that calculates an adaptive learning rate when updating each parameter based on estimates of the first and second moments of the gradients. The first and second moments are the mean and variance of the gradients, respectively. The Adam optimizer is computationally efficient and typically requires little tuning. It also has faster convergence and better generalization than SGD. We use the mini-batch SGD optimizer for the first proposed method and use Adam optimizer for the second and third methods.

2.1.4 Convolutional Neural Networks

Convolutional Neural Network (CNN) is a class of artificial neural networks. CNNs are most commonly applied to computer vision tasks, such as image segmentation [22], image classification [23], and object detection [24]. Our proposed methods for BUS image segmentation and classification use CNNs to build its learning architecture. In this section, we briefly introduce several fundamental concepts of CNN related to our works.

Convolutional Layer. A convolutional layer is where a set of filters (or kernels) are applied to input images or feature maps to generate a new feature map. The parameters of the filters are to be learned during training. Convolutional Layers of a CNN extract features from the input and pass the convolved features to the next layers. Figure 2.1 illustrates how a 3 × 3 filter convolves an input of size 5 × 5 to produce convolved features. The nine parameters of the 3 × 3 filter are learned throughout the training.

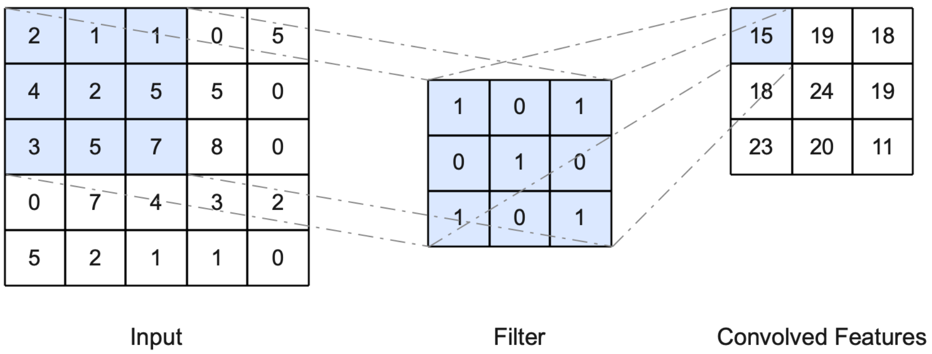


Fig. 2.1: An example of a convolution operation with a 3 × 3 filter and stride of 1. The filter moves across the input and performs a dot product between the nine weights of the filter and the nine pixel values in the input. The result of the dot product is a single value in the output feature map.

Pooling Layer. A pooling Layer always follows a convolutional layer to reduce the spatial dimension of feature maps while retaining the important features. It reduces the number of parameters to learn and the computational cost during the training and alleviates the overfitting problem. In other words, the pooling operation summarises the features in the input feature map. Commonly used pooling operations include max pooling, average pooling, stochastic pooling [25], and spatial pyramid pooling [26]. Figure 2.2 shows two examples of max pooling and average pooling, respectively.

Fully Connected layer. A fully connected layer is also known as a dense layer. It connects all the neurons in one layer to all neurons in the next layer. A fully connected layer

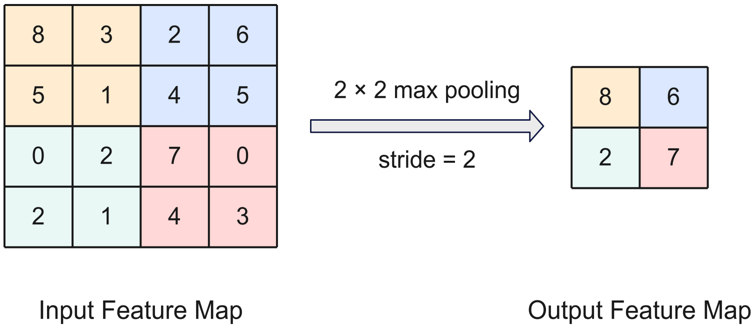


Fig. 2.2: An example of a 2 × 2 max pooling operation with a stride of 2. It selects the maximum value in each 2 × 2 region of the input feature map and outputs the single value for that region. The output feature map is half the size of the input feature map due to the stride size of 2.

always follows a convolutional layer or a pooling layer. Fully connected layers are usually the last few layers in a CNN. They convert high-dimension features to low-dimension features, whose dimension is the same as the number of classes. The value at each dimension gives the probability of an input belonging to the corresponding class.

In a computer vision task, a typical CNN consists of convolutional layers, pooling layers, and fully connected layers. It uses convolutional layers to extract features from the input images, uses pooling layers to reduce the spatial dimension of the extracted features, and uses fully connected layers to flatten the features and perform image classification or regression tasks. Figure 2.3 shows an example of a typical CNN for image classification.

2.2 Attention Mechanisms

In artificial neural networks, the attention mechanism is a technique that mimics cognitive attention in humans. It allows neural networks to focus selectively on certain parts of the input to improve their performance. Attention mechanisms have been commonly used in different tasks, such as natural language processing and computer vision. Spatial attention, channel attention, and self-attention are three representative attention mechanisms. Specifically, spatial attention selectively focuses on different regions of an image or feature map, whereas channel attention selectively focuses on different channels of a feature map. They assign weights to different regions or channels to make the model focus on the

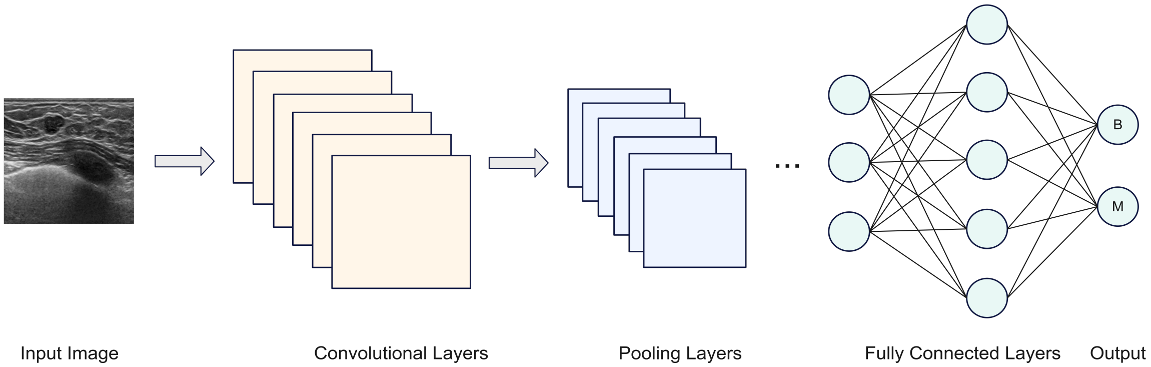


Fig. 2.3: An typical CNN for image classification. It consists of multiple convolutional layers and pooling layers. There may be more convolutional layers and pooling layers before fully connected layers. The output layer has one neuron for each class and outputs the probability of an input image belonging to each class.

most relevant regions or channels to get better results. Self-attention can be used in both spatial attention [27] and channel attention [27, 28] to calculate a set of weights for each position/channel of the input. Self-attention helps a deep learning model focus on the most important regions/channels of the input to improve task performance and can be easily generalized to a wide range of tasks. However, it is computationally expensive and requires powerful equipment to run. Figure 2.4 illustrates a typical spatial attention module and a typical channel attention module.

In BUS image segmentation and classification, self-attention and its variants [29–32] are widely used in neural networks to investigate the importance of features automatically to improve the results. Some other attention mechanisms including the attention-gated unit with soft attention mechanism [33], channel attention [34], spatial-channel attention [35], and global attention upsample [36] have also yielded improved BUS image segmentation or classification performance. These attention mechanisms enhance the useful regions/channels and suppress useless regions/channels in BUS image features to get better feature representation to improve the segmentation and/or classification accuracy.

Two of the three proposed deep learning architectures in this dissertation involve an extension of the spatial self-attention mechanism. Therefore, we provide a brief introduction to spatial self-attention in this section following [27, 28]. The spatial self-attention module

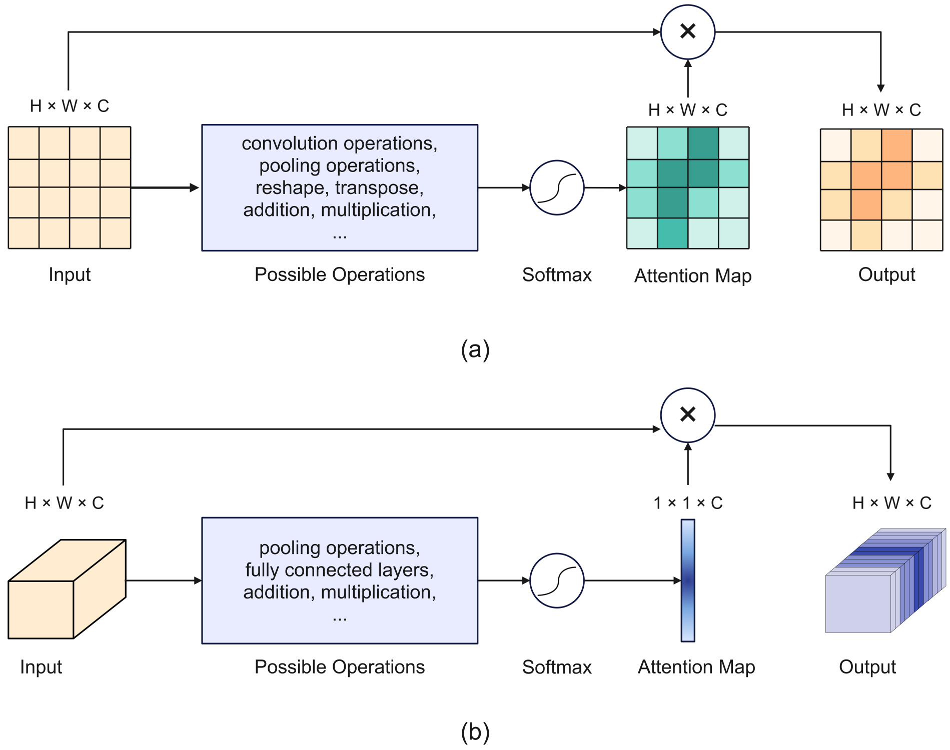

H × W × C
convolution operations, pooling operations, reshape, transpose, addition, multiplication, ...
H × W × C
H × W × C
Input
Possible Operations
Softmax
Attention Map
Output
(a)
H × W × C
pooling operations, fully connected layers, addition, multiplication, ...
1 × 1 × C
H × W × C
Input
Possible Operations
Softmax
Attention Map
Output
(b)

Fig. 2.4: Illustration of (a) a typical spatial attention module and (b) a typical channel attention module. They take a feature of $H \times W \times C$ as the input, where H, W, and C respectively represent the height, width, and channel dimensions, and output a spatial-weighted or channel-weighted feature map. A softmax operation is used to scale the weights in the attention map to the range of (0,1).

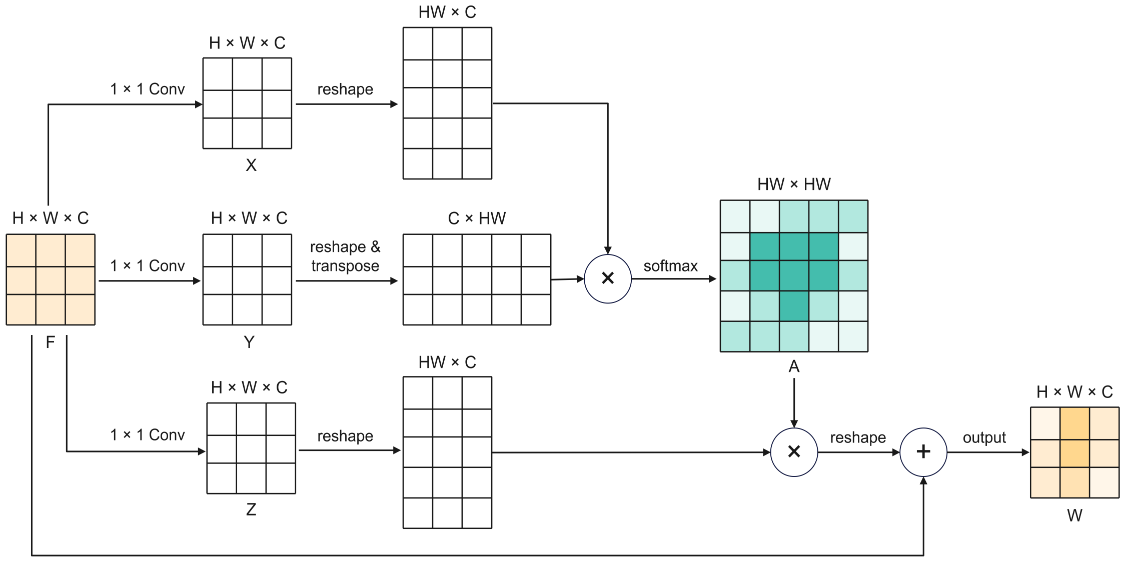


Fig. 2.5: Illustration of the spatial self-attention mechanism.

takes convolutional features as the input and enhances their representation capability by integrating rich contextual information. As shown in Figure 2.5, given a convolutional feature $F \in \mathbb{R}^{H \times W \times C}$, with H, W, and C respectively representing the height, width, and channel dimensions, we first use a 1×1 convolution to transform F into two new feature maps X and Y, respectively, where $\{X, Y\} \in \mathbb{R}^{H \times W \times C}$. Then we reshape X and Y to $\mathbb{R}^{HW \times C}$. A matrix multiplication between the transpose of reshaped X (denoted as X^r) and reshaped Y (denoted as generates a new feature map in $\mathbb{R}^{HW \times HW}$. After that, a softmax layer is performed on this feature map to generate a normalized attention map $A \in \mathbb{R}^{HW \times HW}$:

$$A_{ji} = \frac{exp(X_i^r \cdot Y_j^r)}{\sum_{i=1}^{M} exp(X_i^r \cdot Y_j^r)} \tag{2.6}$$

where A_{ji} measures the i^{th} position's impact on j^{th} position. A large value in A indicates a high correlation.

On a second branch, we use another 1×1 convolution to transform F into a new feature map $Z \in \mathbb{R}^{H \times W \times C}$ and reshape it into $\mathbb{R}^{HW \times C}$. Then we perform a matrix multiplication between A and reshaped Z (denoted as Z^r) to generate a new feature map of size $\mathbb{R}^{HW \times C}$ and then reshape it to $\mathbb{R}^{H \times W \times C}$. Finally, it is multiplied with a learnable parameter μ to gradually assign appropriate weights to A to generate a weighted attention map as in [28], which is further added to the input F to generate a weighted feature map $W \in \mathbb{R}^{H \times W \times C}$:

$$W = \mu \times reshape(\sum_{i=1}^{HW} (A_{ji} Z_i^r)) + A_j \tag{2.7}$$

where each position of W is a weighted sum of the features across all positions and original features. Therefore, the output of the self-attention module integrates global contextual relationships. The similar semantic features achieve mutual gains, which improve intra-class compact and semantic consistency, therefore improving the segmentation accuracy [27].

2.3 BUS Image Segmentation

BUS image segmentation methods can be classified into semi-automated [37–39] and fully automated methods [40, 41] based on human intervention. Fully automated BUS image segmentation is the trend in future BUS CAD systems since it is reproducible and suitable for large-scale tasks [42]. Deep learning-based fully automated methods have recently gained increased popularity because of its improved accuracy and ability to handle large and complex data compared to traditional segmentation methods.

U-Net [43] based methods are particularly popular among all deep learning-based fully automated methods for BUS image segmentation because of their good performance. U-Net uses skip connections to concatenate feature maps in different resolutions and preserve fine details to achieve high segmentation accuracy. Many recent segmentation methods [10, 44–47] are built upon the original U-Net or take advantage of the U-shape encoder-decoder structure. For example, Wang *et al.* [10] propose a fusion deep learning network to address issues of unclear boundaries and large variations in tumors in BUS images. It uses an encoder to capture the context information, a decoder to localize prediction, and a core fusion stream path to combine information from the encoder and the decoder. The fusion stream path takes superpixel images along with the original images as the input and employs four modules to capture various-sized tumor features, coarse-to-fine features, precise boundary features, and consistent features, respectively. These four aggregated feature representations are eventually used for more accurate tumor segmentation. Amiri *et al.* [44] propose a two-stage U-Net architecture that uses the same U-Net architecture for both Region of Interest (ROI) detection and segmentation of BUS images. Specifically, the first U-Net detects where the tumor exists and extracts the ROI of the tumor. The bounding box of each tumor contains the ground truth for the first U-Net. The second U-Net takes the extracted ROI as input and segments the tumor from the ROI. Their results prove that the first stage helps to improve segmentation results in the second stage. Yan *et al.* [47] propose an attention-enhanced U-net with hybrid void convolution to highlight salient features in BUS images to improve the segmentation accuracy. Specifically, they add an attention unit to each skip connection of U-Net to make it focus on learning the tumor area rather than the unnecessary background. They also use hybrid dilated convolution to alleviate the “gridding” effect caused by dilated convolution, and to assign the values in the output feature map using the receptive field region.

2.4 BUS Image Classification

Over recent decades, many methods have been proposed for breast cancer classification. Traditional machine learning classification methods such as Support Vector Machine (SVM) [48, 49], K-Nearest Neighbors (KNN) [50], random forest [51, 52], and Gaussian mixture models [53] have been well studied. For example, Liu *et al.* [48] employ an SVM on three edge-based features (*i.e.*, sum of maximum curvature, sum of maximum curvature and peak, and sum of maximum curvature and standard deviation) extracted from BUS images for breast cancer classification. Ding *et al.* [50] propose a multi-instance learning algorithm to combine local distance and sparseness features and use KNN for classification. Abdel-Nasser *et al.* [52] propose to reconstruct a high-resolution image from a set of input BUS images and then compute ROIs and texture features and finally employ random forests on these features for classification. Huang *et al.* [53] propose to employ a deep neural network to extract features from BUS images, apply principal component analysis to condense extracted features, and use neutrosophic Gaussian mixture models for classification.

Convolutional neural networks (CNNs) have recently achieved superior performance compared to traditional machine learning classification methods. Among them, VGG [54], ResNet [55], and their variants are widely used because the extracted features are efficient for BUS image classification [56]. VGG is popular due to its good performance and its simple and uniform architecture, where all convolutional layers use small 3x3 filters and max-pooling layers to downsample the feature maps. ResNet is designed to alleviate the vanishing gradient problem in very deep neural networks. It introduces the concept of residual connections, where the input is added to the output of a residual block to facilitate the optimization. Many classification methods are based on VGG and ResNet. For example, Liao *et al.* [57] adopt a supervised block-based segmentation algorithm to separate tumor regions from BUS images and then use a VGG-19 to classify segmented tumor regions as benign or malignant. In the tumor region extraction stage, they divide input images into non-overlapping subblocks of the same size for feature extraction. They then use a SVM to classify the tumors in each subblock, and merge adjacent subblocks of the same category into one region. They then use a VGG-19 pre-trained on the Image-Net database to classify segmented tumor regions. Cui *et al.* [58] propose to use ResNet-34 as the backbone feature extractor and design a fused network to combine features of the tumor, peritumoral, and combined-tumoral (combination of tumor and peritumoral) regions to achieve better classification results. Specifically, they propose an enhanced combined-tumoral module to enhance the features of the combined-tumoral region, a region fusion module to extract features of three different regions simultaneously, and a channel attention fusion module to fuse three-region features adaptively. They apply two enhanced combined-tumoral module modules between the last three residual blocks of ResNet-34, apply a region fusion module after the last residual blocks of ResNet-34, and apply a channel attention fusion module before the prediction to improve the overall classification accuracy.

