参赛队员姓名：Geoffrey Sun

中学：深圳国际交流学院

省份：广东

国家/地区：中国

指导教师姓名：Nadine Brown

论文题目：FeaMix: Multi-level Feature Mixer Network for Breast Cancer Segmentation

FeaMix: Multi-level Feature Mixer Network for Breast Cancer Segmentation

Geoffrey Sun

Shenzhen College of International Education
s21038.sun@stu.scie.com.cn

Abstract. Breast cancer is among the deadliest illnesses endangering women's health globally, and computer-aided segmentation of ultrasonic breast tumor images greatly assists its detection. Most existing methods based on convolutional neural network (CNN) focus on extraction of high-level features. Continuous downsampling in the encoder destroys low-level information, making feature maps highly semantic but poor in spatial clues. As a result, relying on high-level but low-resolution representations alone is inadequate for generating pixel-wise prediction. Therefore, effective integration of multi-level information is critical to dense labeling tasks. A number of existing approaches achieve feature fusion through simple operations (e.g. simple concatenation and addition). Nevertheless, these refinement schemes lack flexibility, which leads to suboptimal performance and feature redundancy. In this paper, a feature mixer network (denoted as FMNet) is proposed, which is a simple and effective end-to-end deep learning framework for semantic segmentation that achieves adaptive multi-level feature fusion with high flexibility. The core of this framework is a multi-level feature mixer (FeaMix) that is able to adaptively aggregate feature maps from different stages using two dynamic weights. For each encoding stage, adjacent feature maps with smaller semantic gap are suppressed by assigning them lower weights, as they carry similar-level features compared to the current stage. Vice versa, it is also ensured that higher weights are given to distanced feature maps that carry more complementary information to the current stage. To further enhance model performance, supervised boundary information is aggregated into the fusion scheme in a similarly learnable fashion. Experiments show that FMNet exceeds state-of-the-art methods on the BUSI breast cancer dataset with a 75.69 % mean IoU.

Keywords: Convolutional Neural Network · Semantic Segmentation · Feature Fusion

1 Introduction

Computer-aided segmentation of ultrasound tumor images assists the diagnosis of breast cancer, one of the most fatal threats to women's health around the world. According to statistics, $2,261,400$ female breast cancer cases were reported globally in 2020, accounting for more than $685,000$ deaths [10]. Ultrasonic imaging is among the most widely utilized and critical methods for detecting breast anomalies. It allows tests to be performed rapidly and affordably on patients, yet there still remained several difficulties to overcome. The clarity and quality of ultrasound imaging are at an inferior level compared to other methods such as MRI and CT scans, making diagnosis fallible even for experienced doctors. Shadows and strong noises are exhibited intensively in feature maps. Furthermore, breast lesions exist in varying scales, irregular shapes, and random distributions on yielded images. Reading ultrasonic images is also extremely time-consuming, and the error rate might grow as more graphs are analyzed [6].

The scientific community has extensively explored the segmentation of tumors in ultrasound images. Preliminary approaches towards this problem mainly rely on traditional and hand-crafted feature extractors and are mostly unsupervised. Boundary detection, for example, makes prediction based on edges delimited by hand-crafted boundary extractors such as the Canny operator [4]. Common defects of such traditional approaches include the insufficient extraction of high-level features and a lack of implementation flexibility. With the development of machine learning, supervised machine learning approaches grew dominant because of their flexibility and performance advantages over traditional methods. An example would be the Makarov Random Field (MRF), which models the dependencies within neighbouring pixels to achieve feature extraction. Another example is Random Decision Forest (RDF), which trains a series of semantic classifiers and uses them in combination to output segmentation results.

In the past few years, as deep learning develops continuously, outstanding performance compared to above methods are obtained in the field of computer vision. Recently, Deep Convolutional Neural Networks (DCNN) have made impressive strides in the challenge of medical image segmentation. As a classical network using the encoder-decoder architecture, fully convolutional network (FCN) was the pioneer among other segmentation models in enabling end-to-end training by applying fully-convolutional layers to replace fully-connected layers before output [13] [24]. One problem for the FCN is that for the decoding process, the model only utilizes the latest high-level feature maps in the encoder subnetwork. In the encoding stage, low-level information in shallower layers, such as object boundaries and details, become insufficient and are outweighed by high-level semantic representations when continuous convolution and pooling are applied. UNet and variants of FCN attempt to address this issue through establishing symmetrical skip-connections between the encoder and the decoder, but this only alleviates the problem partially [30]. Only the feature maps of the same level in the encoder subnetwork are concatenated to supplement the corresponding context in the decoder stage. Informative cross-level context can not be utilized effectively, which leads to the failure in bridging the semantic gap between multi-level feature representations. DeepLabV3 and PSPNet respectively utilizes atrous convolution and spatial pyramid pooling to handle multi-scale features [7] [41]. Nevertheless, these proposals exploit low-level features insufficiently since multi-scale feature extractions are performed only on high-level feature maps to prioritize semantic context mining.

In current investigations, more flexible solutions towards the feature imbalance are discussed. BiSeNet designs a separate path that preserves spatial information and later fuses it with high-level feature maps through channel-wise attention [37]. A recently published context contrasted network (CCN) uses a gated sum technique to aggregate multi-level feature maps [9]. This module benefits feature fusion by allowing the network to automatically choose desirable detail-rich feature maps while filtering out inappropriate ones. G-FRNet adopts the common UNet structure but makes information from the encoder pass through a specially designed gate unit before being forwarded to the encoder [2]. ERN [21] improves model performance by supervising model output with an additional boundary loss, which forces the model to take in spatially enhanced representations. MFENet improves feature fusion using pixel-wise attention vectors generated by receptive fields in various scales, therefore strengthening context representation [39]. CGNet proposes a global guidance module that automatically extracts discriminative features such as object saliency and boundary [40]. DANet appends both channel-wise and pixel-wise attention into the fusion scheme to model spatial and channel dependencies independently [12]. However, although the above fusion schemes boast higher

flexibility, it is still not guaranteed that low-level information plays a sufficiently significant role while feature fusion is taking place.