2.5 Multi-task Learning

Multi-task learning (MTL) for simultaneous BUS segmentation and classification has recently been extensively studied in the computer vision community. Benign and malignant breast tumors have different characteristics [59, 60]. For example, benign tumors tend to be smooth, round, and well-circumscribed, whereas malignant tumors are typically rough and spiculated. In addition, malignant tumors tend to have spiculated margins and posterior acoustic shadows. Based on these observations, many MTL studies [29, 30, 33, 61, 62] are proposed to join BUS image segmentation and classification tasks in one network to encourage feature sharing during training to improve both tasks. These MTL methods are mostly based on a U-shape structure (*i.e.*,, an encoder-decoder network for segmentation), and some of them [29, 30, 33] include attention mechanisms to achieve better classification performance. For example, Zhou *et al.* [61] propose an MTL framework for 3D BUS image classification and adopt an iterative feature-refining training strategy to refine features to highlight tumor regions. Their MTL framework consists of a V-Net for segmentation and a lightweight multi-scale network for classification. Low-level features capture shape and boundary information, whereas high-level features summarize attributes of different targets for classification. The authors connect and fuse multi-scale feature maps extracted by different stages of V-Net by using channel-wise global average pooling to improve the classification performance. Their results demonstrate that MTL outperforms single-task segmentation and classification. Chowdary *et al.* [62] propose an MTL framework with a dense branch to combine multi-scale features from different levels of the network for efficient classification of BUS images. The segmentation network is a residual U-Net that replaces each convolutional block of the original U-Net with a residual block. The residual block has a residual connection that adds the input to the output of the original convolutional block, which helps with the propagation of information without degradation. The classification network concatenates features extracted by the last block of the encoder, bridge, and the first block of the decoder for the final classification. Zhang *et al.* [33] propose an MTL framework with soft and hard attention mechanisms to guide the model to pay more attention to tumor regions to boost classification accuracy. The U-shape segmentation network uses DenseNet121 as its backbone. Multi-scale features extracted by the encoder are coalesced by attention-gated units and are flattened to obtain a feature vector for better classification.

2.6 Proposed Methods

In this dissertation, we introduce three deep learning architectures which have been developed during the course of my Ph.D. journey. One of these architectures is for BUS image segmentation, and the other two are for simultaneous BUS image segmentation and classification. Each of the proposed architectures aims to address the drawback of their peers and improve the performance of their peers in terms of segmentation and classification accuracy.

We name the single-task segmentation architecture as Multi-Scale Self-Attention Network (MSSA-Net). The proposed MSSA-Net incorporates an MSSA module in a deep neural network, which uses ResNet-101 as a backbone, to achieve better segmentation results. This MSSA module combines multi-scale features learned by different convolutional blocks to represent the original image at several semantic levels. It integrates both low-level local spatial and high-level semantic contextual information captured in multi-scale features to compute contextual relationships.

We name the two multi-task architectures as Context-Oriented Self-Attention (MTL-COSA) and regional-attentive multi-task learning framework (RMTL-Net). The MTL-COSA incorporates a COSA module in an MTL deep neural network to achieve better segmentation and classification results. It utilizes segmentation outputs to automatically estimate the tumor boundary to learn contextual relationships to improve segmentation and classification accuracy. The RMTL-Net adopts a similar MTL network architecture as MTL-COSA but improves its attention module. Specifically, it employs a more effective regional attention (RA) module to learn corresponding category-sensitive features from three regions (e.g., tumor, peritumoral, and background regions) in BUS images and investigate their influence on BUS image segmentation and classification performance.

CHAPTER 3
MSSA-NET

3.1 Introduction

BUS image segmentation has been well studied. However, many existing BUS image segmentation methods [63–65] simply utilize learned feature maps to segment tumors without considering relationships between pixels. To address this shortcoming, researchers employ self-attention [28] to improve segmentation results by exploring the relationship between pixels and their context. However, it only computes the impact of a pixel on other pixels in one feature map, which is insufficient to represent the contextual relationships. To address the above issues, we propose a novel deep neural network named Multi-Scale Self-Attention Network (MSSA-Net). Our main contributions are:

- Employing ResNet-101 as the backbone to build MSSA-Net to integrate rich spatial and high-level semantic information via multi-scale features.
- Designing a novel MSSA module to explore the rich contextual relationships among pixels in the multi-scale feature maps to boost segmentation performance.

3.2 Architecture Overview of MSSA-Net

The architecture of the proposed MSSA-Net is illustrated in Fig. 3.1. MSSA-Net uses ResNet-101 as its backbone, which consists of five blocks. We use C_i to denote the output of one of the five blocks of ResNet-101, where integer i corresponds to a block number ranging from 1 to 5. It should be noted that C_i contains feature maps of different scales at different depths, where scales decrease and depth increases with increasing i. To integrate both local spatial details and high-level semantics, we employ outputs from five blocks to form a multi-scale feature map F. To maintain local spatial details at the highest resolution,

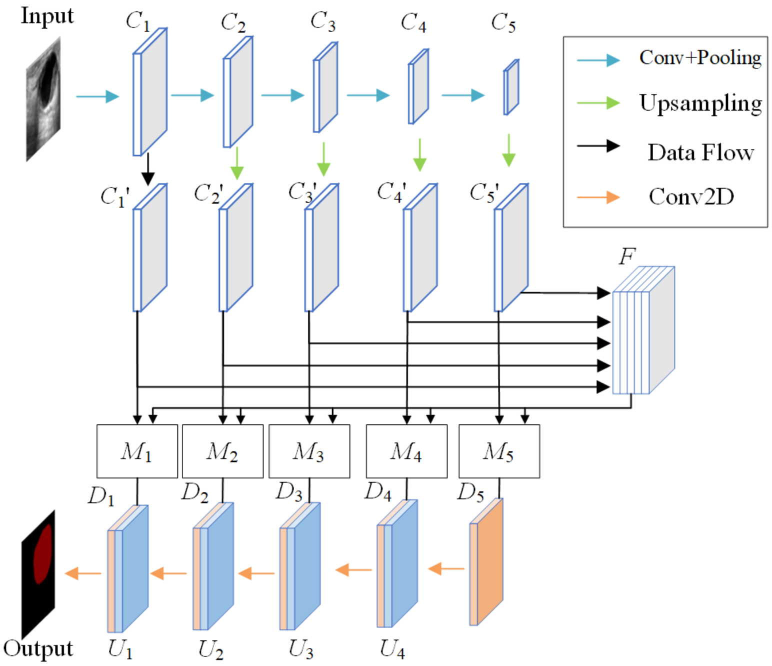


Fig. 3.1: An overview of the proposed MSSA-Net.

we resize each low-resolution output (e.g., C_2, C_3, C_4, and C_5) to a high-resolution output with the same dimension as C_1 by:

$$C_i' = upsample(C_i) \&\& |C_i'| = |C_1| \tag{3.1}$$

where $i = 2$, 3, 4, and 5 and $|x|$ represents the dimension of a feature map x without depth. We then concatenate all resized outputs to construct a multi-scale feature map F by:

$$F = C_1' \oplus C_2' \oplus C_3' \oplus C_4' \oplus C_5' \tag{3.2}$$

where $\oplus$ represents the concatenation operation. Each high resolution C_i' and the multi-scale feature map F are individually fed into the proposed MSSA module, which will be explained in section 3.3, to calculate contextual relationships among pixels and obtain its weighted feature map D_i by:

$$D_i = MSSA(C_i', F) \tag{3.3}$$

Starting with D_5, we convolve it with a 3×3 filter and concatenate the filtered result to integrate spatial and semantic information obtained from blocks 5 and 4. We repeat the same operation to combine spatial and semantic information from blocks 4 and 3, blocks 3 and 2, and blocks 2 and 1. A 3×3 convolution is then applied to U_1, followed by bilinear interpolation and softmax to generate the segmentation result.

3.3 MSSA Module

Self-attention methods [27, 28] have been widely used to compute contextual relationships to better represent features learned by convolutional layers. They take a feature map as the input and output a weighted feature map containing contextual relationships. However, this weighted feature map cannot provide sufficient contextual information. Specifically, a feature map learned by shallow layers contains rich local spatial details while missing high-level semantics. A feature map learned by deep layers contains rich high-level semantic information while missing local spatial details.

To address the aforementioned shortcomings, we propose an MSSA module to integrate both local spatial and high-level semantic contextual information via multi-scale features learned by different convolutional blocks. The MSSA module takes a multi-scale feature map F and a resized local feature map C_i' as inputs and generates a weighted multi-scale feature map D_i that contains contextual relationships among pixels from local spatial and high-level semantic perspectives.

Fig. 3.2 illustrates the proposed MSSA model. For the input of a feature map $C_i' \in \mathbb{R}^{H \times W \times Ch_1}$ with H, W, and Ch_1 respectively representing the height, width, and channel dimensions and i representing the block number, we use a 1×1 convolution to transform C_i' into a new feature map $Y \in \mathbb{R}^{H \times W \times Ch_1/8}$. We use a ratio of 1/8 to reduce the channel number to its 1/8 since this ratio has been empirically determined to be optimal [28]. Similarly, for the input of a multi-scale feature map $F \in \mathbb{R}^{H \times W \times Ch_2}$, we use a 1×1 convolution to generate a new feature map $Z \in \mathbb{R}^{H \times W \times Ch_1/8}$. Since Ch_2 is significantly larger than Ch_1, we reduce the channel number of F to $Ch_1/8$ to conserve time and memory space and enable matrix computations in the next few steps. We then reshape Y to Y_r of

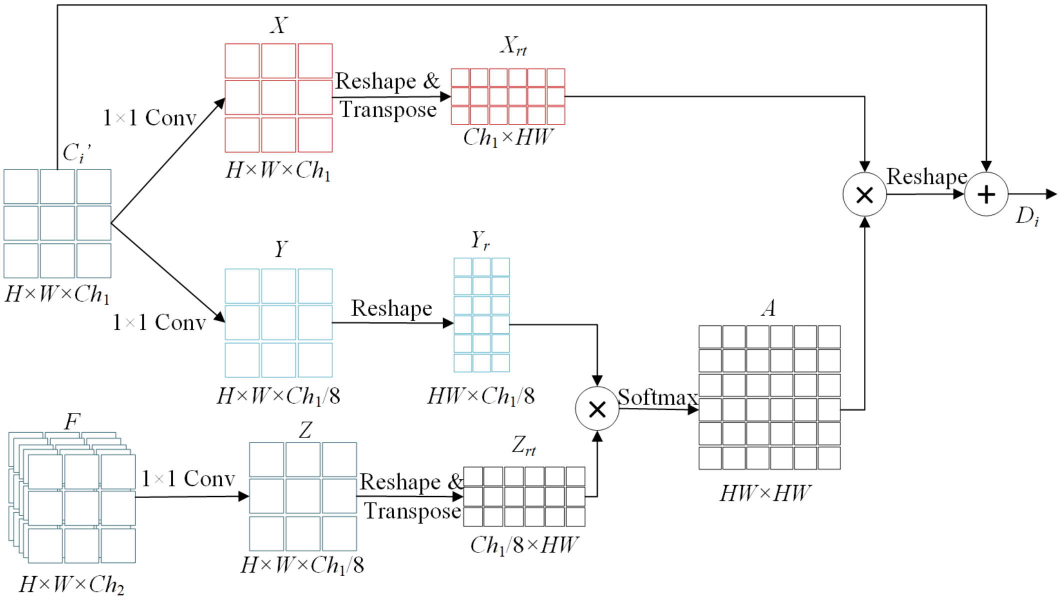


Fig. 3.2: Illustration of the proposed MSSA module.

size $(H \times W) \times Ch_1/8$ and reshape and transpose Z to Z_{rt} of size $Ch_1/8 \times (H \times W)$. A multiplication between Y_r and Z_{rt} generates a map of size $(H \times W) \times (H \times W)$. A softmax is performed on this map to generate a normalized map A, also called the attention map. In other words, the attention map A is computed by:

$$A(m,n) = \frac{exp(Y_r(m,:) \cdot Z_{rt}(:,n))}{\sum_{n=1}^{H \times W} exp(Y_r(m,:) \cdot Z_{rt}(:,n))} \tag{3.4}$$

where : is an operator to get all values in a row or a column and $A(m,n)$ represents the impact of the n^{th} column of Z_{rt} on the m^{th} row of Y_r. A large value in A indicates a high correlation between Y_r and Z_{rt} (i.e., between C_i' and F).

On a second branch, we use another 1×1 convolution to transform C_i' into a new feature map $X \in \mathbb{R}^{H \times W \times Ch_1}$ and reshape and transpose X to X_{rt} of size $Ch_1 \times (H \times W)$. We then perform a matrix multiplication between X_{rt} and A. This result is reshaped to the size $H \times W \times Ch_1$ and multiplied with a learnable parameter μ to gradually assign appropriate weights to A to generate a weighted attention map as in [28], which is further added to the input C_i' to generate a weighted feature map $D_i \in \mathbb{R}^{H \times W \times Ch_1}$.

$$D_i(m,n) = \mu \times reshape((X_{rt}(m,:) \cdot A(:,n))) + C_i'(m,n) \tag{3.5}$$

where $D_i(m,n)$ contains the value of a weighted feature map at location (m,n) and μ is initialized to 0 to allow the network to rely on cues of the local neighborhood to maximize learning.

CHAPTER 4
MTL-COSA

4.1 Introduction

Researchers integrate either attention mechanisms [33, 45] or prior medical knowledge [41] in deep neural networks to achieve better BUS segmentation and classification results. Attention mechanisms make networks focus more on the important parts of BUS images and therefore learn better feature representations and achieve better results. Prior medical knowledge provides helpful information to guide either segmentation or classification. For example, Xu *et al.* [45] design a multi-scale self-attention model to extract rich contextual relationships, which leads to better segmentation results. Zhang [33] *et al.* include soft and hard attention in an MTL network to pay more attention to tumor regions and achieve better classification results. Huang *et al.* [41] incorporate prior medical knowledge to correct any conflicting mistakes. The above three systems achieve better segmentation or classification results. However, none of these systems involves both attention and prior medical knowledge and explores the feasibility of bringing the output of one task to the network to guide the other task. Furthermore, Huang's system is not an end-to-end model.

To address the above issues, we propose a novel end-to-end MTL framework named MTL-COSA (COSA stands for Context-Oriented Self-Attention) to incorporate the COSA module in an MTL deep neural network to achieve better segmentation and classification results. Our major contributions are:

- Proposing an MTL deep neural network for simultaneous breast tumor classification and segmentation.

- Adopting the self-attention model [28] to focus more on each of three regions (background, tumor, and margin) to learn contextual relationships within each region for better feature representations.

- Designing a COSA module that incorporates prior medical knowledge to achieve better segmentation and classification.

4.2 Architecture Overview of MTL-COSA

The architecture of the proposed MTL-COSA is illustrated in Fig. 4.1. It consists of three branches: backbone feature extraction, segmentation, and classification. Segmentation and classification branches use the same feature map extracted by the backbone to produce segmentation and classification results, respectively. The backbone network is ResNet-101 [55], which has five convolutional blocks. We use C_i to denote the output of one of the five blocks, where integer i corresponds to a block number ranging from 1 to 5. Each C_i is at different scales in different depths, where scales decrease and depths increase with increasing i.

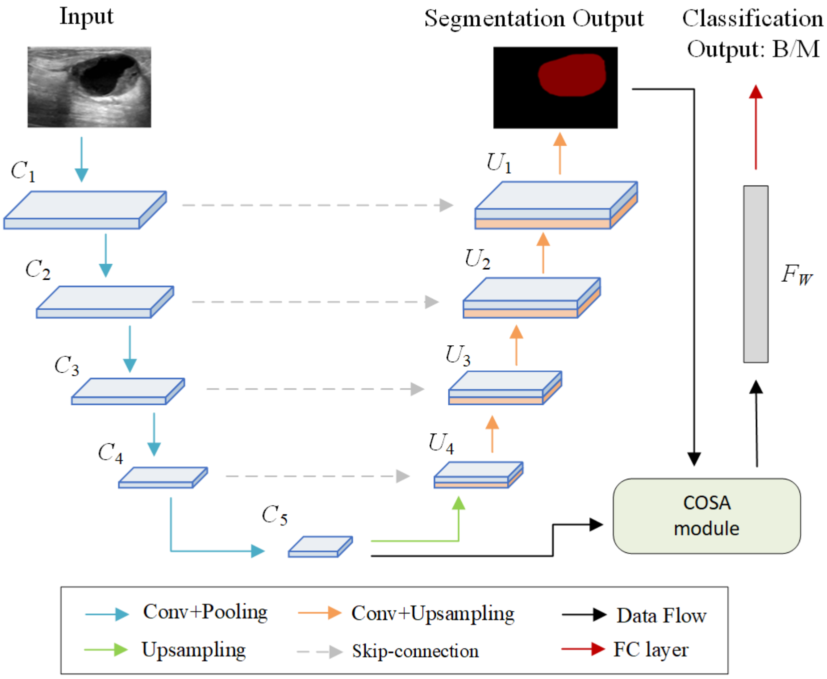


Fig. 4.1: An overview of the proposed MTL-COSA.

The backbone feature extraction branch performs downsampling operations, and the segmentation branch performs upsampling operations. These two branches pair together to form a U-shape structure. We use U_i to denote an upsampled feature map in the segmentation branch, where i ranges from 1 to 4. Starting with C_5, the feature map extracted by the backbone, we concatenate its upsampled feature map with the feature map C_4 to obtain U_4, which integrates local and semantic information from both blocks 5 and 4. We repeat a similar operation to get U_3, U_2, and U_1. A 3×3 convolution is then applied to U_1, followed by bilinear interpolation and softmax layer to generate the final segmentation result.

The feature map C_5 extracted by the backbone feature extraction branch together with the segmentation output is fed into the proposed COSA module to compute rich contextual relationships in BUS images and generate a weighted feature vector F_W, which is passed to a fully connected layer to generate the classification result.

The overall loss of MTL-COSA is the weighted sum of the loss of the segmentation branch $\mathcal{L}_{seg}$ and the loss of the classification branch $\mathcal{L}_{cls}$.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{seg} + \beta \cdot \mathcal{L}_{cls} \tag{4.1}$$

where α and β are weights of losses from segmentation and classification branches, respectively. $\alpha + \beta = 1$. Cross-entropy is employed to compute both $\mathcal{L}_{seg}$ and $\mathcal{L}_{cls}$.

4.3 COSA Module

The self-attention mechanism and prior medical knowledge are commonly used in BUS image segmentation [41, 45]. However, they have not been well studied in the field of BUS image classification and segmentation. To the best of our knowledge, we are the first to incorporate the segmentation results into self-attention [28] to simultaneously segment tumors and classify them as benign or malignant. The segmentation results contain the shape of extracted tumors, which can be used as the estimated prior medical knowledge, to guide the proposed MLT-COSA to learn contextual relationships in BUS to better represent features and therefore achieve better classification and segmentation results.

Fig. 4.2 illustrates the proposed COSA module. It takes the feature map C_5 and

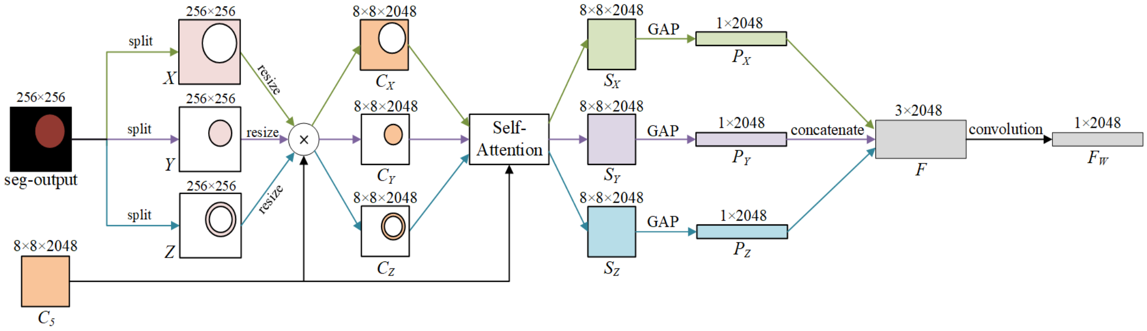


Fig. 4.2: Illustration of the proposed COSA module.

segmentation output as inputs and then outputs a weighted feature vector F_W. Following the findings in [59] that posterior acoustic shadowing in the background region, tumor shape, and tumor margin are three characteristics to differentiate benign and malignant tumors, we split the segmentation output into three regions (background, tumor, and margin) to capture the three vital characteristics. To this end, we apply the Sobel edge detector on the segmentation result to find the tumor contour. We then use a 5×5 square structuring element to perform a dilation operation on the tumor contour to find the inner and outer boundaries. The tumor region falls within the inner boundary. The margin region falls between the inner and the outer boundaries. The background region falls outside the outer boundary. The COSA module employs self-attention to focus on learning contextual relationships to better represent features in each region without the interference of other regions.

To facilitate the description of the COSA, we label the dimension of the data at each step in Fig. 4.2. For the segmentation output of size 256×256, three non-overlapping binary maps X, Y, and Z respectively capture the background, tumor, and margin regions, where the pink region contains values of 1's, and the white region contains values of 0's. The union of three pink regions is a binary map of size 256×256 with all 1's. Region maps X, Y, and Z are then resized to 8×8 and individually multiplied with C_5 to generate three regional feature maps C_X, C_Y, and C_Z, which respectively contain features of background, tumor, and margin regions. C_X, C_Y, and C_Z together with C_5 are individually fed into the self-attention module [28] to compute contextual relationships in the background, tumor, and margin regions, respectively. This self-attention module builds upon the module in [45] to take two inputs and then output an attentive feature map. Attentive feature maps S_X, S_Y, and S_Z produced by the self-attention module are individually fed into a Global Averaging Pooling (GAP) layer to generate their corresponding feature vectors P_X, P_Y, and P_Z, which are concatenated to construct a new feature vector F of size 3×2048. F is resized to $1 \times 2048 \times 3$ and a 1×1 convolution filter is then applied to F to generate F_W of size $1 \times 2048 \times 1$. Lastly, the final weighted feature F_W is resized to 1×2048.

CHAPTER 5
RMTL-NET

5.1 Introduction

Recently, several studies have demonstrated that tumor, peritumoral (the tumor-adjacent area surrounding the tumor), and background regions in BUS images help to improve the diagnosis accuracy of breast cancer in CAD methods [58, 66–68]. Lee *et al.* [67] use the mask R-CNN to extract tumor regions from BUS images and obtain peritumoral regions via a dilation operation. They then use a deep learning model to train tumor, peritumoral, and their combined-tumoral regions to predict axillary lymph node (ALN) metastasis status, which is important in guiding treatment in breast cancer. Sun *et al.* [66] build two models based on tumor, peritumoral, and combined-tumoral regions and compare their performance to show that peritumoral and combined-tumoral regions achieve significantly better performance in predicting ALN metastasis in BUS images for both models.

Tumor, peritumoral, and background regions of a BUS image have been further studied to provide important category-sensitive information to improve the aforementioned methods to achieve better segmentation or classification results. Specifically, the peritumoral region in BUS images was discussed in the BUS image classification task [58] and the ALN metastasis prediction task [66, 67] to further improve their accuracy. Cui *et al.* [58] use an encoder-decoder structure to obtain three tumoral regions at different resolutions to extract tumor features (e.g., component, internal echo, and aspect ratio), peritumoral features (e.g., tumor boundary patterns), and background features (e.g., contextual relationships between the tumor and surrounding tissues). These features lead to higher computational costs but better classification results. Despite the success of the utilization of three tumoral regions, they have hardly been employed in simultaneous BUS image segmentation and classification. To the best of our knowledge, the research work of Xu *et al.* [29] is the pioneer in this direction. They employ three tumoral regions in a BUS image to improve the MTL performance. However, their extracted peritumoral region is small, which may not provide sufficient information for simultaneous BUS image segmentation and classification.

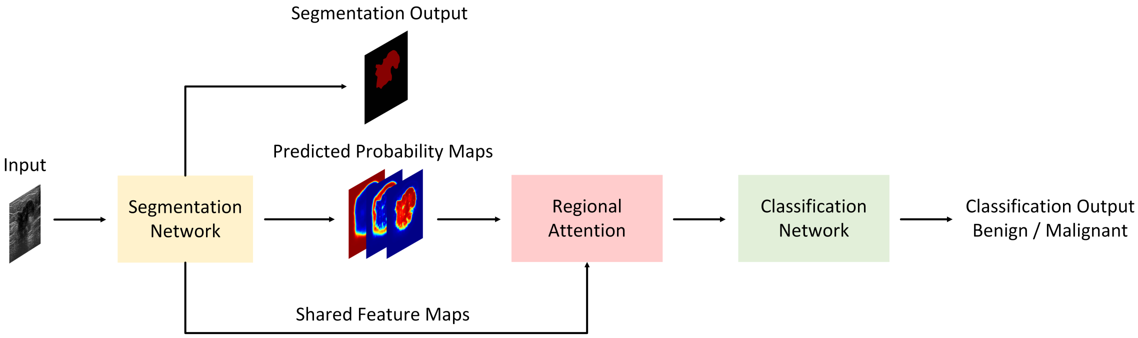


Fig. 5.1: An overview of the proposed RMTL-Net.

To address the shortcomings of the MTL methods, we propose a regional attention (RA) module to learn corresponding category-sensitive features from three regions (e.g., tumor, peritumoral, and background regions) in BUS images and investigate their influence on MTL. We also apply the proposed RA module to a two-stage MTL framework to demonstrate its efficacy in BUS image segmentation and classification. The proposed regional-attentive multi-task learning framework (RMTL-Net) consists of an encoder-decoder network for segmentation and a lightweight network for classification. Both segmentation and classification share features extracted from the encoder. In addition, the RA module utilizes the predicted probability maps to guide the classification network to learn weighted region attentive features for more accurate classification. The overall framework of the proposed RMTL-Net is illustrated in Fig. 5.1. Our main contributions are summarized as follows:

- Designing a novel MTL framework, named RMTL-Net, for simultaneous tumor segmentation and classification in BUS images.
- Proposing a RA module to improve both segmentation and classification performance.
- Employing the predicted probability maps to automatically guide the classifier to learn important category-sensitive information in the tumor, peritumoral, and background

regions.

5.2 Methods

In this section, we first present the proposed pre-processing method to prepare the training images and their pseudo ground truth images. We then describe the proposed method in terms of its network architecture and the regional attention (RA) module.

5.2.1 Pre-processing

In the proposed method, all images are resized to 256 × 256 by bilinear interpolation before being fed into RMTL-Net. Data augmentation techniques are carried out to augment images during the training process using four transformations: (i) rotation of an angle between -5 and 5 degrees at the image center, (ii) random flipping horizontally, vertically, or both, (iii) Gaussian blur, and (iv) Median blur. We perform these four transformations in the above order on each input BUS image to augment the training images during the training procedure.

Given a ground truth BUS image that contains the tumor contour, we generate two pseudo ground truth regions: peritumoral and background regions. First, we employ a Laplace edge detector on the ground truth image to find the contour of the tumor region. Second, we dilate the tumor region by 32 pixels and subtract the tumor region from the dilated result to obtain the peritumoral region. We choose 32 pixels in dilation to ensure the peritumoral region remains at the lowest resolution when a series of down-sampling operations take place in RMTL-Net. Third, we treat the remaining region as the background region. The first three columns in Fig. 5.2 present BUS example images, their ground truth tumor region labeled by radiologists and their pseudo ground truth peritumoral and background regions produced by the proposed pre-processing method, and three regions as shown on the original images. An image containing the ground truth tumor region, the pseudo ground truth peritumoral region, and the pseudo ground truth background region is further used during the training process to learn the boundaries delineating tumor, peritumoral, and background.

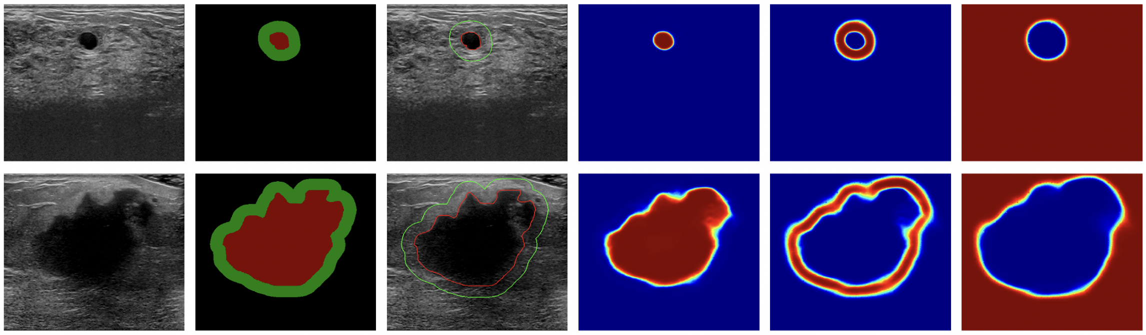


Fig. 5.2: Illustration of two examples of BUS images, their ground truth and pseudo ground truth regions, and three probability maps generated by the proposed RMTL-Net. First column: Original BUS images with a benign tumor shown at the top row and a malignant tumor shown at the bottom row. Second column: Pseudo ground truth regions produced by the proposed pre-processing method, where the peritumoral region is shown in green and the background region is shown in black. The ground truth tumor region is shown in red. Third column: Three regions containing category-sensitive information overlaid on the original image, where the tumor region is within the red line, the peritumoral region is between green and red lines, and the background region is outside the green line. Fourth column: Probability map of the tumor region. Fifth column: Probability map of the peritumoral region. Sixth column: Probability map of the background region.

5.2.2 Architecture Overview of RMTL-Net

The proposed RMTL-Net improves its peer MTL-COSA [29] from the following five aspects:

- Unlike MTL-COSA that generates a binary segmentation result, RMTL-Net generates a binary segmentation result and three probability maps for tumor, peritumoral, and background regions, respectively.

- Unlike MTL-COSA that uses the contour of segmented tumors to find binary segmentation masks for tumor, peritumoral, and background regions, RMTL-Net uses probability maps generated from the network to estimate tumor, peritumoral, and background regions in BUS images and feed them as estimated prior medical knowledge into the RA module to guide the classification task.

- Unlike MTL-COSA that extracts the peritumoral region by dilating the segmented tumor boundary, RMTL-Net is trained to generate respective probability maps for tumor, peritumoral, and background regions to gather more detailed categorical information than the binary masks extracted by MTL-COSA.

- Unlike MTL-COSA whose peritumoral region has a ring area of the width of 5 pixels evenly covering the background and tumor areas, RMTL-Net extracts a bigger peritumoral region with a ring-like area of the width of 32 pixels outside of the tumor to provide sufficient information at the lowest resolution to facilitate classification.
- Unlike MTL-COSA that uses self-attention to learn important classification features, RMTL-Net replaces it with the RA module to significantly reduce network parameters by 14.40% and reduce both training and testing times yet achieve better overall segmentation and classification performance.

The detailed network architecture of the proposed RMTL-Net is illustrated in Fig. 5.3. RMTL-Net is a two-stage framework that consists of a segmentation stage and a classification stage. The segmentation stage utilizes a U-shape architecture consisting of an encoder, a decoder, and skip connections to extract multi-scale features and predict three respective probability maps for tumor, peritumoral, and background regions, as shown in the last three columns in Fig. 5.2. The classification stage uses shared features extracted from the encoder and three probability maps generated from the segmentation stage to produce classification results. Specifically, we use the peritumoral region to capture boundary characteristics, which are useful to differentiate benign and malignant tumors. We use the tumor region to capture the shape properties of tumors, which are useful for both tumor segmentation and classification. We use the background region to capture posterior acoustic shadowing, which is observed more for malignant lesions and less for benign tumors due to attenuation of the sonographic signal [59, 69]. Sharing features makes segmentation and classification promote each other during the training process. In addition, it addresses the problem of having insufficient training images for classification since each pixel is treated as a labeled training data for segmentation. Sharing features with the segmentation stage with sufficient training samples improves the overall accuracy and robustness of the classification stage.

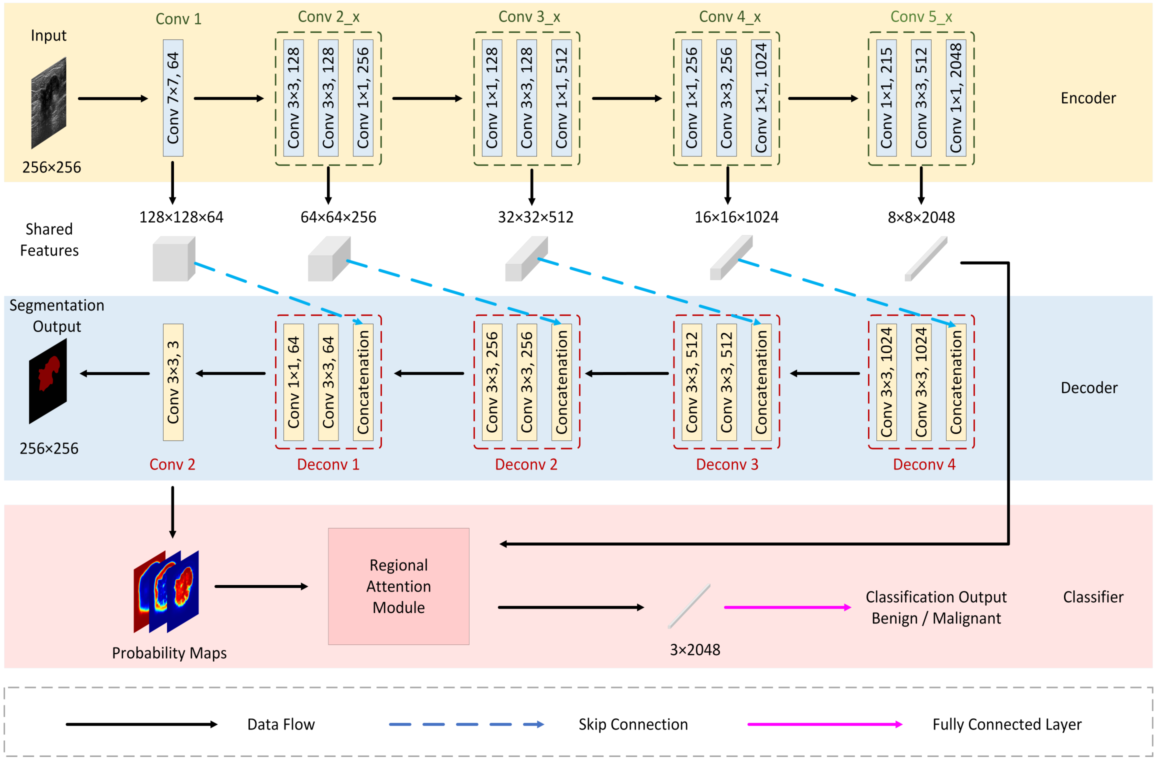


Fig. 5.3: A detailed illustration of the proposed RMTL-Net.

We use ResNet-101 [55] as the backbone of the segmentation stage of RMTL-Net due to its great performance in BUS image segmentation and classification [45, 58]. The architecture of ResNet-101 remains the same. Specifically, the encoder utilizes one convolutional layer $Conv1$ together with four residual blocks ($Conv2\_x$ to $Conv5\_x$) to perform five down-sampling operations to extract multi-scale features from input images. Multi-scale features extracted by $Conv1$ to $Conv5\_x$ are of sizes $128 \times 128 \times 64$, $64 \times 64 \times 256$, $32 \times 32 \times 512$, $16 \times 16 \times 1024$, and $8 \times 8 \times 2048$, respectively. The decoder symmetrically utilizes four deconvolutional blocks ($Deconv4$ to $Deconv1$) and one convolutional layer ($Conv2$) followed by bilinear interpolation and softmax operations to perform up-sampling operations. Skip connections between the encoder and decoder combine feature maps in different scales to compensate for the loss of spatial information during down-sampling operations and to refine segmentation outcomes. As a result, multi-scale features are restored to the original input size and are further interpreted to predict three probability maps.

We use three probability maps generated from the segmentation stage of RMTL-Net and multi-scale high-level features shared by both segmentation and classification stages to produce classification results.

5.2.3 Regional Attention Module

Unlike classical image classification networks (e.g., VGG [54] and ResNet [55]), we add a regional attention (RA) model to further encourage information sharing. This RA model outputs a weighted feature vector of size 1×2048 that is passed to a fully connected layer to generate more accurate classification results.

We observe benign and malignant tumors exhibit different characteristics. For example, benign tumors tend to be smooth and round, and malignant tumors are always rough with an aspect ratio of greater than 1 [59,60]. Benign tumors tend to have smooth, thin, and regular margins, and malignant tumors tend to have spiculated, thick, and irregular margins. Benign tumors tend to have less posterior acoustic shadowing in the background region than malignant lesions. As a result, we propose to utilize tumor, peritumoral, and background regions to learn their inherently important characteristics, including tumor features (e.g., component, internal echo, and aspect ratio), tumor boundary patterns (e.g., smoothness, shape, and contextual texture between tumor and surrounding tissues), and background features (posterior acoustic shadowing) [59, 68] to help with the joint segmentation and classification tasks. In addition, we propose to include a RA module in the classification stage of the RMTL-Net to encourage information sharing and output a weighted feature vector to facilitate classification. This RA module combines multi-scale high-level features with three probability maps generated from the segmentation stage to guide the learning of category-sensitive features from three regions, namely, tumor, peritumoral, and background regions. Category-sensitive features are represented as a weighted feature vector, which is passed to a fully connected layer to generate more accurate classification results. Fig. 1.3 shows six examples of BUS images that contain benign and malignant tumors, respectively. Tumor regions with high variability in shape, size, and location are delineated by red lines. When using these images as training images, we generate their pseudo ground truth peritumoral and background regions using the pre-processing method explained in Section 5.2.1. When using these images as testing images, RMTL-Net predicts their probability maps, as shown in Fig. 5.2.

The structure diagram of the proposed RA module is shown in Fig. 5.4. The algorithmic view of the RA module is summarized below:

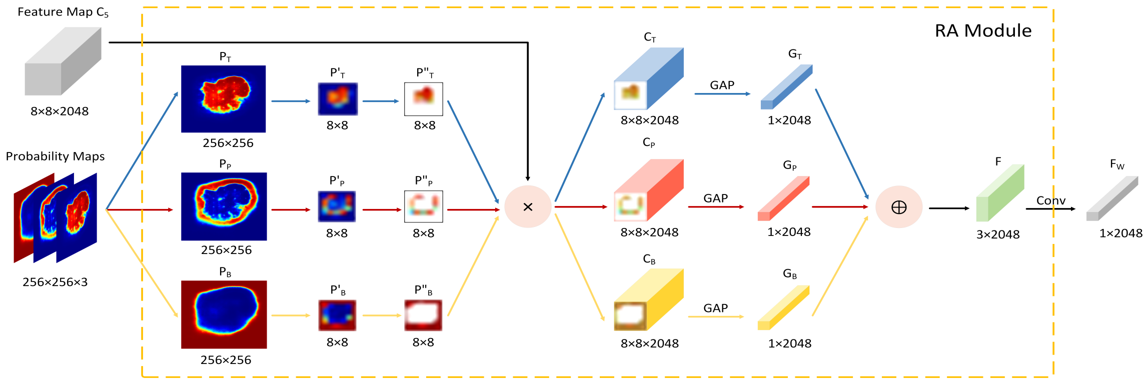


Fig. 5.4: An overview of the proposed Regional Attention (RA) module.

Input: C_5 (the feature map of size $8 \times 8 \times 2048$ extracted by $Conv5\_x$ of the encoder) and P (the probability map of size $256 \times 256 \times 3$ generated by the last convolutional layer $Conv2$ of the decoder).

Output: A weighted feature map F_W of size 1×2048.

1. Split P into three probability maps P_T, P_P, and P_B of size 256×256, where subscripts T, P, and B represent tumor, peritumoral, and background, respectively.

2. Employ the nearest neighbor method to resize P_T, P_P, and P_B to obtain coarse probability maps P'_T, P'_P, and P'_B of size 8×8.

3. Utilize a threshold of 0.5 to filter coarse probability maps P'_T, P'_P, and P'_B to obtain three noise-free probability maps P''_T, P''_P, and P''_B, respectively. Specifically, values greater than 0.5 in coarse probability maps are kept intact, and values smaller than

or equal to 0.5 are set to 0:

$$P''_x = P'_x > 0.5\,?\,P'_x : 0 \tag{5.1}$$

where subscript x can be replaced with T, P, or B.