In this work, a DCNN-based end-to-end deep learning framework is presented, in which multi-level feature fusion is specially modeled to narrow down the semantic gap between multi-level feature maps. For each encoding stage in the encoder subnetwork, feature maps with multi-level information from other stages are supplemented using two adaptive weights. The weights are numerically enlarged for a feature map if it contains information complementary with the current stage. For feature maps carrying analogous feature representation, its corresponding weight is suppressed. To further boost FMNet's performance on generating accurate tumor boundary predictions, supervised edge feature maps are innovatively integrated into the fusion module in a learnable fashion. In addition, FMNet is thoroughly assessed on the difficult BUSI [1] breast tumor dataset. In the experiment, the feature mixer produces noticeable improvements both quantitatively and qualitatively. FMNet reaches state-of-the-art result and outperforms previous state-of-the-art segmentation methods, scoring 75.69 percent in mean IoU. This paper's contributions can be summarized threefold:

- First, this work presents FMNet, a novel DCNN-based framework that dynamically integrates multi-level information to narrow down their semantic gap. In the feature mixer, for each stage in the encoder, feature maps are enhanced by adaptively aggregate features from feature maps of other stages, which is achieved by applying two flexible weights. The weights allow for the emphasis of complementary features while suppressing redundant representations during feature fusion.

- Second, FMNet's performance is further improved via appending supervised boundary information learnably into the feature mixer. Experiments suggest that this act assists the model in outputting accurate predictions surrounding tumor boundaries.

- Third, an exhaustive experiment is carried out to provide a thorough analysis of FMNet's performance both qualitatively and quantitatively. FMNet's results in various evaluations illustrate that this paper's proposal outperforms previous state-of-the-art semantic segmentation methods and obtains greatly improved segmentation quality on the BUSI [1] dataset.

2 Related Work

2.1 Computer aided breast cancer diagnosis

Breast cancer refers to any form of malignant tumors that originate from unconfined cell multiplications in human breasts [31]. Ultrasonic breast imaging has become one of the key components in the process of breast cancer diagnosis. In the past, pathologists manually examine images and search for any malignant tissues to diagnose breast cancer. The biggest deficiency of this procedure is its high reliance on experts' visual inspection. Manpower can be worn out after reading a large number of images, which might lead to fatal misjudgements with considerable costs. This process is also expenditure-unwarranted as double reading might be performed depending on the doctors' confidence in their judgements. Computer-aided diagnosis (CAD) has gained their significance in the process of cancer diagnosis. Specifically in the field of breast cancer diagnosis, solutions based on computer vision enable automatic and rapid classification and segmentation of breast cancer based on ultrasonic images. This approach comes with less human participation, which reduces fatigue-related misjudgements. Much lower inference time is another advantage of CAD compared to manual approaches.

Numerous algorithmic approaches have been adopted for breast cancer segmentation and classification. Early segmentation methods, which mainly depend on hand-crafted extractors and operators, include region-based methods [23] [28] [32], edge-based methods [29] and threshold-based approaches [17] [26]. Region-based proposals mainly use watershed or region growing as tools to obtain a rough area for each component in the image. Edge-based methods obtain segmentation results via applying edge extraction operators such as Sobel and Canny edge detector [4] [22]. Threshold-based methods, on the other hand, often apply Otsu's thresholding [26] as preprocessing procedures. Later, machine learning approaches gradually substitute preliminary methods due to their outstripping performance. Their greatest advantage over traditional methods is that they are learnable and can be adapted to almost any given semantic segmentation task. Support Vector Machine (SVM) is an example of a solution based on machine learning. This method optimizes its parameters by maximizing the distance between the predicted segmentation boundaries and both of the segmentation classes splitted by the boundary line. Many investigations focus on improvements based on the original SVM model, such as [5] [8] [25]. Classification of breast cancer is another area in which machine learning can be used. K-Nearest-Neighbourhood (KNN), for example, is a supervised machine learning method which makes classifications by taking proximity relationships in the sample space into account [11].

2.2 Medical semantic segmentation based on CNN

Segmentation of medical images, though an effective paradigm for diagnosis, remains a challenging task because of speckle noises, low image quality, and high dependence on examiner experience [3] [20]. Deep learning methods have taken the lead in lesion segmentation tasks in recent years. UNet is among the earliest and most favored end-to-end semantic segmentation architectures for medical applications [19] [30]. It novelly utilizes skip-connections to link feature maps of the same level in the encoder and decoder subnetworks. Its major goal is to aggregate multi-scale features by stacking together feature maps with various receptive fields. To enhance its performance, various upgrades based on the original UNet are proposed. UNet++ [42] constructs denser skip-connections linking the encoder and decoder to obtain fine-grained foreground details. The additional connections narrow the semantic gap between encoder and decoder feature maps before they are fused, which is better than the plain skip-connections in the original UNet. Res-UNet [35] applies a weighted attention mechanism on the original UNet. This module enhances the model's discriminative ability when dealing with small segmentation areas by highlighting helpful parts of feature maps while suppressing useless areas.

Other networks have also reached outstanding results in the challenge of medical image segmentation. DeepLabV3 [7] was the pioneer in using dilated convolution with various kernel sizes to detect objects present at various scales. Based on the same exigence of multi-scale detection, PSPNet applies an atrous spatial pyramid pooling module applied after decoding. DANet [12] excavates spatial dependencies using an innovative dual-attention module. This mechanism consists of both spatial-wise attention and channel-wise attention, which respectively connects pixels in the same feature map and builds inter-channel connection between feature maps. To cope with the loss of context information in the encoder, CE-Net [16] novelly proposes a dense atrous convolution module to capture wider and deeper context information. In addition, to further model the multi-scale context information obtained, it applies a residual multi-kernel pooling block to carry out various pooling operations.

2.3 Feature fusion in semantic segmentation

Feature fusion for different purposes is researched extensively in the field of semantic segmentation. The architecture in the Laplacian Pyramid Reconstruction Network [14] fuses low-level feature maps with high-level representations. Its proposed mechanism uses spatial information rich in low-level feature maps to fix residual errors found in high-level feature maps, which enhances model performance. Global Convolutional Network [27] fuses low-level feature maps into the model pipeline in an attempt to refine segmentation boundaries. Some other methods fuses feature maps obtained from their own proposed module to boost model performance. ISHFNet [36] aggregates representations obtained from two separable convolution branches to construct a comprehensive feature overview; which is later decoded in a hierachical fashion. DASSNet [18] applies channel-wise attention to fuse high-level features obtained, which is later upsampled to higher spatial resolution and supervised.

3 Methodology

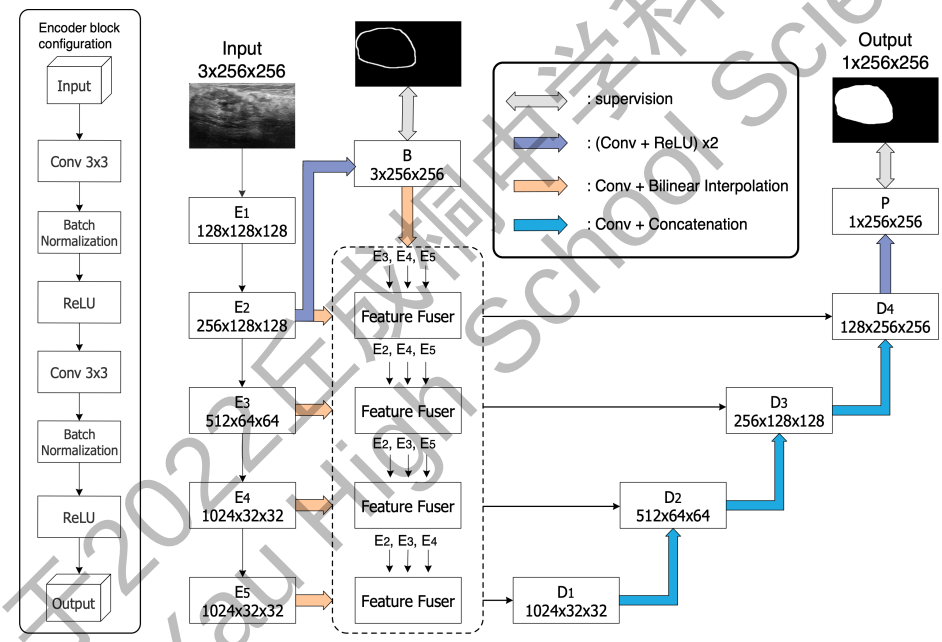

Fig. 1. A schematic illustration of FMNet's architecture

3.1 Motivation and overview

In a typical semantic segmentation network based on CNN, low-level information such as object boundary and spatial clues are mostly abundant in shallow layers of the encoder subnetwork. High-level information, on the other hand, is highly semantic and abstract, and can be found in deeper encoding layers. Current dense-prediction proposals have obtained improved performance when

compared against preliminary approaches. However, many of these methods solely rely on skip-connections or simple concatenation to achieve the integration of cross-level or same-level feature maps, which is inadequately adaptive and flexible. In the feature mixer proposed in this paper, two adaptive weights are used to generate a dynamic "ingredient table" for multi-level feature maps, which flexibly manipulate their participation in the process of feature mixing. In addition, to further improve the prediction accuracy on object boundaries, a supervised boundary map is integrated into the mixer's workflow.

An illustration of FMNet's architecture is displayed in Fig.1. The model workflow can be summarized as follows: (1) Ultrasound breast cancer images are fed into a CNN encoder to extract multi-level feature maps. (2) To effectively aggregate feature maps of one stage with information from other stages, two adaptive weights are simultaneously assigned to ensure that complementary information are emphasized while redundant representations are suppressed. (3) Supervised edge information is integrated into the fusion scheme using an independent learnable coefficient to enhance the FMNet's accuracy in predicting boundary pixels.

3.2 Initial extractor

As shown in Fig.1, FMNet adopts the encoder-decoder architecture with a replaceable backbone model. UNet is selected as the CNN backbone, but it's worth noting that implementation of the feature mixer is independent of the backbone network. The encoder of FMNet consists of five encoding blocks each with the configuration displayed on the left of Fig.1. Each encoding block contains two downsampling packages, each composed of a 3×3 convolution, a batch normalization and a ReLU activation function. The outputs of the blocks are defined using the set $E_i, i \in [1, 2, 3, 4, 5]$, where each E_i denotes a different encoding stage. A similar approach is used to define the four decoding stages in FMNet with $D_i, i \in [0, 1, 2, 3, 4]$, where each D_i denotes a different decoding stage.