4. Individually and elementwisely multiply P''_x with each channel of C_5 to generate multi-channel weighted regional feature maps C_x.

$$C_x = C_5 \cdot P''_x \tag{5.2}$$

5. Apply the global average pooling (GAP) on C_x to capture weights of each region in its corresponding G_x of size 1 × 2048:

$$G_x = GAP(C_x) \tag{5.3}$$

6. Concatenate G_T, G_P, and G_B to construct a new feature vector F of size 3 × 2048:

$$F = Concatenate(G_T, G_P, G_B) \tag{5.4}$$

7. Apply a 1 × 1 convolution filter to F to generate a weighted feature map F_W of size 1 × 2048.

$$F_W = f^{1\times1}(F) \tag{5.5}$$

It should be noted that all non-zero pixels in P''_T, P''_P, and P''_B have high likelihood values larger than 0.5, which indicate high strength of tumor, peritumoral, and background features, respectively. We choose 0.5 as the threshold because it classifies a pixel into one of the three classes. The multiplication of C_5 and P''_T, P''_P, and P''_B leads to multi-channel weighted tumor, peritumoral, and background features C_T, C_P, and C_B. The GAP operation further finds the features in each channel of C_T, C_P, and C_B to best represent three respective regions. The concatenation operation followed by the 1 × 1 convolution constructs a weighted sum of multi-view features from three parallel channels (*i.e.*, G_T, G_P, and G_B), which can be formulated as:

$$F_W = w_1 \cdot G_T + w_2 \cdot G_P + w_3 \cdot G_B \tag{5.6}$$

where w_1, w_2, and w_3 indicate the importance of tumor, peritumoral, and background regions, respectively. These weights are automatically learned during the training process. Finally, F_W is passed to a fully connected layer followed by a softmax activation function for automated tumor classification. F_W captures the importance of each region for better feature representation and therefore leads to better classification results than using a non-weighted feature map (*i.e.*, convolving C_5 with a feature vector of 1×2048). In summary, the proposed RA module follows the perspectives of radiologists to learn multi-view features from three regions in BUS images to achieve better segmentation and classification performance. Specifically, the tumor region helps to extract the basic features of breast tumors. The peritumoral region helps to capture tumor boundary patterns. The background region helps to collect contextual information.

5.2.4 Loss Function

The overall loss of RMTL-Net is computed by the weighted sum of the loss of the segmentation task $\mathcal{L}_{seg}$ and the loss of the classification task $\mathcal{L}_{cls}$.

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{seg} + (1 - \lambda) \cdot \mathcal{L}_{cls} \tag{5.7}$$

where λ and $1 - \lambda$ are contribution weights of losses from segmentation and classification tasks, respectively. Cross entropy is employed to compute both $\mathcal{L}_{seg}$ and $\mathcal{L}_{cls}$.

Let K denote the number of classes in a given task, N denote the number of images, and P denote the number of pixels in an image. In the segmentation task, there are 3 classes representing tumor, peritumoral, and background regions. In other words, $K = 3$.

The pixel-wise cross entropy $\mathcal{L}_{seg}$ of the segmentation task is computed as follows:

$$\mathcal{L}_{seg} = -\frac{1}{P}\sum_{p}^{P}\sum_{k}^{K} y_{p,k} \cdot \log \hat{y}_{p,k} \tag{5.8}$$

where $y_{p,k}$ and $\hat{y}_{p,k}$ represent the true and predicted probability of pixel p belonging to class k, respectively. The true probability $y_{p,k}$ is either 0 or 1 since each pixel belongs to one of the three classes. The predicted probability $\hat{y}_{p,k}$ is in the range of $[0, 1]$.

In the classification task, there are 2 classes representing benign and malignant tumors. In other words, $K = 2$. The image-wise cross-entropy $\mathcal{L}_{cls}$ of the classification task is computed as follows:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{n}^{N}\sum_{k}^{K} y_{n,k} \cdot \log \hat{y}_{n,k} \tag{5.9}$$

where $y_{n,k}$ and $\hat{y}_{n,k}$ represent the true and predicted category of image n belonging to class k, respectively. Both $y_{n,k}$ and $\hat{y}_{n,k}$ are either 0 or 1.

CHAPTER 6
EXPERIMENTAL RESULTS

In this section, we first introduce two major public datasets in BUS image segmentation and classification in section 6.1. We then introduce evaluation metrics of BUS segmentation and classification in section 6.2. We next provide the experimental setup and experimental results of the proposed MSSA-Net, MTL-COSA, and RMTL-Net in sections 6.3, 6.4, and 6.5, respectively. Each of them is compared with various state-of-the-art BUS segmentation and classification methods.

6.1 Datasets

Dataset UDIAT [11]: This dataset was collected by the UDIAT Diagnostic Centre of the Parc Taul´ı Corporation, Sabadell (Spain), using a Siemens ACUSON Sequoia C512 system 17L5 HD linear array transducer (8.5 MHz). It contains 163 BUS images with an average size of 760 × 570 pixels, where 110 images have benign tumors and 53 images have malignant tumors. These BUS images are obtained from different female patients, and each BUS image presents one tumor. Ground truth is generated by experienced radiologists.

Dataset BUSI [12]: This dataset was collected by Baheya Centre for Early Detection and Treatment of Women's Cancer, Egypt using LOGIQ E9 ultrasound and LOGIQ E9 Agile ultrasound system. It contains 780 BUS images with an average size of 500 × 500 pixels, where 437 images have benign tumors, 210 images have malignant tumors, and 133 images do not have any tumors. These BUS images are obtained from 600 female patients between the ages of 25 and 75 years old. We use 647 images with benign or malignant tumors in this dataset for binary classification in this study. Ground truth is generated by radiologists from Baheya.

Because the size of dataset UDIAT is small, there is 3% classification performance differences between multiple runs even if we use five-fold cross-validation to train and test on it. To increase the credibility of experimental results, we train all competing methods on two datasets together and test on two datasets separately. Specifically, for each dataset, we split the data into five groups, where each group keeps the same proportion of benign and malignant cases as in the original dataset. In each fold experiment, four groups of each dataset are combined and used as the training set, and the other group is used as the testing set. In this study, all experimental results are reported by averaging the five-fold cross-validation performance.

6.2 Evaluation Metrics

In this section, we introduce evaluation metrics of BUS segmentation and classification that we used throughout this dissertation. The three proposed methods employ different metrics depending on their tasks.

We employ commonly-used BUS segmentation metrics [10, 33, 42, 44, 45, 62] including sensitivity (SEN), specificity (SPE), accuracy (ACC), dice similarity coefficient (DSC), and intersection over the union of tumor (tumor IoU) to quantitatively evaluate the segmentation performance. Higher values of these metrics represent better segmentation performance. Specifically, SEN and SPE measure the ability of a model to correctly identify all tumor pixels and background pixels in BUS images, respectively; ACC reports the percent of correctly segmented tumor pixels in BUS images; both DSC and tumor IoU are positively correlated and measure the spatial overlap between the predicted segmentation result and ground truth. However, DSC tends to measure the average-case performance and tumor IoU tends to measure the worst-case performance. $FPR^{\star}$ and AER are proposed in [70] specifically for measuring the ratio of wrongly classified pixels, and the error rate of BUS segmentation. These metrics are calculated as follows:

$$SEN = \frac{TP}{TP + FN} \tag{6.1}$$

$$SPE = \frac{TN}{TN + FP} \tag{6.2}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.3}$$

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{6.4}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{6.5}$$

$$FPR^{\star} = \frac{FP}{TP + FN} \tag{6.6}$$

$$AER = \frac{FN + FP}{TP + FN} \tag{6.7}$$

where TP represents true positives (*i.e.*, the number of true tumor pixels that are correctly predicted to be tumor pixels), FP represents false positives (*i.e.*, the number of true background pixels that are wrongly predicted to be tumor pixels), FN represents false negatives (*i.e.*, the number of true tumor pixels that are wrongly predicted to be background pixels), and TN represents true negatives (*i.e.*, the number of true background pixels that are correctly predicted to be background pixels). Since only two kinds of pixels (tumor and background) are involved in evaluating the segmentation performance, we consider all the pixels in the predicted background and peritumoral regions as background pixels and all the pixels in the predicted tumor region as tumor pixels.

We employ commonly-used BUS classification metrics [33,48,53,58,62] including SEN, SPE, ACC, precision (PRE), F1-score (F1), and area under receiver operating characteristic curve (AUC) to quantitatively evaluate the classification performance. Higher values of these metrics represent better classification performance. Specifically, SEN, SPE, and ACC are computed in the same manner as the segmentation metrics of the same names. However, TP, TN, FP, and FN are defined differently when evaluating classification. TP and TN respectively represent the number of BUS images that are correctly predicted as benign images (*i.e.*, a positive class) and malignant images (*i.e.*, a negative class). FP and FN respectively represent the number of BUS images that are incorrectly predicted as benign and malignant images. F1-score is the same as DSC. AUC is a summary of the receiver operating characteristic (ROC) curve, which shows the performance of a model at all classification thresholds. A higher AUC value represents better classification performance. PRE computes the ratio of correctly predicted positive samples to the total predicted positive samples. It is computed as follows:

$$PRE = \frac{TP}{TP + FP} \tag{6.8}$$

6.3 MSSA-Net Results

In this section, we evaluate the performance of the proposed MSSA-Net on the above two public BUS datasets. All experiments are conducted on Ubuntu 18.04 system, Intel(R) Xeon(R) CPU E5-2620 2.00 GHz, and two NVIDA GeForce 1080 graphics cards. Input images and ground truths are resized to 128×128. The Stochastic Gradient Descent (SGD) optimizer utilizes a learning rate of 0.001, a momentum of 0.99, a batch size of 12, and epochs of 100. Cross-entropy is employed in the loss function. To ensure a fair comparison, we set these parameters to be the same for all compared methods. We also employ 10-fold cross-validation to evaluate the performance of all compared methods on two datasets.

We compare the performance of MSSA-Net with six state-of-the-art deep neural network-based segmentation methods on two aforementioned datasets. The six compared methods are U-Net [43] with ResNet-101 as a backbone [55] (denoted as U-ResNet), U-ResNet with self-attention [28] applied on five blocks (denoted as U-ResNet SA), ResNet-101 [55] by resizing the output of the 5^{th} block to the input size, FCN8s [71], PSPNet [72], and Deeplabv3+ [73]. We use five metrics to evaluate segmentation results, including SEN, FPR, IoU, DSC, and AER.

Table 6.1 summarizes segmentation results of MSSA-Net and six peer methods in terms of five measures on two datasets. MSSA-Net has the highest SEN, JI, and DSC values and the lowest $FPR^{\star}$ and AER values on Dataset BUSI and therefore achieves the best performance. Specifically, it improves the second-best method by 2.35%, 1.42%, 1.32%, 2.82%, and 5.67% for SEN, JI, DSC, FPR, and AER, respectively. MSSA-Net achieves the best performance in terms of JI, DSC, FPR, and AER and a comparable SEN on dataset BUSI. MSSA-Net also achieves the smallest standard deviation for all five metrics (e.g.,

Table 6.1: Summary of tumor segmentation results of MSSA-Net and its peer methods (%)

Datasets	Methods	SEN	$FPR^{\star}$	IoU	DSC	AER
Dataset UDIAT	U-ResNet	**85.67**	24.12	74.70	82.83	38.45
	U-ResNet SA	84.32	24.98	74.87	82.85	40.67
	ResNet-101	46.52	26.70	37.35	46.77	80.18
	FCN8s	77.95	32.98	62.27	72.90	55.03
	PSPNet	81.06	23.77	70.65	79.73	42.71
	Deeplabv3+	63.44	76.20	50.57	60.78	112.77
	proposed	85.63	**19.48**	**76.05**	**83.78**	**33.85**
Dataset BUSI	U-ResNet	79.20	34.82	70.59	79.03	55.63
	U-ResNet SA	79.02	29.80	70.89	79.60	50.78
	ResNet-101	53.30	36.94	44.37	54.20	83.64
	FCN8s	78.19	44.94	64.00	74.28	66.75
	PSPNet	78.76	33.96	69.79	78.56	55.20
	Deeplabv3+	57.34	44.51	48.12	57.98	87.18
	proposed	**81.06**	**28.96**	**71.90**	**80.65**	**47.90**

0.21 for SEN, 1.30 for FPR, 0.21 for JI, 0.21 for DSC, and 1.36 for AER). MSSA-Net and U-ResNet SA, respectively, have $71,534,626$ and $98,374,562$ trainable parameters. On average, it takes 0.031 seconds for MSSA-Net and 0.035 seconds for U-ResNet SA to segment an image.

Fig. 6.1 compares the performance of MSSA-Net and its five variants on two datasets in terms of five aforementioned metrics. MSSA-Net involves combined attention layers U_5 through U_1, while its variants involve some selected attention layers or no attention. Five variants of MSSA-Net are as follows: V1 for variant 1 without involving attention layers; V2 for variant 2 involving one attention layer U_1; V3 for variant 3 involving combined attention layers U_2 and U_1; V4 for variant 4 involving combined attention layers U_3 through U_1; V5 for variant 5 involving combined attention layers U_4 through U_1; V6 for the proposed MSSA-Net. We compute the average values of each metric for two datasets to compare segmentation performance. Specifically, we present TPR, JI, and DSC results in the left plot of Fig. 6.1 since larger values indicate better segmentation results and present FPR and AER results in the right plot of Fig. 6.1 since smaller values indicate better segmentation results. It clearly shows that MSSA-Net yields the largest TPR, JI, and DSC values and

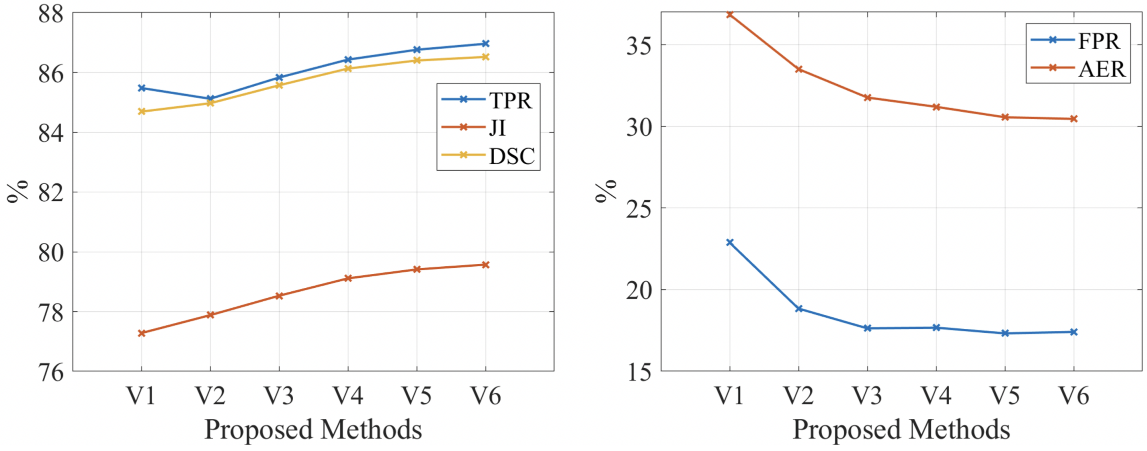


Fig. 6.1: Comparison of MSSA-Net and its variants in terms of five metrics: TPR, JI, and DSC (left); FPR and AER (right).

the smallest FPR and AER values. Variant 1 yields the smallest JI and DSC values and the largest FPR and AER values. With the exception of the TPR metric, JI and DSC values gradually increase, and FPR and AER values gradually decrease as more attention layers are employed. In other words, segmentation results gradually improve as more attention layers are employed.

Fig. 6.2 presents segmentation results of MSSA-Net and six compared methods for one representative BUS image in Dataset UDIAT (top row) and Dataset BUSI (bottom row). For the BUS image in Dataset UDIAT containing a small tumor and a large tumor-like region with a clear contour, Deeplabv3+, PSPNet, ResNet-101, and U-ResNet mistakenly segment the tumor-like region and ResNet-101 mistakenly segments the tumor region. FCN8s and U-ResNet SA segment a single tumor with a JI value of 63.28% and 72.46%, respectively. MSSA-Net gives a more accurate segmentation result with the highest JI value of 82.17%. For the BUS image in Dataset BUSI containing one irregular tumor without a clear contour, MSSA-Net achieves the highest JI and DSC values of 74.43% and 84.68%, and the lowest AER value of 28.79%. The other six methods fail to segment the tumor since their JI values are less than 55%, DSC values are less than 70%, and AER values are larger than 65%.

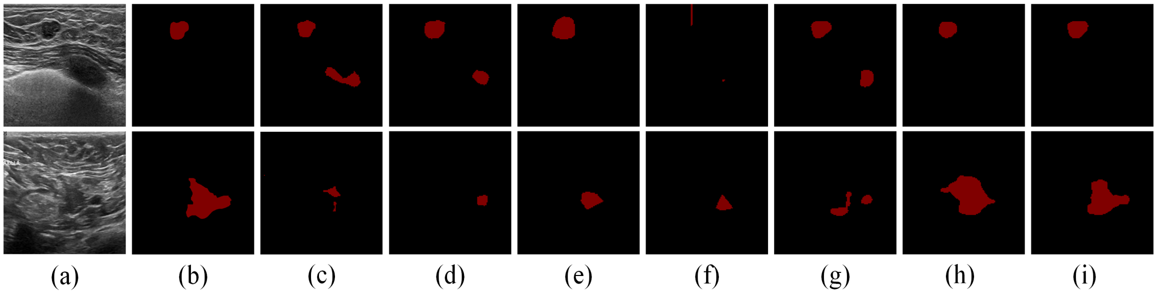


Fig. 6.2: Illustration of segmentation results. (a) BUS images; (b) Ground truth; Segmentation results obtained by (c) Deeplabv3+; (d) PSPNet; (e) FCN8s; (f) ResNet-101; (g) U-ResNet; (h) U-ResNet SA; (i) MSSA-Net.

6.4 MTL-COSA Results

In this section, we evaluate the performance of the proposed MTL-COSA on Dataset UDIAT and Dataset BUSI. All experiments are conducted on Ubuntu 18.04 system, Intel(R) Core(TM) CPU i5-11600K 3.9 GHz, and 2 NVIDIA GeForce 1080Ti graphics cards. To train all networks, Adam optimizer is used with learning rate of 0.0001, momentum β_1 of 0.9, momentum β_2 of 0.99, and weight decay of 0.0005. The batch size is 12 and the number of training epochs is 100. All BUS images are resized to 256 × 256 as the input. Weights of two branches α and β are empirically set to be 0.8 and 0.2, respectively. Other values significantly reduce the mIoU value of segmentation results.

We compare the proposed MTL-COSA with several state-of-the-art methods in terms of segmentation and classification accuracy. The compared segmentation methods are U-shape ResNet-101 [55] (UResNet), Multi-Task Learning (MTL) with a classification branch added to UResNet, MTL-SA with conventional self-attention [28] added to MTL, and the proposed MTL-COSA with COSA added to MTL. MTL feeds C_5 into a fully connected layer for classification while passing C_5 into U_4 to U_1 for segmentation. MTL-SA applies self-attention [28] to C_5, and the attentive C_5 is used to generate classification and segmentation results in the same way as MTL. The compared classification methods are VGG-16 [54], LeNet [74], ResNet-101, MTL, MTL-SA, and MLT-COSA.