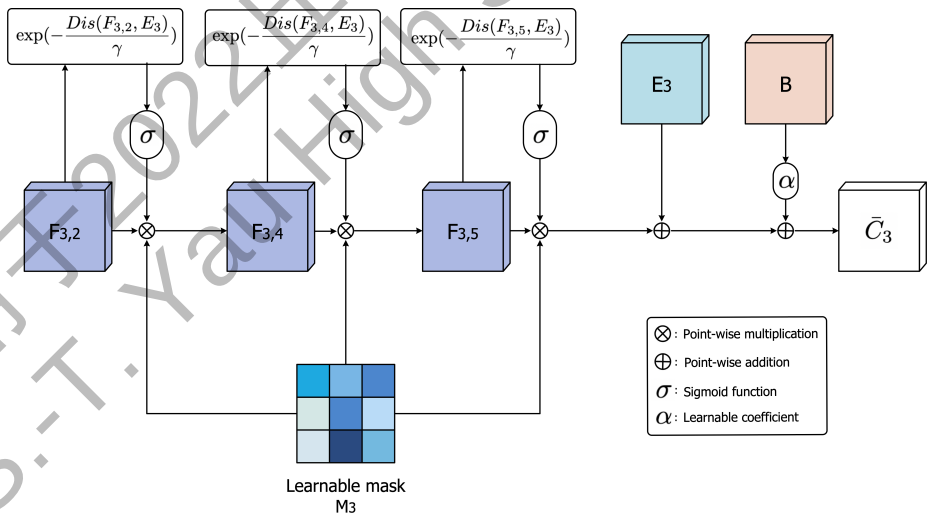

Fig. 2. A detailed diagram illustrating the work flow of the feature mixer

3.3 Multi-level feature mixer

This section focuses on the implementation of two different forms of adaptive weight, which are the core of the feature mixer. It's worth noting that as E_1 is too shallow and contain noises that can potentially jeopardize feature fusion, this stage is ignored during feature fusion and boundary extraction. In order to aggregate multi-stage feature maps to the current stage, their spatial resolution and channel number are first adjusted to fit the current stage. A rescaler function is performed on the encoding stages to be fused as follows:

$$F_{i,j} = \phi(Up(E_j), \theta_F), \forall j \in [2,3,4,5], j \neq i \tag{1}$$

where $Up(\cdot)$ is bilinear interpolation that upsample feature maps to the same size as E_i, and $\phi(\cdot)$ denotes 1×1 convolution with parameter θ_F.

Adaptive weight According to previous assumptions, features carried by distant feature maps in the encoder subnetwork tend to be complementary in nature. This results in the conception that they can be integrated to maximize feature efficiency and minimize feature redundancy. Based on such exigence, an adaptive weight is proposed that optimizes feature fusion based on the level of redundancy between two feature maps. In an attempt to enhance E_i, the weight $W_{i,j}$ assigned to resized feature maps from other levels $F_{i,j}$ is defined as:

$$W_{i,j} = \exp(-\frac{Dis(F_{i,j}, E_i)}{\gamma}), \forall j \in [2,3,4,5], j \neq i \tag{2}$$

where γ is an artificially defined coefficient, here given the value of 0.7. The distance function $Dis(\cdot)$ measures the complementarity between two feature maps, here defined as:

$$Dis(X,Y) = \sum_{i=0}^{H}\sum_{j=0}^{W} ||X_{i,j} - Y_{i,j}||^2 \tag{3}$$

Now the FMNet is able to obtain strongly enhanced feature representations by applying the defined weight to each resized feature maps at different stages in the following way:

$$C_i = E_i + \sum_{n=2}^{5} F_{i,n} \cdot \sigma(F_{i,n})^{W_{i,n}}, \forall i \in [2,3,4,5], n \neq i \tag{4}$$

where C_i represents feature maps belonging to stage i after enhanced with multi-level features adaptively and $\sigma(\cdot)$ denotes the sigmoid function. For a feature map containing representations complementary with the current stage, the distance function would generate a higher output which results in a smaller W. However, since all values from other stages are mapped into $(0,1)$ by the sigmoid function, a smaller W as its power would exert a hyperbolic effect on distinctive feature maps, thus ensuring that complementary information participates in the fusion process more actively .

Pixel-wise guidance Although the above mechanism can be deployed alone with small additional computational cost, the fusion process needs further refinement. An additional mechanism is applied

to provide a more flexible pixel-wise guidance to multi-level feature maps during their aggregation using a learnbale mask. The learnable mask for stage i is defined using the denotation:

$$M_i, i \in [2, 3, 4, 5] \tag{5}$$

All M_i possesses the same spatial resolution as E_i and has only 1 channel. In order to ensure that adjacent feature maps are suppressed while distant feature maps containing complementary information are emphasized, this mask is integrated using the following form:

$$C_i = E_i + \sum_{n=2}^{5} (1 - \sigma(M_i)^{|n-i|}) \cdot F_{i,n} \cdot \sigma(F_{i,n})^{W_{i,n}}, \forall i \in [2, 3, 4, 5], n \neq i \tag{6}$$

In this equation, M_i is first suppressed between $(0, 1)$ using the sigmoid function and assign it with the distance between the current stage and other stages. To ensure that distant feature maps are assigned larger weights than adjacent ones, the values in the mask obtained by the last operation are subtracted from one. This mechanism further narrows the semantic gap between multi-level feature representations and contributes to the flexibility of the fusion scheme simultaneously.

3.4 Edge Enhancement

In this section, the aim is to refine feature maps with edge details by combining supervised boundary information into the workflow. Object boundary information, as a low-level representation, is more abundant in shallower layers of the CNN backbone because of large spatial resolution, so E_1 seems to be the optimal choice for reconstructing tumor boundaries. However, the trade-off between spatial information concentration and image quality makes E_1 too noisy to be implemented without hampering segmentation quality. Consequently, the extraction of boundary information denoted as B is carried out based on E_2:

$$B = \sigma(\phi(E_2, \theta_B)) \tag{7}$$

where $\phi(\cdot)$ denotes a series of 1×1 convolution with parameter θ_B and $\sigma(\cdot)$ represents the sigmoid function. To specifically model tumor boundary, the obtained edge map is supervised using Binary Cross Entropy (BCE) loss:

$$\mathcal{L}_B(B, \bar{B}) = -\frac{1}{N} \sum_{i=0}^{N} (\bar{B}_i \log(B_i) + (1 - \bar{B}_i) \log(1 - B_i)) \tag{8}$$

where $\bar{B}$ denotes tumor edges extracted from the ground truth during training using the canny operator [4] (which can be considered the ground truth for tumor boundary). After obtaining the supervised boundary map B, bilinear interpolation and several 1×1 convolution are applied for suitable channel number and shape . The guided boundaries are then integrated into the fusion path in the following form:

$$\bar{C}_i = C_i + \alpha B \tag{9}$$

where $\bar{C}_i$ denotes the further upgraded feature maps of stage i and α represents a learnable coefficient that adds to the flexibility of edge enhancement.

3.5 Decoding scheme

UNet-style upsampling is adopted to obtain a dense prediction map for breast lesions. The application of the feature mixer effectively avoids UNet's defect that skip-connections only link same-level information, as all feature maps involved in upsampling are itself multi-level. The decoder in FMNet aggregates enhanced feature maps $\bar{C}_i$ step by step to obtain representations D_i:

$$D_i = \begin{cases} \bar{C}_5, i = 0 \\ \phi^T(Cat(\phi(D_{i-1}, \theta_i), \bar{C}_{6-i}), \theta_i^T), \forall i \in \{1, 2, 3, 4\} \end{cases} \tag{10}$$

where $\phi(\cdot)$ and $\phi^T(\cdot)$ respectively denotes 1×1 convolution with parameter θ_i and transposed convolution (deconvolution) with parameter θ_i^T, and $Cat(\cdot)$ represents matrix concatenation. A series of convolutions are applied to the last upsampling layer D_4 before a prediction mask is generated by FMNet. The following formula is used to determine the prediction mask:

$$P = \arg\max(\phi(Up(D_4), \theta_P)) \tag{11}$$

where $Up(\cdot)$ denotes bilinear interpolation and $\phi(\cdot)$ represents a series of 1×1 convolution with parameters θ_P.

3.6 Optimization

A supervision using the BCE loss is added between the model output and the ground truth and defined as below:

$$\mathcal{L}_P(P, Y) = -\frac{1}{N}\sum_{i=0}^{N}(Y_i \log(P_i) + (1 - Y_i)\log(1 - P_i)) \tag{12}$$

where Y is the ground truth. In the framework, a composite loss is adopted to supervise both the extracted object boundaries and segmentation result. The weight assigned to both loss is one, and thus the total loss is defined as the direct addition between the two losses:

$$\mathcal{L}_T = \mathcal{L}_B + \mathcal{L}_P \tag{13}$$

where $\mathcal{L}_T$ denotes the total loss obtained, which is optimzed using standard back-propagation.

4 Experiments

4.1 Dataset

To prove that gratifying improvements are made, various experiments are conducted on the BUSI [1] dataset to evaluate FMNet's performance. The BUSI [1] dataset collects 780 breast ultrasound images from 600 female patients in 2018, which contains 133 normal cases, 487 benign cases and 210 malignant cases. Images in the dataset have an average resolution of 500×500, and are cropped to 256×256 before input to reduce computational cost. During a standard diagnostic procedure for breast cancer, the existence of a tumor is first proved by clinicians before they are segmented in greater detail. As a result, the normal cases with no masks is removed to simulate this situation, but an additional experiment where all three categories are present is also carried out.

4.2 Training and inference

All experiments are performed on PyTorch deep learning framework with Intel Xeon Platinum 8255C 2.50GHz CPU and Nvidia RTX 3090 GPU. The software environment consists of PyTorch 1.12.0, Python 3.8.10, CUDA 11.6 and cuDNN 8.3.0. The Adams optimizer [33] with weight decay of 0.0005 and initial learning rate of 0.0001 is selected to optimize FMNet on the BUSI [1] dataset. Each model tested is optimized for 100 epochs after being initialized with random weights, and the batch size of both training set and validation set is 2. The evaluation metrics are obtained by calculating the mean result from three successive tests on the entire validation set. The deviation value is calculated by subtracting the highest results with the lowest from the three tests, which measures the stability of a specific model on individual metric.

4.3 Comparison scheme

In horizontal comparisons, performance of FMNet is compared against previous state-of-the-arts, including UNet [30], DeepLabV3 [7], PSPNet [41], CENet [16], AGNet [34] and PyDiNet [15]. To ensure that comparisons are fair, the same iteration scheme is applied on every model experimented. In the table, the best result of each metric is bolded and the second-best is underlined.

4.4 Evaluation metric

Four commonly used evaluation metrics are employed to evaluate FMNet and other methods quantitatively. They are respectively the Jaccard index (denoted as JA, also known as mIoU), Dice coefficient (denoted as DI), sensitivity (denoted as SE) and specificity (denoted as SP). The following formulae may be used to calculate the metrics assessing model performance:

$$\begin{aligned} JA &= \frac{TP}{TP+FN+FP} \\ DI &= \frac{2TP}{2TP+FN+FP} \\ SE &= \frac{TP}{TP+FN} \\ SP &= \frac{TN}{TN+FP} \end{aligned} \tag{14}$$

It's worth noting that, following preexisting contributions, the Jaccard index is regarded as the major criteria evaluating a model's performance.