Table 6.2 summarizes the segmentation results of all compared methods on each dataset in terms of segmentation metrics SEN, $FPR^{\star}$, mIoU, DSC, and AER. Note that mIoU represents the mean value of IoU of the tumor and background regions in this work. The Single Task (ST) segmentation network, UResNet, achieves better SEN, mIoU, and DSC values than all MTL methods on both datasets. It is reasonable because adding a classification branch to UResNet leads to a weight reduction of the segmentation branch in the loss function. This reduction is determined by α in Eq. (4.1), with smaller α leading to more weight reduction in segmentation and therefore leading to worse segmentation results. Among three MTL methods, MTL-COSA achieves the highest mIoU of 85.87%, the highest DSC of 81.82%, and the lowest AER of 34.97% on Dataset UDIAT. Among three MTL methods, MTL-COSA outperforms the other two methods in all five metrics on Dataset BUSI. Overall, MTL-COSA achieves the best segmentation performance on both datasets compared to other MTL methods. The segmentation performance is dropped for all MTL methods when comparing with the ST method since less weight is employed in training to reduce segmentation error. However, MTL-COSA maintains the smallest drop in segmentation performance due to its integration of both attention mechanisms and prior medical knowledge.

Table 6.2: Summary of tumor segmentation results of MTL-COSA and its peer methods (%)

Datasets	Methods		SEN	$FPR^{\star}$	mIoU	DSC	AER
Dataset UDIAT	ST	UResNet	84.39	34.03	86.11	82.25	49.64
	MT	MTL	**82.18**	23.03	85.00	79.92	40.85
		MTL-SA	78.54	**15.50**	84.44	78.90	36.96
		MTL-COSA	82.17	17.14	**85.87**	**81.82**	**34.97**
Dataset BUSI	ST	UResNet	77.09	33.34	82.27	77.63	56.25
	MT	MTL	77.52	39.11	81.48	76.38	61.59
		MTL-SA	75.28	37.05	81.09	75.61	61.77
		MTL-COSA	**77.85**	**37.01**	**82.13**	**77.48**	**59.15**

Table 6.3 summarizes classification results of all compared methods on each dataset. Among the ST classification methods, ResNet achieves the best overall performance. Among the MTL classification methods, MTL-COSA achieves the best performance in terms of

FPR, ACC, PRE, and F_1 scores. All MTL classification methods achieve better performance than all ST classification methods in terms of ACC, PRE, and F_1 scores. At least one MTL classification method achieves better performance than all ST classification methods in terms of SEN and $FPR^\star$ scores. It is clear that classification results are significantly improved for MTL methods. Adding a segmentation branch, which has enough training samples, makes the network learn better feature representations and therefore achieve better classification results. This improvement surpasses the performance drop caused by the weight reduction of the segmentation branch in the loss function.

Table 6.3: Summary of tumor classification results of MTL-COSA and its peer methods (%)

Datasets	Methods		SEN	SPE	ACC	PRE	F_1
Dataset UDIAT	ST	VGG-16	85.33	66.73	79.16	84.10	84.60
		LeNet	**92.73**	48.18	77.99	78.26	84.74
		ResNet	90.91	**70.18**	**84.09**	**86.23**	**88.38**
	MT	MTL	**92.73**	68.36	84.69	86.09	88.88
		MTL-SA	90.82	74.36	85.34	88.24	89.22
		MTL-COSA	91.73	**77.64**	**87.08**	**89.36**	**90.41**
Dataset BUSI	ST	VGG-16	89.22	79.05	85.92	89.83	89.51
		LeNet	90.84	56.67	79.74	81.44	85.81
		ResNet	**93.56**	**82.86**	**90.10**	**91.90**	**92.68**
	MT	MTL	93.57	85.24	90.86	92.98	93.22
		MTL-SA	**94.27**	83.81	90.87	92.39	93.30
		MTL-COSA	92.20	**90.00**	**91.48**	**95.05**	**93.59**

Fig. 6.3 shows ROC curves with values of Area Under the ROC Curve (AUC) for each method listed in Table 6.3. Among three ST classification methods, ResNet yields the highest AUC of 0.89 on Dataset UDIAT and 0.96 on Dataset BUSI. Among three MTL methods, the proposed MTL-COSA achieves the highest AUC of 0.93 on Dataset UDIAT and 0.97 on Dataset BUSI. The MTL method without attention and prior medical knowledge achieves the worst classification results on both datasets, which are comparable with the classification results obtained by the best ST method. It is clear from Table 6.3 and Figure 6.3 that the COSA module guides the MLT network to utilize the estimated prior

medical knowledge in the attention mechanism to learn better feature representations and achieve better classification results and comparable segmentation results than ST methods.

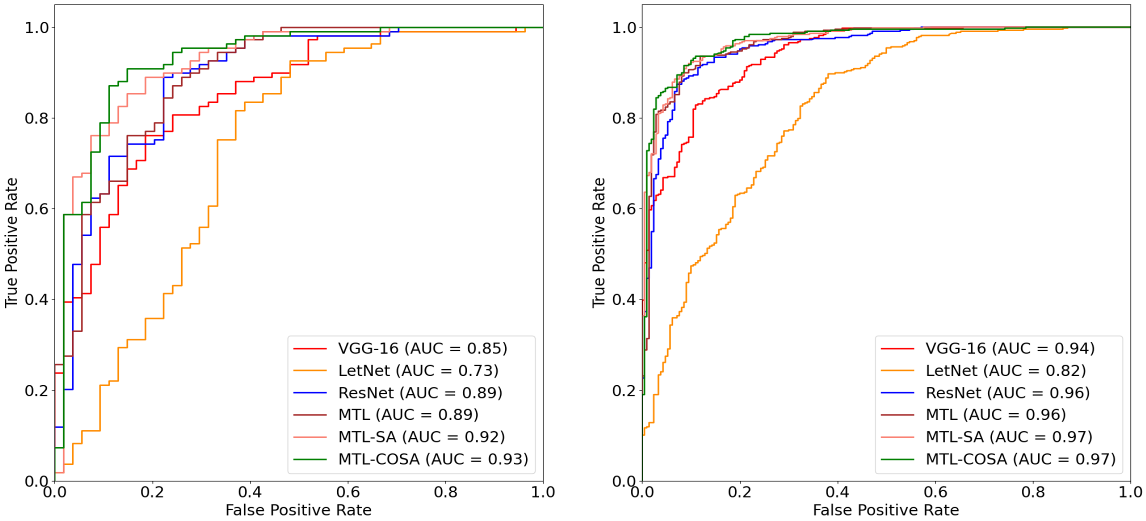


Fig. 6.3: ROC curves of six compared classification methods on Dataset UDIAT (left) and Dataset BUSI (right).

6.5 RMTL-Net Results

In this section, we introduce all competing methods, evaluate the performance of multi-task learning with different hyperparameter λ values in equation 5.7, present a detailed ablation study of the proposed RA module, and evaluate the performance of the proposed RMTL-Net on Dataset UDIAT and Dataset BUSI.

The implementation of the proposed method is based on the public platform PyTorch [75]. All experiments are conducted on Ubuntu 18.04 system, Intel(R) Core(TM) CPU i5-11600K 3.9. All models are trained and tested on a GeForce RTX 3080 Ti graphics card with 12GB memory using the Adam optimizer with momentum β_1 of 0.9, momentum β_2 of 0.99, a weight decay of 0.0001, and a learning rate initialized at 0.0001 and decayed at 10% after every 20 epochs. In the training procedure, the batch size is set as 16 and the number of training epochs is set as 100. Following the empirically optimal setup [55], we adopt batch normalization right after each convolution and before activation. To reduce overfitting, we adopt dropout with a probability of 0.5 in the fully connected layer of the classification network. The contribution weight of loss from the segmentation task (*i.e.*, λ) is empirically set to be 0.9. All competing methods, including ResNet, UResNet, MTL-Net, MTL-COSA, and RMTL-Net models, are pre-trained on ImageNet and fine-tuned with training images selected from datasets UDIAT and BUSI. In this study, all experimental results are reported by averaging the five-fold cross-validation performance.

Competing Methods

Table 6.4 briefly summarizes the task nature and enhanced features of the proposed RMTL-Net and 11 State-Of-The-Art (SOTA) methods. Specifically, we compare RMTL-Net with three recent single-task classification methods (e.g., VGG-16 [54], ResNet-101 [55], and DenseNet [76]), four recent single-task segmentation methods (e.g., FCN [71], PSPNet [72], Deeplab v3+ [73], and U-ResNet), and four recent MTL methods (e.g., MTL-Net, MTL-COSA [29], SHA-MTL [33], and Residual U-Net [62]). U-ResNet is a U-Net [43] with ResNet-101 as its backbone. MTL-Net passes features extracted by $Conv5\_x$ of U-ResNet into a GAP layer followed by a fully connected layer for classification. Table 6.4 shows that some of these compared methods employ feature enhancement strategies such as attention mechanisms and skip connections to improve segmentation and classification performance.

Multi-Task Learning

All compared multi-task learning (MTL) methods including MTL-Net, MTL-COSA [29], SHA-MTL [33], Residual U-Net [62], and the proposed RMTL-Net compute their total loss as the weighted sum of both segmentation and classification losses. In other words, they use the hyperparameter λ in equation 5.7 to balance segmentation and classification performance during MTL. In this section, we evaluate the segmentation and classification performance of RMTL-Net under different λ values. We anticipate observing similar trends for the other compared multi-task methods since MTL-Net, MTL-COSA, and RMTL-Net use U-ResNet and others use a similar network as their backbones.

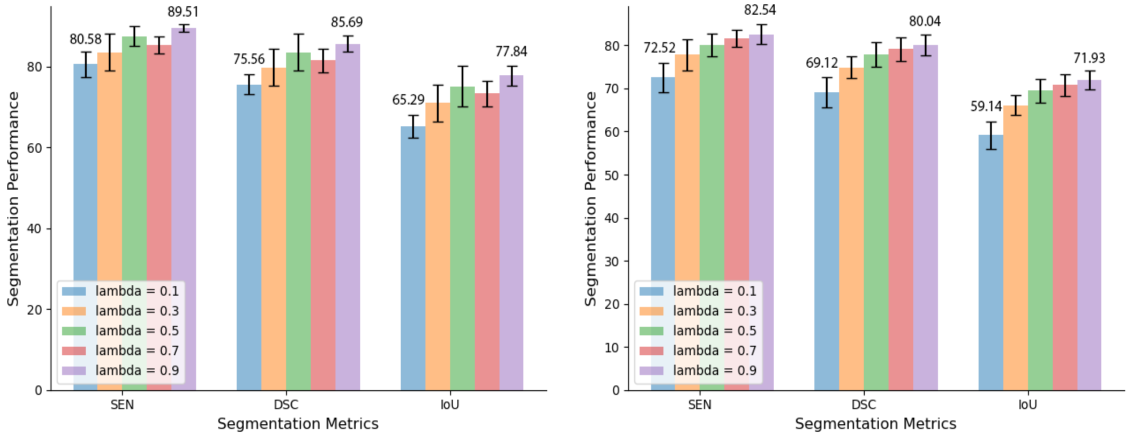
Fig. 6.4 compares the segmentation results of RMTL-Net under five λ values (e.g., 0.1,

Table 6.4: A brief comparison of 11 SOTA methods and the proposed RMTL-Net

Methods	Tasks			Feature Enhancement	
	Single Classification	Single Segmentation	Multi Task	Attention Mechanism	Skip Connection
VGG-16	✓				
DenseNet	✓				
ResNet-101	✓				
FCN		✓			
PSPNet		✓			
Deeplab v3+		✓			✓
U-ResNet		✓			✓
MTL-Net			✓		✓
MTL-COSA			✓	✓	✓
SHA-MTL			✓	✓	✓
Residual-U-Net			✓		✓
RMTL-Net (Proposed)			✓	✓	✓

0.3, 0.5, 0.7, and 0.9) on two datasets. We calculate all five segmentation metrics to evaluate the segmentation results on two datasets under five λ values. It is interesting to observe that SPE and ACC segmentation metrics yield similar values when using different λ values. Specifically, SPE oscillates between a range of 98.97% and 99.25% on dataset UDIAT and between a range of 97.75% and 98.02% on dataset BUSI. Similarly, ACC oscillates between a range of 98.20% and 98.79% on dataset UDIA and between a range of 94.96% and 96.28% on dataset BUSI. As a result, we remove SPE and ACC results in Fig. 6.4 to show values of segmentation metrics SEN, DSC, and IoU, where the narrow bar near the top of each bar indicates the standard deviation and the values above two selected narrow bars present the largest and smallest metric values obtained under five λ values in five-fold experiments. It demonstrates that SEN, DSC, and IoU values increase on both datasets when λ increases, except for $\lambda = 0.7$ on dataset UDIAT.

Fig. 6.5 compares the classification results of RMTL-Net under five λ values (e.g., 0.1, 0.3, 0.5, 0.7, and 0.9) on two datasets. We calculate all six classification metrics to evaluate the classification results on two datasets under five λ values. We re-scale AUC to the range of [0, 100] to ensure all classification values are in the same range for easy display and better understanding. Similar to Fig. 6.4, we use a narrow bar to indicate the standard



(a) Segmentation performance on UDIAT. (b) Segmentation performance on BUSI.

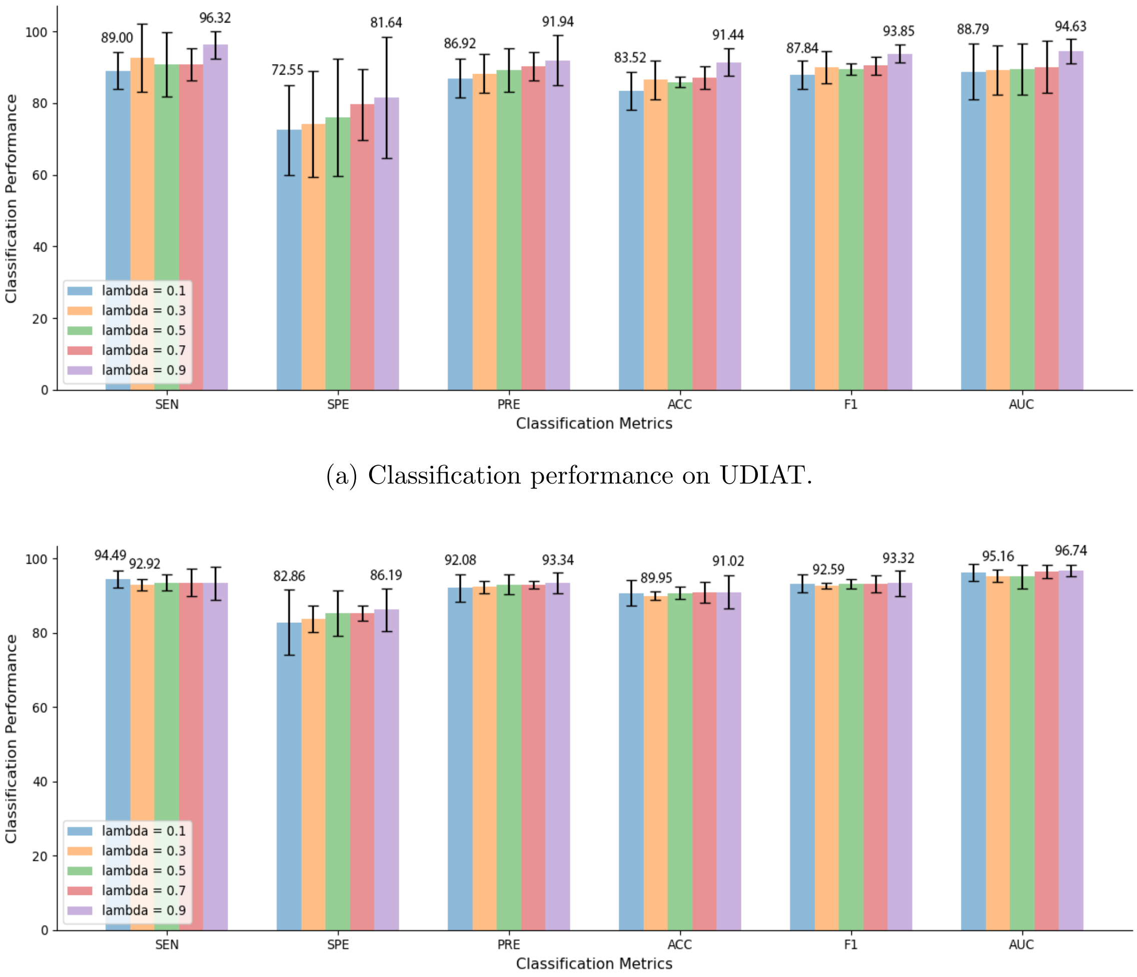
Fig. 6.4: Segmentation results of RMTL-Net on two datasets under different λ values.

deviation for each metric and present the largest and smallest metric values obtained under five λ values in five-fold experiments. It is clear that the overall classification performance of RMTL-Net tends to increase on both datasets when λ increases, except for the SEN values on both datasets.

RMTL-Net uses predicted probability maps to guide the classification task to learn better feature representations and achieve better classification results. As a result, accurate segmentation may lead to a better classifier. Fig. 6.4 and Fig. 6.5 confirm that both segmentation and classification accuracy tends to improve hand in hand when λ increases. Therefore, we set $\lambda = 0.9$ for RMTL-Net to ensure that more weights are given to the dominating task in the MTL framework. We also use the same setting for all MTL methods to ensure a fair comparison.

Ablation Study of RA Module

The regional attention (RA) module is a crucial component of RMTL-Net. It utilizes predicted probability maps to guide the classification network to learn multi-view features from tumor, peritumoral, and background regions in BUS images. To validate the effectiveness of the proposed RA module, we conduct a detailed ablation study by combining information from different region combinations. We list all variants of RMTL-Net below:



(a) Classification performance on UDIAT.

(b) Classification performance on BUSI.

Fig. 6.5: Classification results of RMTL-Net on two datasets under different λ values.

- **Variant 1 (MTL-Net):** None of the three regions are used.
- **Variant 2 (MTL-Net + P):** The peritumoral region is used.
- **Variant 3 (MTL-Net + T):** The tumor region is used.
- **Variant 4 (MTL-Net + B):** The background region is used.
- **Variant 5 (MTL-Net + T + P):** The tumor and peritumoral regions are used.
- **Variant 6 (MTL-Net + P + B):** The peritumoral and background regions are used.
- **Variant 7 (MTL-Net + T + B):** The tumor and background regions are used.
- **Variant 8 (proposed RMTL-Net):** The tumor, peritumoral, and background regions are used.

For Variant 1, the feature map extracted by $Conv5\_x$ of the encoder is directly passed to a GAP layer followed by a fully connected layer for classification. For Variants 2, 3, and 4, the weighted regional feature maps C_P, C_T, and C_B are respectively passed to a GAP layer to obtain a new feature vector G_P, G_T, and G_B of size 1×2048, which are then respectively passed to a fully connected layer for classification. For variants 5, 6, and 7, multi-channel weighted regional feature maps C_T and C_P, C_P and C_B, and C_T and C_B are respectively passed to a GAP layer and concatenated to obtain a new feature vector F of 2×2048. Their corresponding F is then filtered by a 1×1 convolution to get their associated weighted feature vector F_w of 1×2048. Lastly, their corresponding F_w is passed to a fully connected layer for classification.

Tables 6.5 and 6.6 present the segmentation results of eight systems in the ablation study in terms of SEN, SPE, DSC, ACC, and Tumor IoU on datasets UDIAT and BUSI, respectively. Tables 6.7 and 6.8 present the classification results of eight systems in the ablation study in terms of SEN, SPE, PRE, ACC, F_1, and AUC on datasets UDIAT and BUSI, respectively. We observe the following from the results shown in these four tables:

1. Variant 1, which does not incorporate RA, achieves the worst overall segmentation performance when compared with the other seven variant systems. It achieves comparable overall classification performance as the other seven variant systems.

2. Variant 8, which uses tumor, peritumoral, and background regions in the RA module, achieves the best overall segmentation and classification performance when compared with the other seven variant systems.

3. Comparing three variants that use a single region in the RA module, variant 4 involving the background region achieves the best overall performance. Variant 3 involving the tumor region achieves the second-best performance. Variant 2 involving peritumoral regions achieves the worst performance.

4. Comparing three variants that use two of the three regions in the RA module, variant 7 involving tumor and background regions achieves the best performance. Variant 5 involving tumor and peritumoral regions achieves the worst performance.

Table 6.5: Segmentation performance (Mean ± SD) of ablation study on dataset UDIAT.