4.5 Results

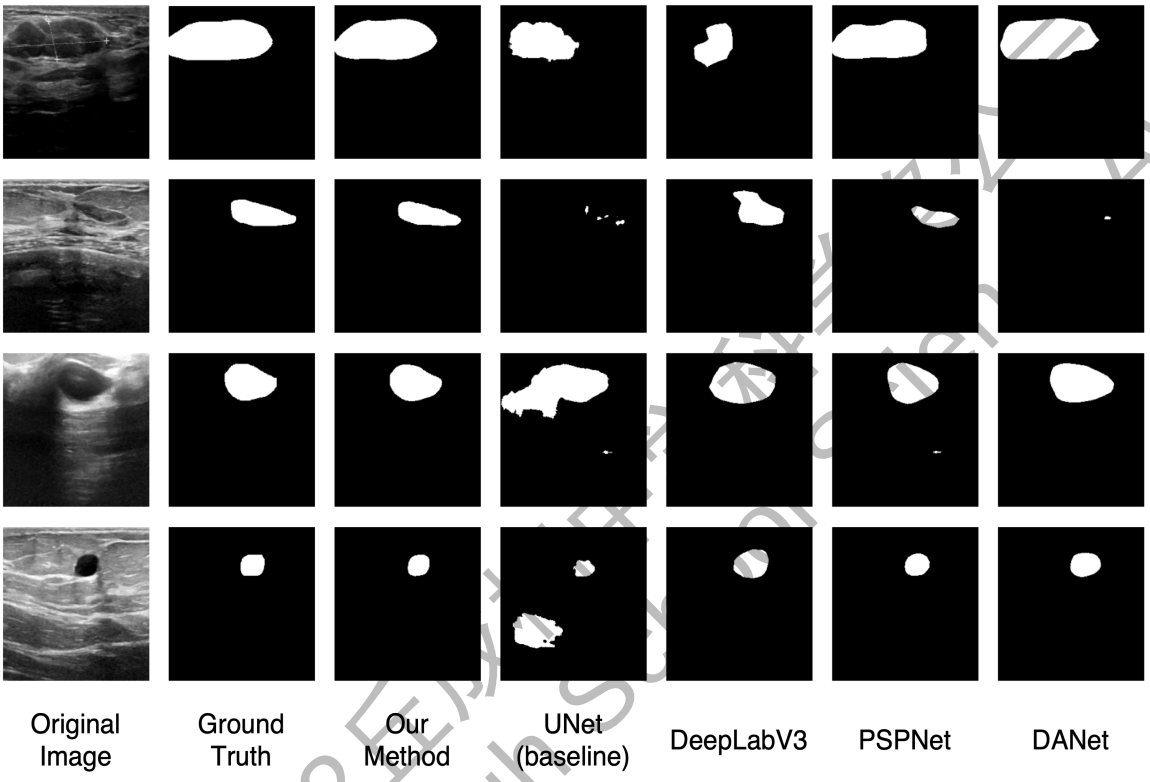

Fig. 3. A visual comparison between FMNet's segmentation result with other semantic segmentation methods on four samples from the dataset

Qualitative evaluation In Fig.3, four samples are chosen at random from the BUSI [1] dataset to qualitatively evaluate the performance of FMNet and other methods. In example 1, UNet and DeepLabV3 neglect the majority of the area containing breast tumor and generate unsmooth boundaries, while PSPNet and DANet tend to produce inaccurate boundary prediction. FMNet, by contrast, generates much more accurate prediction on tumor edge because of the integrated boundary map. In example 2, UNet and DANet barely detect the presence of tumor, while DeepLabV3 outputs improper tumor shape. UNet respectively generates prediction of wrong shape and location in example 3 and 4. Predictions generated by FMNet remain stable and high-quality in all four examples.

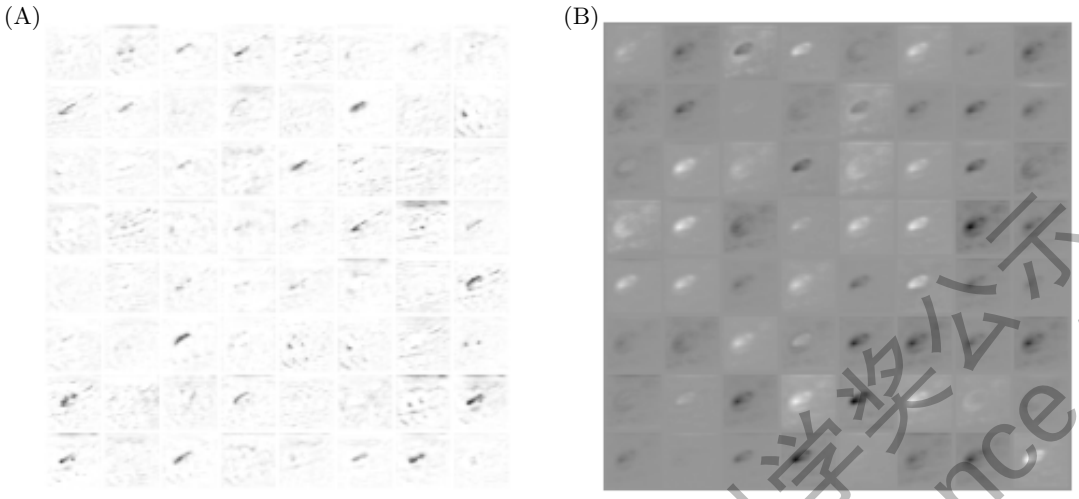

Fig. 4. A visual comparison between mid-layer feature maps from (A) E_3, before the feature mixer is applied; (B) $\bar{C}_3$, after the feature mixer is applied.

In Fig.4, 64 feature maps respectively from E_3 and $\bar{C}_3$ are demonstrated. Before the feature mixer is applied, features represented by middle layers are rough, inaccurate and noisy. The majority of feature maps from E_3 are devoid of any indication of object lesions. Even for those maps where the tumor is recognizable, its saliency is inadequate and there is low contrast between the tumor and the background. However, after E_3 is enhanced using out feature mixer and turned into $\bar{C}_3$, the quality of mid-layer feature maps is greatly improved. The target tumor is recognizable on most of the feature maps with higher contrast with the background. In addition, the variance of tumor shape on different maps is significantly reduced, marking a more stable model performance.

Quantitative analysis Table 1 records the mean and deviation of the four metrics tested on models trained on the BUSI [1] dataset with the normal class removed. Compared against other models, FMNet obtains the best Jaccard, Dice, sensitivity and specificity score, which demonstrates that FMNet can segment breast lesions with higher precision and accuracy. An improvement of 10.11% mIoU, 12.45% Dice, 9.09% sensitivity and 0.96% specificity over the baseline is achieved. Table 2 records the evaluation of models trained with the presence of the normal class. Again, FMNet ranked the top among all the 7 models compared in all four metrics. Fig.5 provides a visual illustration of the above experiment.

Table 1. Quantitative comparison of FMNet with other approaches on the BUSI dataset with normal cases removed

method	JA(%)	DI(%)	SE(%)	SP(%)
UNet (baseline) [30]	65.58±1.96	75.73±1.09	76.16±1.27	98.77±0.01
DeepLabV3 [38]	68.41±3.31	71.60±3.84	72.93±3.72	98.54±0.04
PSPNet [41]	69.81±3.02	78.78±2.57	77.05±2.54	99.16±0.04
DANet [12]	68.32±2.58	76.59±1.47	76.07±1.14	99.04±0.00
CENet [16]	67.49±1.49	75.96±1.37	75.89±1.73	98.93±0.02
AGNet [34]	69.19±2.53	74.88±1.42	79.49±1.32	99.64±0.01
PyDiNet [15]	69.78±1.59	76.49±1.92	82.46±1.88	99.43±0.03
FMNet	**75.69 ± 2.15**	**88.18±1.15**	**85.25±0.92**	**99.73±0.03**

Table 2. Quantitative comparison of FMNet with other approaches on the BUSI dataset without removing normal cases

method	JA(%)	DI(%)	SE(%)	SP(%)
UNet (baseline) [30]	58.64±2.07	63.39±2.73	69.16±2.44	94.77±0.03
DeepLabV3 [38]	60.43±3.52	59.06±2.18	66.93±1.79	94.54±0.02
PSPNet [41]	60.37±2.78	66.13±1.46	70.45±1.62	95.76±0.01
DANet [12]	58.91±1.59	65.28±1.57	66.07±2.31	95.04±0.01
CENet [16]	58.02±2.47	64.17±1.62	66.89±1.72	94.93±0.00
AGNet [34]	60.47±2.38	63.08±2.05	70.29±1.49	95.64±0.02
PyDiNet [15]	61.77±1.62	61.49±2.72	73.46±1.68	95.56±0.04
FMNet	**66.81 ± 1.53**	**75.18±1.77**	**76.25±0.67**	**95.83±0.02**

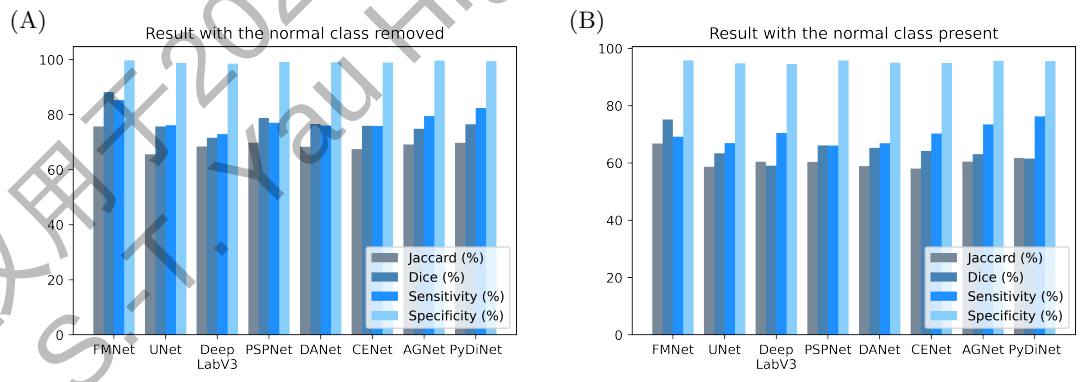

Fig. 5. A schematic illustration of the qualitative evaluation result (A) when the normal cases are removed; (B) when the normal cases are present.

Table 3. Comparison of FMNet's computational efficiency with other methods

method	Flops(G)	Params(M)
UNet (baseline) [30]	135.93	37.66
DeepLabV3 [38]	34.09	58.63
PSPNet [41]	131.44	65.70
DANet [12]	289.27	65.18
CENet [16]	109.43	32.85
AGNet [34]	123.89	37.96
PyDiNet [15]	163.48	47.64
FMNet	148.88	38.24

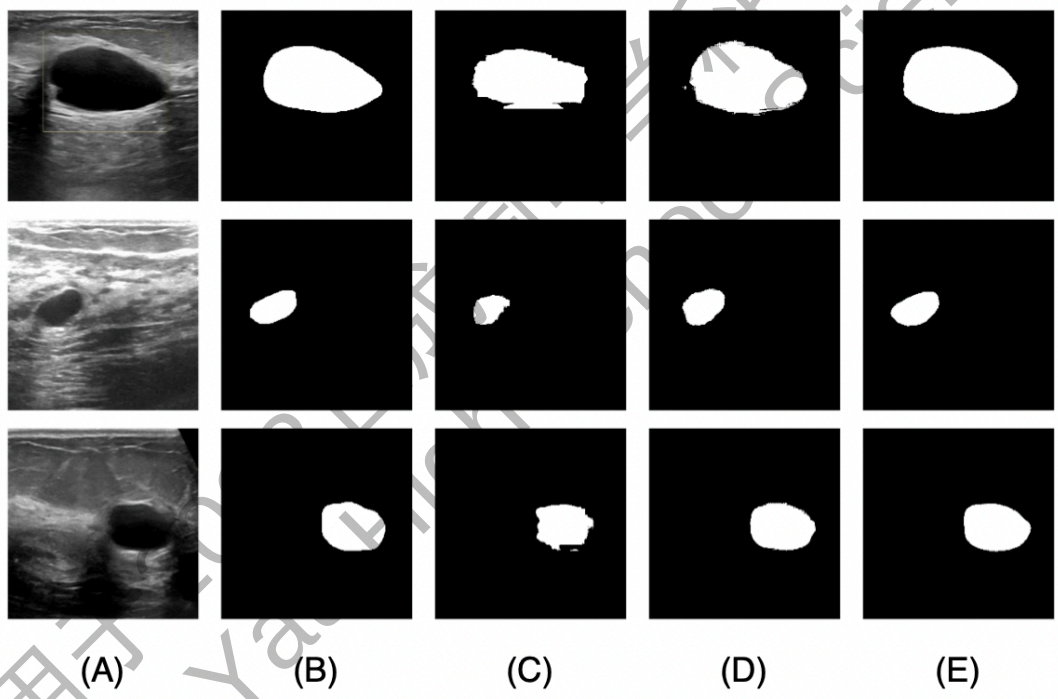

Fig. 6. A visual presentation of the ablation experiment. (A) the original image; (B) the ground truth; (C) baseline; (D) adaptive weight + pixel-wise guidance; (E) adaptive weight + pixel-wise guidance + boundary enhancement.