Variants	SEN	SPE	DSC	ACC	Tumor IoU
MTL-Net	84.28 ± 5.05	99.25 ± 0.14	80.95 ± 5.00	98.65 ± 0.38	72.73 ± 5.49
MTL-Net + P	86.52 ± 1.06	99.20 ± 0.23	82.73 ± 2.52	98.68 ± 0.30	74.92 ± 2.86
MTL-Net + T	87.15 ± 2.74	99.23 ± 0.22	84.03 ± 2.91	98.71 ± 0.27	75.84 ± 3.24
MTL-Net + B	88.43 ± 2.99	99.09 ± 0.31	84.48 ± 3.23	98.65 ± 0.18	76.06 ± 3.57
MTL-Net + T + P	87.97 ± 3.07	99.22 ± 0.24	84.21 ± 4.62	98.69 ± 0.29	76.19 ± 5.17
MTL-Net + P + B	88.68 ± 2.29	99.17 ± 0.23	84.61 ± 2.84	98.68 ± 0.21	76.25 ± 3.16
MTL-Net + T + B	87.87 ± 3.76	99.24 ± 0.34	85.09 ± 2.33	98.72 ± 0.29	76.88 ± 2.51
RMTL-Net	**89.51 ± 0.91**	**99.25 ± 0.19**	**85.69 ± 2.00**	**98.79 ± 0.24**	**77.84 ± 2.45**

For most BUS images, we observe that the background region has the biggest size and the peritumoral region has the smallest size. As a result, we assume that the larger the region, the more information it can provide for both segmentation and classification tasks. The experimental results shown in Tables 6.5, 6.6, 6.7, and 6.8 seem to support this assumption. First, either background, tumor, or peritumoral region plays an important role in the segmentation task since variants 2, 3, and 4 outperform variant 1 without using

Table 6.6: Segmentation performance (Mean ± SD) of ablation study on dataset BUSI.

Variants	SEN	SPE	DSC	ACC	Tumor IoU
MTL-Net	78.91 ± 2.22	98.30 ± 0.25	77.76 ± 3.11	96.18 ± 0.15	69.33 ± 2.89
MTL-Net + P	81.02 ± 2.08	98.05 ± 0.58	79.46 ± 2.84	96.25 ± 0.44	71.31 ± 2.82
MTL-Net + T	81.63 ± 1.92	98.08 ± 0.51	79.55 ± 2.04	96.28 ± 0.27	71.31 ± 1.93
MTL-Net + B	81.84 ± 2.71	98.02 ± 0.43	79.53 ± 2.39	96.25 ± 0.21	71.44 ± 2.20
MTL-Net + T + P	81.30 ± 2.43	**98.08 ± 0.32**	79.64 ± 2.13	96.26 ± 0.20	71.47 ± 1.97
MTL-Net + P + B	81.98 ± 2.25	97.99 ± 0.44	79.79 ± 2.85	96.30 ± 0.24	71.72 ± 2.83
MTL-Net + T + B	81.84 ± 3.26	97.96 ± 0.40	79.91 ± 2.74	96.32 ± 0.30	71.82 ± 2.54
RMTL-Net	**82.54 ± 2.31**	98.00 ± 0.30	**80.04 ± 2.47**	**96.41 ± 0.27**	**71.93 ± 2.15**

Table 6.7: Classification performance (Mean ± SD) of ablation study on dataset UDIAT.

Variants	SEN	SPE	PRE	ACC	F_1	AUC
MTL-Net	89.96 ± 7.43	74.00 ± 13.63	87.99 ± 5.08	84.69 ± 2.85	88.65 ± 2.39	90.82 ± 6.46
MTL-Net + P	92.64 ± 4.12	68.91 ± 16.22	86.09 ± 6.62	84.73 ± 5.52	89.09 ± 3.79	88.26 ± 9.38
MTL-Net + T	91.73 ± 6.76	72.55 ± 17.96	87.66 ± 6.65	85.34 ± 4.82	89.34 ± 3.37	89.54 ± 8.75
MTL-Net + B	92.68 ± 6.87	70.36 ± 20.71	87.24 ± 7.79	85.30 ± 4.40	89.45 ± 2.81	89.90 ± 7.17
MTL-Net + T + P	93.55 ± 7.65	74.18 ± 07.20	88.10 ± 2.55	87.12 ± 3.94	90.56 ± 3.30	91.28 ± 4.65
MTL-Net + P + B	93.55 ± 6.15	77.82 ± 16.43	90.09 ± 6.65	88.33 ± 4.00	91.50 ± 2.84	91.87 ± 7.09
MTL-Net + T + B	94.50 ± 2.01	79.64 ± 11.81	90.62 ± 4.96	89.58 ± 3.41	92.43 ± 2.27	93.02 ± 6.50
RMTL-Net	**96.32 ± 3.82**	**81.64 ± 16.89**	**91.94 ± 6.97**	**91.44 ± 3.90**	**93.85 ± 2.58**	**94.63 ± 3.44**

Table 6.8: Classification performance (Mean ± SD) of ablation study on dataset BUSI.

Variants	SEN	SPE	PRE	ACC	F_1	AUC
MTL-Net	93.36 ± 2.37	84.67 ± 6.65	92.61 ± 3.18	90.18 ± 3.25	93.07 ± 2.41	96.20 ± 2.08
MTL-Net + P	92.21 ± 1.54	80.48 ± 6.16	90.79 ± 2.79	88.40 ± 2.94	91.49 ± 2.11	93.53 ± 2.49
MTL-Net + T	92.45 ± 0.67	82.38 ± 6.86	91.68 ± 2.99	89.17 ± 2.23	92.04 ± 1.52	94.31 ± 1.15
MTL-Net + B	91.30 ± 1.06	85.24 ± 3.91	92.81 ± 1.76	89.33 ± 1.42	92.04 ± 1.03	94.96 ± 1.56
MTL-Net + T + P	92.41 ± 4.98	83.81 ± 4.88	92.28 ± 2.13	89.63 ± 3.44	92.28 ± 2.67	95.87 ± 1.98
MTL-Net + P + B	91.28 ± 3.89	**88.10 ± 4.76**	94.13 ± 2.18	90.25 ± 3.01	92.64 ± 2.36	95.89 ± 1.63
MTL-Net + T + B	92.19 ± 3.51	88.10 ± 6.07	**94.21 ± 2.76**	90.87 ± 3.10	93.15 ± 2.33	95.92 ± 1.22
RMTL-Net	**93.34 ± 4.42**	86.19 ± 5.68	93.34 ± 2.80	**91.02 ± 4.42**	**93.32 ± 3.35**	**96.74 ± 1.48**

the RA module in all segmentation metrics. Second, the background region of C_5 provides the most valuable information for both segmentation and classification tasks since variant 4 achieves the best performance among variants involving one region in the RA module. The tumor region of C_5 provides the second most valuable information followed by the peritumoral region. Third, variants involving two regions in the RA module outperform variants involving one region in the RA module since a combined larger region provides more information to facilitate the learning process. Fourth, the variant involving three regions in the RA module achieves the best performance. Fifth, the weighted feature vector F_W, which obtains valuable information from multiple regions, better represents BUS images than C_5 without using the RA module.

Comparison with Competing Methods

We implement all compared methods except for SHA-MTL and Residual U-Net and conduct experiments using the same parameters to ensure a fair comparison. The authors of SHA-MTL and Residual U-Net did not provide sufficient details on their methods and did not publish their code either. As a result, we directly use their reported segmentation and classification results on dataset BUSI in our comparison. We use the symbol of ”—” to represent a missing result since they did not report their results on each metric. Both methods did not provide any results on dataset UDIAT. So they are not included when comparing segmentation and classification results on dataset UDIAT.

Table 6.9 summarizes the segmentation results of RMTL-Net and six methods in terms of five metrics on the dataset UDIAT. Among four single-task segmentation methods, Deeplabv3+ achieves the best overall segmentation performance with the highest values of SEN, DSC, ACC, and tumor IoU. PSPNet achieves the second-best overall segmentation performance, followed by UResNet and FCN. Among three MTL methods, the proposed RMTL-Net achieves the best segmentation performance in all metrics except for SPE. It improves the second-best method MTL-COSA by 2.54%, 1.62%, 0.02%, and 1.79% for SEN, DSC, ACC, and tumor IoU, respectively.

Table 6.10 summarizes the segmentation results of RMTL-Net and eight methods in

Table 6.9: Segmentation performance (Mean ± SD) of all compared methods on Dataset UDIAT.

Methods	SEN	SPE	DSC	ACC	Tumor IoU
FCN	78.78 ± 6.54	99.08 ± 0.24	76.90 ± 4.69	98.37 ± 0.27	66.49 ± 4.93
PSPNet	83.19 ± 4.60	**99.44 ± 0.16**	83.08 ± 4.58	98.74 ± 0.31	74.59 ± 5.30
Deeplabv3+	85.15 ± 3.54	99.34 ± 0.11	83.53 ± 4.28	98.77 ± 0.32	74.60 ± 4.88
UResNet	85.38 ± 3.84	99.20 ± 0.10	81.38 ± 4.86	98.57 ± 0.35	73.08 ± 5.28
MTL-Net	84.28 ± 5.05	99.25 ± 0.14	80.95 ± 5.00	98.65 ± 0.38	72.73 ± 5.49
MTL-COSA	86.97 ± 2.76	99.27 ± 0.25	84.07 ± 3.25	98.77 ± 0.26	76.05 ± 3.71
RMTL-Net	**89.51 ± 0.91**	99.25 ± 0.19	**85.69 ± 2.00**	**98.79 ± 0.24**	**77.84 ± 2.45**

terms of five metrics on the dataset BUSI. Single-task segmentation methods exhibit similar performance trends on dataset BUSI as on dataset UDIAT. The three MTL methods including MTL-Net, MTL-COSA, and RMTL-Net exhibit similar performance trends on dataset BUSI as on dataset UDIAT. The proposed RMTL-Net achieves the best overall segmentation performance and improves the second-best method MTL-COSA by 3.23%, 1.14%, 0.06%, and 1.28% for SEN, DSC, ACC, and tumor IoU, respectively. Two MTL methods residual-U-Net and SHA-MTL seem to lack credibility since residual-U-Net did not report its standard deviation values for five runs on all evaluation metrics and SHA-MTL reported different values for two equivalent metrics DSC and F_1 without giving any explanation. In addition, residual-U-Net seems to have an overfitting issue since its AUC values of five runs are 0.98, 1, 0.99, 0.97, and 1. As a result, we do not include these two methods here for comparison and list their results in tables for completeness.

Table 6.10: Segmentation performance (Mean ± SD) of all compared methods on Dataset BUSI.

Methods	SEN	SPE	DSC	ACC	Tumor IoU
FCN	78.40 ± 3.33	98.02 ± 0.20	76.87 ± 2.88	96.08 ± 0.12	67.05 ± 2.88
PSPNet	78.28 ± 2.36	**98.34 ± 0.28**	78.48 ± 2.73	96.31 ± 0.46	70.11 ± 2.51
Deeplabv3+	80.71 ± 2.40	98.15 ± 0.42	79.14 ± 2.84	96.37 ± 0.31	70.51 ± 3.04
UResNet	79.65 ± 2.16	98.05 ± 0.40	77.98 ± 2.91	96.06 ± 0.18	69.65 ± 2.92
MTL-Net	78.91 ± 2.22	98.30 ± 0.25	77.76 ± 3.11	96.18 ± 0.15	69.33 ± 2.89
MTL-COSA	79.31 ± 2.48	98.31 ± 0.11	78.90 ± 2.03	96.35 ± 0.16	70.65 ± 2.01
SHA-MTL	81.21 ± 4.83	97.36 ± 1.93	81.42 ± 1.53	95.56 ± 1.08	—
Residual-U-Net	**86.13**	—	**84.81**	88.08	—
RMTL-Net	82.54 ± 2.31	98.00 ± 0.30	80.04 ± 2.47	**96.41 ± 0.27**	**71.93 ± 2.15**

Table 6.11 summarizes the classification results of RMTL-Net and five methods in terms of six metrics on the dataset UDIAT. Among three single-task classification methods, ResNet achieves the best overall classification performance with the highest values of SEN, ACC, F_1, and AUC. DenseNet achieves the second-best overall classification performance, followed by VGG-16. Among three MTL methods, the proposed RMTL-Net achieves the best classification performance in all metrics. It improves the second-best method MTL-COSA by 3.68%, 5.82%, 2.83%, 4.36%, 3.34%, and 1.02% for SEN, SPE, PRE, ACC, F_1, and AUC, respectively.

Table 6.11: Classification performance (Mean ± SD) of all compared methods on Dataset UDIAT.

Methods	SEN	SPE	PRE	ACC	F_1	AUC
VGG16	85.37 ± 4.86	63.09 ± 13.99	82.67 ± 5.49	77.99 ± 4.66	83.85 ± 3.28	86.24 ± 5.03
DenseNet	89.00 ± 2.42	66.73 ± 13.57	84.57 ± 5.62	81.62 ± 5.95	86.69 ± 4.02	86.93 ± 7.49
ResNet	91.69 ± 7.59	64.91 ± 14.65	84.52 ± 4.61	82.80 ± 1.93	87.65 ± 1.63	90.52 ± 5.08
MTL-Net	89.96 ± 7.43	74.00 ± 13.63	87.99 ± 5.08	84.69 ± 2.85	88.65 ± 2.39	90.82 ± 6.46
MTL-COSA	92.64 ± 7.63	75.82 ± 14.02	89.11 ± 5.41	87.08 ± 2.79	90.51 ± 2.29	93.61 ± 4.55
RMTL-Net	**96.32 ± 3.82**	**81.64 ± 16.89**	**91.94 ± 6.97**	**91.44 ± 3.90**	**93.85 ± 2.58**	**94.63 ± 3.44**

Table 6.12 summarizes the classification results of RMTL-Net and seven methods in terms of six metrics on the dataset BUSI. Single-task classification methods exhibit similar performance trends on dataset BUSI as on dataset UDIAT. The three MTL methods including MTL-Net, MTL-COSA, and RMTL-Net exhibit similar performance trends on dataset BUSI as on dataset UDIAT. The proposed RMTL-Net achieves the second-best overall classification performance and MTL-COSA outperforms RMTL-Net by a little bit in all metrics. Due to the lack of credibility, residual U-Net and SHA-MTL are not included here for comparison and are listed in tables for completeness.

Tables 6.9, 6.10, 6.11, and 6.12 demonstrate that RMTL-Net achieves the best overall segmentation and classification results on both datasets. It incorporates the RA module to improve MTL-COSA by learning the importance of three predicted probability maps representing tumor, peritumoral, and background regions. MTL-COSA incorporates self-attention to improve MTL-Net by learning the importance of three regions constructed

Table 6.12: Classification performance (Mean ± SD) of all compared methods on Dataset BUSI.

Methods	SEN	SPE	PRE	ACC	F_1	AUC
VGG16	93.80 ± 3.32	77.63 ± 3.61	89.71 ± 1.66	88.55 ± 2.86	91.69 ± 2.15	94.59 ± 2.67
DenseNet	94.26 ± 3.73	83.33 ± 5.32	92.23 ± 2.24	90.71 ± 2.42	93.18 ± 1.85	95.66 ± 2.123
ResNet	93.81 ± 2.41	80.95 ± 9.37	91.22 ± 3.93	89.63 ± 3.52	92.45 ± 2.49	95.74 ± 2.201
MTL-Net	93.36 ± 2.37	84.67 ± 6.65	92.61 ± 3.18	90.18 ± 3.25	93.07 ± 2.41	96.20 ± 2.08
MTL-COSA	93.57 ± 4.04	87.14 ± 3.19	93.81 ± 1.52	91.49 ± 3.02	93.66 ± 2.36	96.77 ± 1.57
SHA-MTL	96.13 ± 2.33	89.93 ± 5.59	—	94.12 ± 2.45	92.93 ± 3.31	96.28
Residual-U-Net	**98.79**	**94.65**	**98.12**	**97.86**	**98.45**	**99.99**
RMTL-Net	93.34 ± 4.42	86.19 ± 5.68	93.34 ± 2.80	91.02 ± 3.42	93.32 ± 3.35	96.74 ± 1.48

from the predicted binary segmentation mask. MTL-Net decreases the values of three segmentation metrics including SEN, DSC, and tumor IoU (*i.e.*, decreasing the segmentation performance) when compared with the best single-task segmentation method UResNet. This decrease in performance is caused by reduced segmentation weight, which was added to the classification task. Therefore, less weight is employed in training to reduce segmentation errors. However, incorporating attention to MTL-Net addresses this issue to achieve comparable or better segmentation results than UResNet and achieve comparable or better classification results than ResNet.

Table 6.13 lists the number of trainable parameters of all compared methods. It shows that MTL-Net increases trainable parameters of UResNet by 0.004% via adding a lightweight classification task. This simple addition utilizes segmentation results to guide the classification task, which leads to comparable segmentation results as single-task segmentation methods and better classification results than single-task classification methods. Table 6.13 also shows that both MTL-COSA and RMTL-Net increase the different amounts of trainable parameters in networks such as ResNet and UResNet by adding attention modules to learn important regions. RMTL-Net has a simpler attention mechanism than MTL-COSA and therefore leads to a reduction of 16.8% trainable parameters when compared with MTL-COSA. It also outperforms MTL-COSA in segmentation on both datasets and in classification on dataset UDIAT.

Fig. 6.6 presents the segmentation results of RMTL-Net and six compared methods on four representative BUS images: two in Dataset UDIAT as shown in the top two rows

Table 6.13: Summary of the number of trainable parameters of all compared methods.

Methods	Number of Trainable Parameters
VGG-16	138,357,544
DenseNet	52,166,124
ResNet-101	44,677,034
FCN	134,270,278
PSPNet	70,295,620
Deeplabv3+	59,339,426
UResNet	93,500,842
MTL-Net	93,504,940
MTL-COSA	109,241,266
SHA-MTL	—
Residual U-Net	—
RMTL-Net	93,506,099

and two in Dataset BUSI as shown in the bottom two rows. The UDIAT BUS image on the first row contains a small tumor with an irregular boundary. All methods fail to predict a clear and accurate tumor boundary. FCN and MTL-Net completely fail to detect the tumor region. UResNet mistakenly segments a tumor-like region as a tumor. PSPNet, Deeplabv3+, and MTL-COSA segment a tumor partially overlapping with the ground truth. They achieve a tumor IoU value of 61.89%, 60.40%, and 69.23%, respectively. RMTL-Net yields a more accurate segmentation result with the highest IoU value of 76.65%. The UDIAT BUS image on the second row contains a small tumor. FCN, UResNet, and MTL-COSA segment a much bigger tumor region than the ground truth and yield a low tumor IoU value of 35.00%, 31.27%, and 40.98%, respectively. PSPNet, Deeplabv3+, and MTL-Net achieve better segmentation results with tumor IoU values of 59.42%, 51.01%, and 63.33%, respectively. RMTL-Net achieves the best segmentation result and the highest tumor IoU value of 81.60%. The BUSI BUS image on the third row contains a small tumor and a big tumor-like region. All methods except for RMTL-Net mistakenly segment the tumor-like region as tumor region and therefore yield low tumor IoU values less than 55.00%. RMTL-Net segments the correct tumor region and achieves large values close to 1 in almost all segmentation metrics (*i.e.*, 99.90% for SPE, 94.95% for DSC, 99.84% for ACC, and 90.39% for IoU). The BUSI BUS image on the last row contains a small tumor

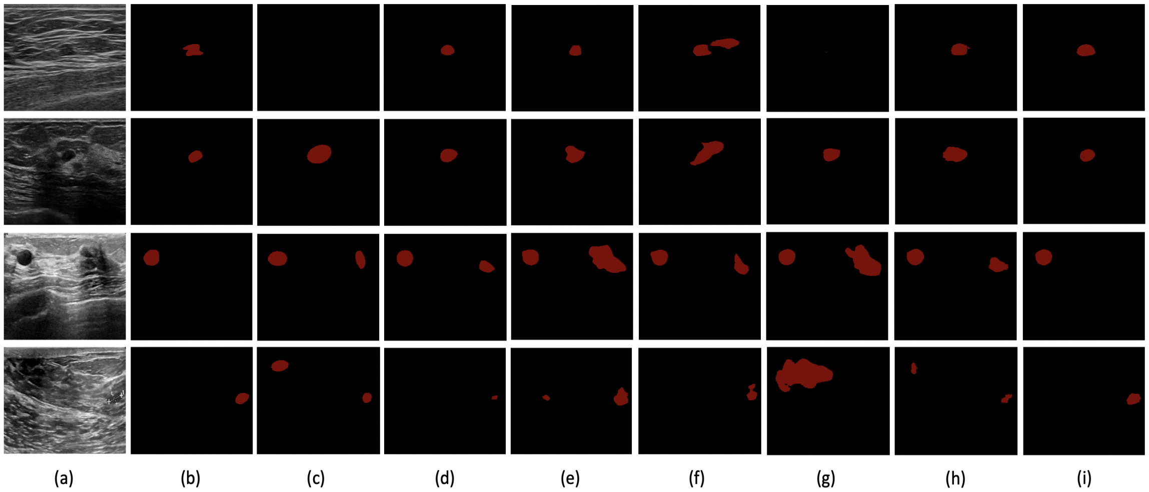


Fig. 6.6: Illustration of segmentation results. (a) BUS images; (b) Ground truth; Segmentation results obtained by (c) FCN; (d) PSPNet; (e) Deeplabv3+; (f) UResNet; (g) MTL-Net; (h) MTL-COSA; (i) RMTL-Net.

with a blurry boundary. This small tumor locates on the right side towards the middle row. MTL-Net segments a completely wrong tumor region and obtains the lowest IoU value of 0.00%. FCN, Deeplabv3+, and MTL-COSA segment a partial tumor region and mistakenly segment another tumor-like region. Their tumor IoU values are 23.56%, 52.04%, and 32.67%, respectively. PSPNet and UResNet segment a partial tumor region with a low IoU value of 17.49%, and 32.58%, respectively. RMTL-Net segments the most accurate tumor region and achieves the largest values on all five segmentation metrics (94.98% for SEN, 99.91% for SPE, 92.37% for DSC, 99.86% for ACC, and 85.82% for IoU).

CHAPTER 7
DISCUSSION

7.1 Comparison of the Proposed Methods

In this section, we compare three proposed methods, including MSSA-Net for BUS image segmentation and MTL-COSA and RMTL-Net for multi-task learning (MTL) of BUS image segmentation and classification. Specifically, we compare the segmentation results of MSSA-Net, MTL-COSA, and RMTL-Net and the classification results of MTL-COSA and RMTL-Net on two publicly available datasets.

Table 7.1 summarizes the segmentation results of the three proposed methods in terms of five metrics on dataset UDIAT and dataset BUSI. All three methods yield good segmentation accuracy with tumor IoU greater than 74% and 69% on dataset UDIAT and dataset BUSI, respectively. RMTL-Net achieves the best overall segmentation result on both datasets, followed by MTL-COSA and MSSA-Net. For the most important segmentation metric, tumor IoU, RMTL-Net improves MTL-COSA by 1.59% and 1.28% on dataset UDIAT and dataset BUSI, respectively. RMTL-Net improves MSSA-Net by 3.28% and 2.51% on dataset UDIAT and dataset BUSI, respectively. Two MTL methods outperform the single-task MSSA-Net, which indicates that the feature sharing between two tasks leads to improved feature extraction and thus improves the segmentation performance. In addition, two MTL methods, RMTL-Net and MTL-COSA, have similar backbone network architectures but different attention modules. RMTL-Net outperforms MTL-COSA, which indicates the RA module of RMTL-Net outperforms the COSA module of MTL-COSA.

Fig. 7.1 presents segmentation results of MSSA-Net, MTL-COSA, and RMTL-Net for two representative BUS images in Dataset UDIAT (top two rows) and two representative BUS images in Dataset BUSI (bottom two rows). For the BUS image on the first row containing a small tumor with an irregular boundary, all methods fail to predict a clear

Table 7.1: Segmentation performance (Mean ± SD) of three proposed methods.

Datasets	Methods	SEN	SPE	DSC	ACC	Tumor IoU
UDIAT	MSSA-Net	84.54 ± 2.49	99.25 ± 0.20	82.96 ± 2.11	98.63 ± 0.51	74.56 ± 2.59
	MTL-COSA	86.97 ± 2.76	**99.27 ± 0.25**	84.07 ± 3.25	98.77 ± 0.26	76.05 ± 3.71
	RMTL-Net	**89.51 ± 0.91**	99.25 ± 0.19	**85.69 ± 2.00**	**98.79 ± 0.24**	**77.84 ± 2.45**
BUSI	MSSA-Net	79.22 ± 2.07	97.93 ± 0.47	78.14 ± 2.84	96.19 ± 0.45	69.42 ± 2.75
	MTL-COSA	79.31 ± 2.48	**98.31 ± 0.11**	78.90 ± 2.03	96.35 ± 0.16	70.65 ± 2.01
	RMTL-Net	**82.54 ± 2.31**	98.00 ± 0.30	**80.04 ± 2.47**	**96.41 ± 0.27**	**71.93 ± 2.15**

tumor boundary. MSSA-Net and MTL-COSA yield tumor IoU values of 69.76% and 69.23%, respectively. RMTL-Net segments the tumor region more accurately with a tumor IoU value of 76.65%. For the BUS image on the second row containing a small tumor, MTL-COSA segments a much bigger tumor region than the ground truth and yields a low tumor IoU value of 40.98%. MSSA-Net segments the tumor region but mistakenly segments a non-tumor region as well. It yields a higher tumor IoU value of 71.07% than MSSA-Net. RMTL-Net achieves the best segmentation result and the highest tumor IoU value of 81.60%. For the BUS image on the third row containing a small tumor and a big tumor-like region, MTL-COSA mistakenly segments a tumor-like region, yielding the lowest tumor IoU values of 47.93%. MSSA-Net and RMTL-Net accurately segment the correct tumor region with high tumor IoU values of 90.35% and 90.39%, respectively. For the BUS image on the last row containing a small tumor with a blurry boundary, MSSA-Net and MTL-COSA mistakenly segment the tumor-like region as the tumor and therefore yield low tumor IoU values of 4.35% and 32.67%, respectively. Overall, RMTL-Net achieves the best segmentation results on all presented BUS images. The segmentation results indicate that RMTL-Net can accurately segment tumor regions and ignore tumor-like regions in BUS images. However, MSSA-Net and MTL-COSA seem to have difficulty differentiating tumor and tumor-like regions.

Fig. 7.2 shows confusion matrices of ResNet, MTL-COSA, and RMTL-Net on dataset UDIAT and dataset BUSI for binary BUS classification. In each confusion matrix, the top-left, top-right, bottom-left, and bottom-right entries represent TN, FP, FN, and TP, respectively. On dataset UDIAT, RMTL-Net has the highest TP and TN values and the lowest FP and FN values and therefore achieves the best classification result. On Dataset

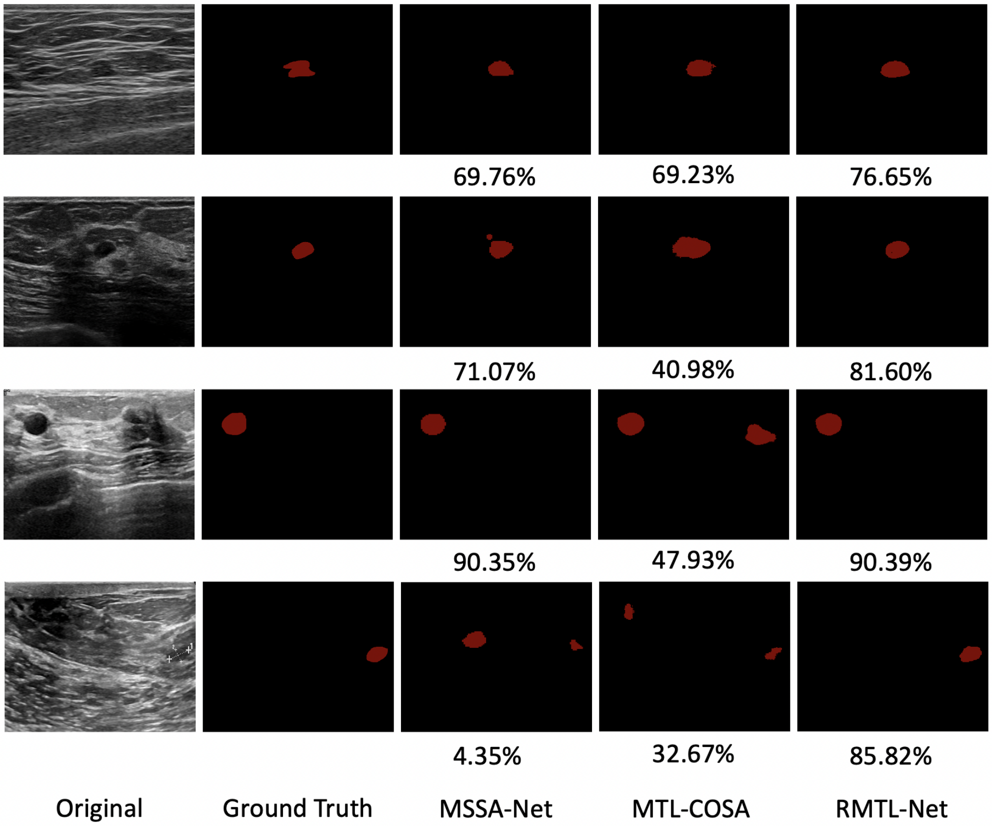


Fig. 7.1: Illustration of segmentation results. The tumor IoU values for each method are displayed below the segmentation results.

BUSI, MTL-COSA has the highest TN and the second highest TP values and the lowest FP and the second lowest FN values. Overall, it achieves the best classification result. cResNet has lower TP and TN values and higher FP and FN values than MTL-COSA and RMTL-Net on both datasets, which implies that MTL is more efficient than individual BUS image classification on both datasets.

Table 7.2 summarizes the classification results of ResNet, MTL-COSA, and RMTL-Net in terms of six metrics on dataset UDIAT and dataset BUSI. ResNet is a single-task classification network. MTL-COSA and RMTL-Net are MTL networks that use ResNet as their encoder. The classification results indicate that MTL of BUS image segmentation and classification is more efficient than individual classification on small datasets. Feature sharing with the segmentation task, which has sufficient training data, improves the feature representation and therefore boosts the classification performance. MTL-COSA and RMTL-

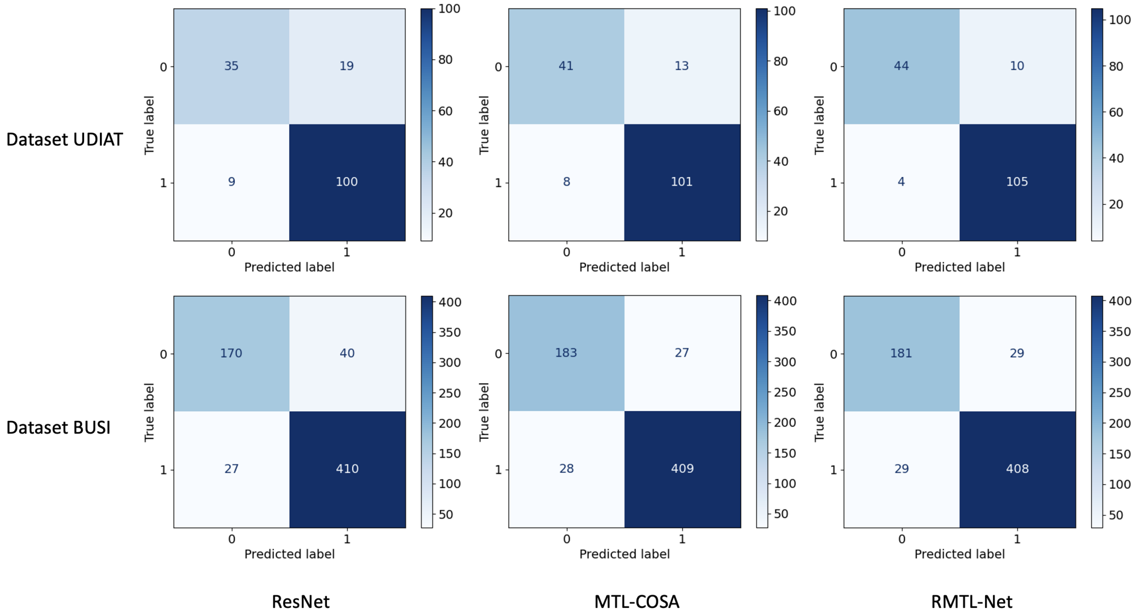


Fig. 7.2: Confusion matrices of ResNet, MTL-COSA, and RMTL-Net on dataset UDIAT and dataset BUSI for binary BUS classification.

Net both achieve good classification results on two datasets with high SEN, PRE, ACC, F_1, and AUC values, mostly over 90%. RMTL-Net outperforms MTL-COSA on dataset UDIAT in terms of all metrics. However, MTL-COSA outperforms RMTL-Net on dataset BUSI in terms of all metrics. One possible reason is that dataset UDIAT and dataset BUSI have different characteristics. For example, dataset UDIAT may have simpler features that can be effectively modeled with a simpler CNN such as RMTL-Net. In contrast, dataset BUSI may have more complex features that require a more complex CNN architecture, such as MTL-COSA. A similar trend can be found in Table 6.11 and Table 6.12. DenseNet has a more complex architecture and higher computational cost than ResNet. ResNet outperforms DenseNet on dataset UDIAT, whereas DenseNet outperforms ResNet on dataset BUSI.

Table 7.2: Classification performance (Mean ± SD) of two multi-task learning methods.

Datasets	Methods	SEN	SPE	PRE	ACC	F_1	AUC
UDIAT	ResNet	91.69 ± 7.59	64.91 ± 14.65	84.52 ± 4.61	82.80 ± 1.93	87.65 ± 1.63	90.52 ± 5.08
	MTL-COSA	92.64 ± 7.63	75.82 ± 14.02	89.11 ± 5.41	87.08 ± 2.79	90.51 ± 2.29	93.61 ± 4.55
	RMTL-Net	**96.32** ± **3.82**	**81.64** ± **16.89**	**91.94** ± **6.97**	**91.44** ± **3.90**	**93.85** ± **2.58**	**94.63** ± **3.44**
BUSI	ResNet	93.81 ± 2.41	80.95 ± 9.37	91.22 ± 3.93	89.63 ± 3.52	92.45 ± 2.49	95.74 ± 2.201
	MTL-COSA	**93.57** ± **4.04**	**87.14** ± **3.19**	**93.81** ± **1.52**	**91.49** ± **3.02**	**93.66** ± **2.36**	**96.77** ± **1.57**
	RMTL-Net	93.34 ± 4.42	86.19 ± 5.68	93.34 ± 2.80	91.02 ± 3.42	93.32 ± 3.35	96.74 ± 1.48

Fig. 7.3 shows ROC curves with values of Area Under the ROC Curve (AUC) for ResNet, MTL-COSA, and RMTL-Net, where AUC values are computed over all BUS images in each dataset. MTL-COSA and RMTL-Net yield high and nearly the same AUC values on both datasets. Specifically, MTL-COSA and RMTL-Net yield AUC values of 0.9268 and 0.9283 on dataset UDIAT and AUC values of 0.9662 and 0.9643 on dataset BUSI, respectively. ResNet achieves the worst AUC values on both datasets. These results further demonstrate the superiority of MTL over single-task classification. It should be noted that these AUC values are not equivalent to the AUC values with standard deviation in Table 7.2, which are computed among five folds in the cross-validation process for each dataset.

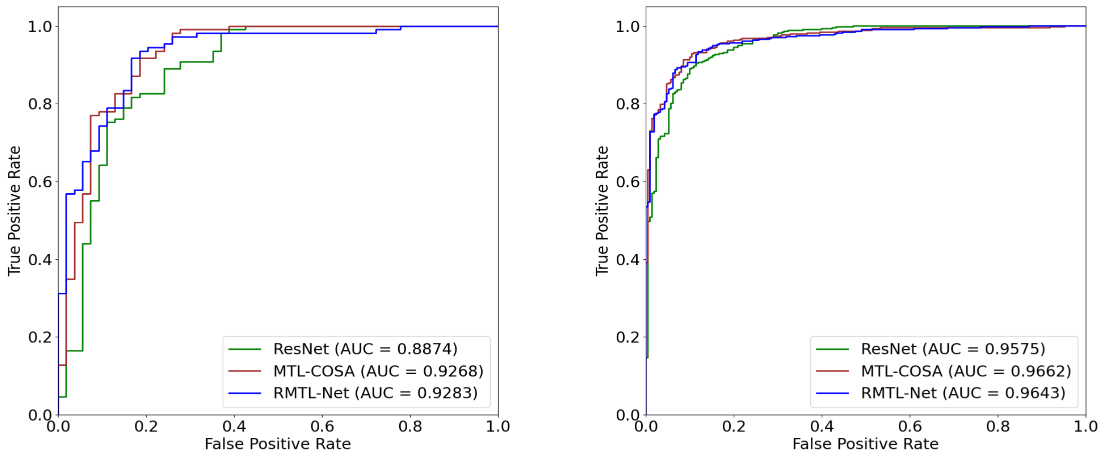


Fig. 7.3: ROC curves of ResNet, MTL-COSA, and RMTL-Net on Dataset UDIAT (left) and Dataset BUSI (right).

Fig. 7.4 presents classification results of ResNet, MTL-COSA, and RMTL-Net for three representative BUS images in Dataset UDIAT (left three columns) and three representative BUS images in Dataset BUSI (right three columns). For the first and fourth images with a benign tumor with a clear boundary, three methods all make a correct prediction. The second and fifth images contain multiple black areas in the background, which makes the tumor hard to identify. ResNet fails to classify them, whereas MTL-COSA and RMTL-Net make correct predictions. The tumors in the third and last images are particularly hard to identify, even for humans. Unfortunately, all three methods fail to make a correct prediction.

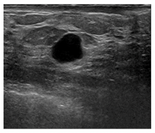
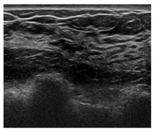
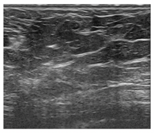
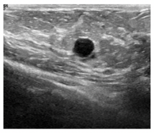
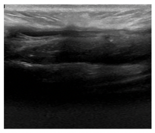
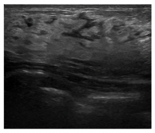
BUS images						
Label	B	M	M	B	B	B
ResNet	B	B	B	B	M	M
MTL-COSA	B	M	B	B	B	M
RMTL-Net	B	M	B	B	B	M

Fig. 7.4: Illustration of classification results. “B” represents benign, and “M” represents malignant. Incorrect predictions are highlighted in red.

Table 7.3 lists the number of trainable parameters and training time of three proposed methods. The training time is for five-fold cross-validation on two datasets. It shows that MSSA-Net has the least trainable parameters but the longest training time of 18 hours. It employs a spatial self-attention module to improve the feature extraction, which is computationally expensive because the spatial self-attention is applied to a high-resolution multi-scale feature map. Two MTL methods have more trainable parameters than MSSA-Net because they add a classification network to the segmentation network. MTL-COSA employs a COSA attention module, which applies a spatial self-attention module to a low-resolution feature map. It takes about 4 hours to train MTL-COSA. RMTL-Net employs a RA attention module, which does not contain a spatial self-attention module and thus has fewer parameters to learn. The RA module is more efficient and much less computationally expensive. It takes only about 3 hours to train and achieves the best overall segmentation and classification results on both datasets.

7.2 Advantages and Potential Usefulness

The advantages and potential usefulness of the proposed three methods are summarized below:

Table 7.3: Summary of the number of trainable parameters and training time of three proposed methods.

Methods	Trainable Parameters	Training Time
MSSA-Net	89,515,791	$\approx$ 18h
MTL-COSA	109,241,266	$\approx$ 4h
RMTL-Net	93,506,099	$\approx$ 3h

1. MSSA-Net improves the original self-attention module by making it take two different inputs to fuse more information. Specifically, one input is a high-level feature map extracted by the deepest layer of the network, and the other input is a multi-scale feature map that is the concatenation of feature maps in different resolutions. The improved self-attention module improves the segmentation performance with a relatively high computational cost compared to the original self-attention module. It can be easily applied to any deep neural network for BUS image segmentation.

2. MTL-COSA performs simultaneous BUS image segmentation and classification by adding a classification branch to U-ResNet. It also proposes a more lightweight and effective COSA module to improve both segmentation and classification performance.

3. RMTL-Net simultaneously performs segmentation and classification by utilizing predicted probability maps to guide the classification task to focus on regions of different importance. It also proposes a more lightweight and effective RA module to improve both segmentation and classification performance.

4. RMTL-Net incorporates a three-region-based attention module (*i.e.*, RA module) to automatically assign appropriate weights to tumor, peritumoral, and background regions during the training procedure. The learned weights help to find regions of importance for better feature representations and therefore improve both the segmentation and classification performance of an MTL method. The RA module aligns well with doctors' clinical perspectives on the importance of tumor, peritumoral, and background regions. The proposed RA module can be easily applied to any existing

MTL methods to incorporate prior medical knowledge into the attention model to improve the performance of multiple tasks.

5. The proposed COSA module and RA module can be easily applied to any MTL network. Our study clearly shows that adding a lightweight classification branch on most existing segmentation methods, at least U-Net-based ones (*e.g.*, UResNet), increases very few parameters but yields both good segmentation and classification results.

6. The study of MTL-COSA and RMTL-Net proves that MTL can achieve better classification results than a standalone classification network on a dataset with a limited number of images. Sharing features with the segmentation task, which has enough training data, can compensate for the lack of training data for the classification task.

7. From a clinical perspective, simultaneous BUS image segmentation and classification are more practical and appealing than single segmentation and classification tasks, as they can provide both tumor boundary as well as tumor category. As a result, MTL in BUS image segmentation and classification is a promising direction that is worthy of more exploration.

7.3 Limitation and Future Work

Our proposed methods have some limitations, as summarized below:

1. MSSA-Net has a high computational cost because it uses a high-resolution multi-scale feature map and a self-attention module.

2. MTL-COSA extracts a small tumor margin region of BUS images, which is not enough to provide the needed information from the peritumoral regions of BUS images.

3. RMTL-Net requires a pre-processing step to generate pseudo ground truths of peritumoral and background regions, which are indispensable in the training procedure to help the network to learn and produce three regions in any test images.

4. The effectiveness of the COSA module and the RA module has not been thoroughly evaluated by comparing them with other traditional spatial or channel attention modules.

5. We do not have a separate testing set and use five-fold cross-validation to have every BUS image in the dataset validated and tested due to the limited number of public BUS images.

In the future, we will test the three proposed methods on larger nuclei segmentation and classification datasets and explore more strategies to improve their generalization ability. We will also compare the proposed COSA module and RA module with more recent spatial and channel attention modules to not only validate its effectiveness but also find a new perspective to improve it.

CHAPTER 8
CONCLUSIONS

In this dissertation, we introduced three different methods for BUS image segmentation and classification and compared their performance with recent state-of-the-art methods. Each of the proposed methods aims to address some of the drawbacks of their peers. Specifically, we can summarize the strategy and performance of each proposed method as follows:

- We propose a novel MSSA-Net for BUS image segmentation. It integrates rich spatial and high-level semantic information via multi-scale feature maps and designs an MSSA mechanism to explore the rich contextual relationships among pixels to boost segmentation performance. MSSA-Net outperforms six state-of-the-art deep neural network-based methods in terms of FPR, JI, DSC, and AER and achieves a comparable performance in TPR on two public datasets.

- We propose a novel MTL-COSA network for simultaneous BUS image segmentation and binary classification. The COSA module utilizes the segmentation output to gain estimated prior medical knowledge and use it to learn contextual relationships for better feature representations in BUS images. MTL-COSA achieves significant classification improvement and comparable segmentation performance on two datasets compared to other state-of-the-art deep learning-based methods.

- We propose a novel RMTL-Net for simultaneous BUS image segmentation and classification. It adopts ResNet-101 as the backbone feature extractor and utilizes a RA module to automatically learn weighted useful information from the tumor, peritumoral, and background regions in BUS images for better segmentation and classification performance. We conduct extensive experiments on two public BUS datasets

UDIAT and BUSI, and the results show that RMTL-Net outperforms recent state-of-the-art single-task segmentation and classification methods and most MTL methods on two datasets.

The contributions of our work include:

- We propose three novel deep learning architectures, including MSSA-Net for BUS image segmentation and MTL-COSA and RMTL-Net for multi-task learning (MTL) of BUS image segmentation and classification.
- We propose three attention modules, including the MSSA module to improve segmentation performance and COSA module and RA module to improve both the segmentation and classification performance of MTL. The proposed three attention modules can be easily applied to any existing BUS image segmentation and MTL methods to improve their performance.
- We evaluate the performance of the three proposed deep learning architectures with attention modules on two public BUS image datasets in terms of several commonly used evaluation metrics. The proposed three methods all outperform recent state-of-the-art methods.
- We prove the effectiveness of MTL of BUS image segmentation and classification on a dataset with a limited number of images. Feature sharing with the segmentation task can compensate for the lack of training data for the classification task.

REFERENCES

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

[2] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling *et al.*, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, 2022.

[3] M. M. Rivera-Franco and E. Leon-Rodriguez, "Delays in breast cancer detection and treatment in developing countries," *Breast cancer: basic and clinical research*, vol. 12, p. 1178223417752677, 2018.

[4] C. A. Anyigba, G. A. Awandare, and L. Paemka, "Breast cancer in sub-saharan africa: The current state and uncertain future," *Experimental Biology and Medicine*, vol. 246, no. 12, pp. 1377–1387, 2021.

[5] O. Ginsburg, C. Yip, A. Brooks, A. Cabanes, M. Caleffi, J. A. Dunstan Yataco, B. Gyawali *et al.*, "Breast cancer early detection: A phased approach to implementation," *Cancer*, vol. 126, pp. 2379–2393, 2020.

[6] R. Sood, A. F. Rositch, D. Shakoor, E. Ambinder, K.-L. Pool, E. Pollack, D. J. Mollura, L. A. Mullen, and S. C. Harvey, "Ultrasound for breast cancer detection globally: a systematic review and meta-analysis," *Journal of global oncology*, vol. 5, pp. 1–17, 2019.

[7] Q. Huang, F. Zhang, and X. Li, "Machine learning in ultrasound computer-aided diagnostic systems: a survey," *BioMed research international*, vol. 2018, 2018.

[8] S. Yang, X. Gao, L. Liu, R. Shu, J. Yan, G. Zhang, Y. Xiao, Y. Ju, N. Zhao, and H. Song, "Performance and reading time of automated breast us with or without computer-aided detection," *Radiology*, vol. 292, no. 3, pp. 540–549, 2019.

[9] V. K. Singh, H. A. Rashwan, M. Abdel-Nasser, M. Sarker, M. Kamal, F. Akram, N. Pandey, S. Romani, and D. Puig, "An efficient solution for breast tumor segmentation and classification in ultrasound images using deep adversarial learning," *arXiv preprint arXiv:1907.00887*, 2019.

[10] K. Wang, S. Liang, S. Zhong, Q. Feng, Z. Ning, and Y. Zhang, "Breast ultrasound image segmentation: A coarse-to-fine fusion convolutional neural network," *Medical Physics*, vol. 48, no. 8, pp. 4262–4278, 2021.

[11] M. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís *et al.*, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.

[12] W. Al-Dhabyani, M. Gomaa, and A. Khaled, H.and Fahmy, “Dataset of breast ultrasound images,” *Data in Brief*, vol. 28, p. 104863, 2020.

[13] Y. Lai, “A comparison of traditional machine learning and deep learning in image recognition,” in *Journal of Physics: Conference Series*, vol. 1314, no. 1. IOP Publishing, 2019, p. 012148.

[14] N. K. Chauhan and K. Singh, “A review on conventional machine learning vs deep learning,” in *2018 International conference on computing, power and communication technologies (GUCON)*. IEEE, 2018, pp. 347–352.

[15] P. Wang, E. Fan, and P. Wang, “Comparative analysis of image classification algorithms based on traditional machine learning and deep learning,” *Pattern Recognition Letters*, vol. 141, pp. 61–67, 2021.

[16] O. Chapelle, B. Scholkopf, and A. Zien, “Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews],” *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[17] S. J. Russell, *Artificial intelligence a modern approach.* Pearson Education, Inc., 2010.

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

[19] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” in *International conference on machine learning.* PMLR, 2013, pp. 1139–1147.

[20] A. Krogh and J. Hertz, “A simple weight decay can improve generalization,” *Advances in neural information processing systems*, vol. 4, 1991.

[21] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[22] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, “Doubleu-net: A deep convolutional neural network for medical image segmentation,” in *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS).* IEEE, 2020, pp. 558–564.

[23] T. Guo, J. Dong, H. Li, and Y. Gao, “Simple convolutional neural network on image classification,” in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA).* IEEE, 2017, pp. 721–724.

[24] C. Szegedy, A. Toshev, and D. Erhan, “Deep neural networks for object detection,” *Advances in neural information processing systems*, vol. 26, 2013.

[25] M. D. Zeiler and R. Fergus, “Stochastic pooling for regularization of deep convolutional neural networks,” *arXiv preprint arXiv:1301.3557*, 2013.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[27] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[28] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *International conference on machine learning.* PMLR, 2019, pp. 7354–7363.

[29] M. Xu, K. Huang, and X. Qi, “Multi-task learning with context-oriented self-attention for breast ultrasound image classification and segmentation,” in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI).* IEEE, 2022, pp. 1–5.

[30] ——, “A regional-attentive multi-task learning framework for breast ultrasound image segmentation and classification,” *IEEE Access*, 2023.

[31] B. Lei, S. Huang, H. Li, R. Li, C. Bian, Y.-H. Chou, J. Qin, P. Zhou, X. Gong, and J.-Z. Cheng, “Self-co-attention neural network for anatomy segmentation in whole breast ultrasound,” *Medical image analysis*, vol. 64, p. 101753, 2020.

[32] H. Li, J.-Z. Cheng, Y.-H. Chou, J. Qin, S. Huang, and B. Lei, “Attentionnet: Learning where to focus via attention mechanism for anatomical segmentation of whole breast ultrasound images,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019).* IEEE, 2019, pp. 1078–1081.

[33] G. Zhang, K. Zhao, Y. Hong, X. Qiu, K. Zhang, and B. Wei, “Sha-mtl: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 10, pp. 1719–1725, 2021.

[34] Y. Luo, Q. Huang, and X. Li, “Segmentation information with attention integration for classification of breast tumor in ultrasound image,” *Pattern Recognition*, vol. 124, p. 108427, 2022.

[35] P. Pan, H. Chen, Y. Li, N. Cai, L. Cheng, and S. Wang, “Tumor segmentation in automated whole breast ultrasound using bidirectional lstm neural network and attention mechanism,” *Ultrasonics*, vol. 110, p. 106271, 2021.

[36] X. Qu, Y. Shi, Y. Hou, and J. Jiang, “An attention-supervised full-resolution residual network for the segmentation of breast ultrasound images,” *Medical physics*, vol. 47, no. 11, pp. 5702–5714, 2020.

[37] G. Pons, J. Martí, R. Martí, S. Ganau, and J. A. Noble, “Breast-lesion segmentation combining b-mode and elastography ultrasound,” *Ultrasonic imaging*, vol. 38, no. 3, pp. 209–224, 2016.

[38] Z. Zhou, W. Wu, S. Wu, P.-H. Tsui, C.-C. Lin, L. Zhang, and T. Wang, “Semi-automatic breast ultrasound image segmentation based on mean shift and graph cuts,” *Ultrasonic imaging*, vol. 36, no. 4, pp. 256–276, 2014.

[39] H.-C. Kuo, M. L. Giger, I. Reiser, K. Drukker, J. M. Boone, K. K. Lindfors, K. Yang, A. V. Edwards, and C. A. Sennett, “Segmentation of breast masses on dedicated breast computed tomography and three-dimensional breast ultrasound images,” *Journal of Medical Imaging*, vol. 1, no. 1, p. 014501, 2014.

[40] Z. Hao, Q. Wang, X. Wang, J. B. Kim, Y. Hwang, B. H. Cho, P. Guo, and W. K. Lee, “Learning a structured graphical model with boosted top-down features for ultrasound image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2013, pp. 227–234.

[41] K. Huang, Y. Zhang, H. Cheng, P. Xing, and B. Zhang, “Semantic segmentation of breast ultrasound image with fuzzy deep learning network and breast anatomy constraints,” *Neurocomputing*, vol. 450, pp. 319–335, 2021.

[42] M. Xian, Y. Zhang, H.-D. Cheng, F. Xu, B. Zhang, and J. Ding, “Automatic breast ultrasound image segmentation: A survey,” *Pattern Recognition*, vol. 79, pp. 340–355, 2018.

[43] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention.* Springer, 2015, pp. 234–241.

[44] M. Amiri, R. Brooks, B. Behboodi, and H. Rivaz, “Two-stage ultrasound image segmentation using u-net and test time augmentation,” *International journal of computer assisted radiology and surgery*, vol. 15, no. 6, pp. 981–988, 2020.

[45] M. Xu, K. Huang, Q. Chen, and X. Qi, “Mssa-net: Multi-scale self-attention network for breast ultrasound image segmentation,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI).* IEEE, 2021, pp. 827–831.

[46] G. Chen, Y. Dai, and J. Zhang, “C-net: Cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation,” *Computer Methods and Programs in Biomedicine*, vol. 225, p. 107086, 2022.

[47] Y. Yan, Y. Liu, Y. Wu, H. Zhang, Y. Zhang, and L. Meng, “Accurate segmentation of breast tumors using ae u-net with hdc model in ultrasound images,” *Biomedical Signal Processing and Control*, vol. 72, p. 103299, 2022.

[48] Y. Liu, L. Ren, X. Cao, and Y. Tong, “Breast tumors recognition based on edge feature extraction using support vector machine,” *Biomedical Signal Processing and Control*, vol. 58, p. 101825, 2020.

[49] L. Cai, X. Wang, Y. Wang, Y. Guo, J. Yu, and Y. Wang, “Robust phase-based texture descriptor for classification of breast ultrasound images,” *Biomedical engineering online*, vol. 14, no. 1, pp. 1–21, 2015.

[50] J. Ding, H. Cheng, M. Xian, Y. Zhang, and F. Xu, “Local-weighted citation-knn algorithm for breast ultrasound image classification,” *Optik*, vol. 126, no. 24, pp. 5188–5193, 2015.

[51] N. Uniyal, H. Eskandari, P. Abolmaesumi, S. Sojoudi, P. Gordon, L. Warren, R. N. Rohling, S. E. Salcudean, and M. Moradi, "Ultrasound rf time series for classification of breast lesions," *IEEE transactions on medical imaging*, vol. 34, no. 2, pp. 652–661, 2014.

[52] M. Abdel-Nasser, J. Melendez, A. Moreno, O. A. Omer, and D. Puig, "Breast tumor classification in ultrasound images using texture analysis and super-resolution methods," *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 84–92, 2017.

[53] K. Huang, M. Xu, and X. Qi, "Ngmms: Neutrosophic gaussian mixture models for breast ultrasound image classification," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3943–3947.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[56] J. Virmani, R. Agarwal *et al.*, "Deep feature extraction and classification of breast ultrasound images," *Multimedia Tools and Applications*, vol. 79, no. 37, pp. 27 257–27 292, 2020.

[57] W.-X. Liao, P. He, J. Hao, X.-Y. Wang, R.-L. Yang, D. An, and L.-G. Cui, "Automatic identification of breast ultrasound image based on supervised block-based region segmentation algorithm and features combination migration deep learning model," *IEEE journal of biomedical and health informatics*, vol. 24, no. 4, pp. 984–993, 2019.

[58] W. Cui, Y. Peng, G. Yuan, W. Cao, Y. Cao, Z. Lu, X. Ni, Z. Yan, and J. Zheng, "Fmrnet: A fused network of multiple tumoral regions for breast tumor classification with ultrasound images," *Medical Physics*, vol. 49, no. 1, pp. 144–157, 2022.

[59] S. Gokhale, "Ultrasound characterization of breast masses," *The Indian Journal of Radiology & Imaging*, vol. 19, no. 3, p. 242, 2009.

[60] W. Yang, S. Zhang, Y. Chen, W. Li, and Y. Chen, "Measuring shape complexity of breast lesions on ultrasound images," in *Medical Imaging 2008: Ultrasonic Imaging and Signal Processing*, vol. 6920. SPIE, 2008, pp. 169–178.

[61] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, and D. Shen, "Multitask learning for segmentation and classification of tumors in 3d automated breast ultrasound images," *Medical Image Analysis*, vol. 70, p. 101918, 2021.

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

[63] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning," *Ultrasonics*, vol. 91, pp. 1–9, 2019.

[64] R. Almajalid, J. Shan, Y. Du, and M. Zhang, "Development of a deep-learning-based method for breast ultrasound image segmentation," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 1103–1108.

[65] M. Xian, Y. Zhang, and H.-D. Cheng, "Fully automatic segmentation of breast ultrasound images based on breast characteristics in space and frequency domains," *Pattern Recognition*, vol. 48, no. 2, pp. 485–497, 2015.

[66] Q. Sun, X. Lin, Y. Zhao, L. Li, K. Yan, D. Liang, D. Sun, and Z.-C. Li, "Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region," *Frontiers in oncology*, vol. 10, p. 53, 2020.

[67] Y.-W. Lee, C.-S. Huang, C.-C. Shih, and R.-F. Chang, "Axillary lymph node metastasis status prediction of early-stage breast cancer using convolutional neural networks," *Computers in Biology and Medicine*, vol. 130, p. 104206, 2021.

[68] T. Liu, Q. Guo, C. Lian, X. Ren, S. Liang, J. Yu, L. Niu, W. Sun, and D. Shen, "Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks," *Medical image analysis*, vol. 58, p. 101555, 2019.

[69] K. Drukker, M. L. Giger, and E. B. Mendelson, "Computerized analysis of shadowing on breast ultrasound for improved lesion detection," *Medical physics*, vol. 30, no. 7, pp. 1833–1842, 2003.

[70] Y. Zhang, M. Xian, H.-D. Cheng, B. Shareef, J. Ding, F. Xu, K. Huang, B. Zhang, C. Ning, and Y. Wang, "Busis: A benchmark for breast ultrasound image segmentation," in *Healthcare*, vol. 10, no. 4. MDPI, 2022, p. 729.

[71] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

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

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

[74] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[76] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

CURRICULUM VITAE

Meng Xu

Education

- Ph.D., Computer Science, Utah State University, Logan, Utah, US, Adviser: Dr. Xiaojun Qi, May 2023.
- B.S., Management Information Systems, Tianjin University of Technology, Tianjin, China. June 2017.

Research Interests

- Computer Vision
- Deep Learning
- Medical Image Analysis

Published Journal Articles

- **M. Xu**, K. Huang, X. Qi, A Regional-Attentive Multi-Task Learning Framework for Breast Ultrasound Image Classification and Segmentation, *IEEE Access*, vol. 11, pp. 5377-5392, 2023.

Published Conference Papers

- **M. Xu**, K. Huang, X. Qi, Multi-Task Learning with Context-Oriented Self-Attention for Breast Ultrasound Image Classification and Segmentation, *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2022.

- K. Huang, **M. Xu**, X. Qi, NGMMs: Neutrosophic Gaussian Mixture Models for Breast Ultrasound Image Classification, *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021.
- Q. Chen, P. Li, **M. Xu**, X. Qi, Sparse Activation Maps for Interpreting 3D Object Detection, *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.
- **M. Xu**, K. Huang, Q. Chen, X. Qi, MSSA-Net: Multi-scale Self-attention Network for Breast Ultrasound Image Segmentation, *IEEE International Symposium on Biomedical Imaging 2021 (ISBI)*, 2021.