Ablation study A comparison of the number of parameters and float operations between these mentioned models is provided in Table 3. Although FMNet possesses neither the least parameters

nor float operations, the introduction of the feature mixer does not bring a significant increase in the number of either parameter or float. FMNet attains a considerable performance upgrade with an affordable increase in computational cost.

Table 4. Ablation experiment on FMNet to evaluate each component's utility

Adaptive weight	Pixel-wise guidance	Boundary enhancement	JA(%)
			65.58±1.96
✓			70.89±1.47
✓	✓		73.43±1.19
✓	✓	✓	**75.69 ± 2.15**

To illustrate the effectiveness of three key modules of FMNet, this section shows the result of an ablation experiment conducted also on the BUSI [1] dataset. The baseline method in this comparison is the original UNet [30]. Adaptive weight, pixel-wise guidance and boundary enhancement are added step-by-step into the baseline in order to demonstrate each component's significance in the final result. Fig.6 displays the qualitative result of the ablation experiment based on three random samples from the BUSI [1] dataset. Table 4 presents the outcomes of the ablation experiment. All the proposed components have a positive effect on segmentation result, and it can be clearly seen that addressing the semantic gap between multi-level information brings a considerable quantitative advantage in segmentation quality. The boundary detection module also provides a positive effect on the performance of FMNet. The combination of the three components overall results in an improvement of 10.11% mIoU.

5 Conclusion

In this work, an end-to-end DCNN-based deep learning network is presented for breast cancer segmentation. With the help of the proposed feature mixer, the semantic gap between high-level and low-level representations is effectively addressed. Specifically, this module automatically adjusts the weight assigned to different-level feature maps to maximize the introduction of complementary information. In addition, supervised lesion boundary information is adaptively aggregated into the feature mixer to enhance FMNet's segmentation accuracy in tumor boundaries. It is proven that firm improvements are brought by the introduction of the feature mixer. Exhaustive experiments also illustrate that FMNet surpasses previous state-of-the-art results both qualitatively and quantitatively on the BUSI [1] dataset.

References

1. Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.
2. Md Amirul Islam, Mrigank Rochan, Neil DB Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3751–3759, 2017.

3. Bahareh Behboodi, Mina Amiri, Rupert Brooks, and Hassan Rivaz. Breast lesion segmentation in ultrasound images with limited annotated data. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1834–1837. IEEE, 2020.
4. John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.
5. Jaime S Cardoso, Nuno Marques, Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. Mass segmentation in mammograms: A cross-sensor comparison of deep and tailored features. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1737–1741. IEEE, 2017.
6. Yusuf Celik, Muhammed Talo, Ozal Yildirim, Murat Karabatak, and U Rajendra Acharya. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. *Pattern Recognition Letters*, 133:232–239, 2020.
7. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
8. Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. Deep structured learning for mass segmentation from mammograms. In *2015 IEEE international conference on image processing (ICIP)*, pages 2950–2954. IEEE, 2015.
9. Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2393–2402, 2018.
10. Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, Donald M Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. Cancer statistics for the year 2020: An overview. *International journal of cancer*, 149(4):778–789, 2021.
11. Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.
12. Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.
13. Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018.
14. Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European conference on computer vision*, pages 519–534. Springer, 2016.
15. Mourad Gridach. Pydinet: Pyramid dilated network for medical image segmentation. *Neural Networks*, 140:274–281, 2021.
16. Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.
17. Rafika Harrabi and Ezzedine Ben Braiek. Color image segmentation using multi-level thresholding approach and data fusion techniques: application in the breast cancer cells images. *EURASIP Journal on Image and Video Processing*, 2012(1):1–11, 2012.
18. Xin Li, Feng Xu, Xin Lyu, Hongmin Gao, Yao Tong, Sujin Cai, Shengyang Li, and Daofang Liu. Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images. *International Journal of Remote Sensing*, 42(9):3583–3610, 2021.
19. Yang Li, Yue Zhang, Weigang Cui, Baiying Lei, Xihe Kuang, and Teng Zhang. Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation. *IEEE Transactions on Medical Imaging*, 2022.
20. Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep learning in medical ultrasound analysis: a review. *Engineering*, 5(2):261–275, 2019.

21. Shuo Liu, Wenrui Ding, Chunhui Liu, Yu Liu, Yufeng Wang, and Hongguang Li. Ern: Edge loss reinforced semantic segmentation network for remote sensing images. *Remote Sensing*, 10(9):1339, 2018.
22. Epimack Michael, He Ma, Hong Li, Frank Kulwa, and Jing Li. Breast cancer segmentation methods: current status and future potentials. *BioMed Research International*, 2021, 2021.
23. Muhammad Muhammad, Diyar Zeebaree, Adnan Mohsin Abdulazeez Brifcani, Jwan Saeed, and Dilovan Asaad Zebari. Region of interest segmentation based on clustering techniques for breast cancer ultrasound images: A review. *Journal of Applied Science and Technology Trends*, 1(3):78–91, 2020.
24. Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.
25. Arnau Oliver, Meritxell Tortajada, Xavier Lladó, Jordi Freixenet, Sergi Ganau, Lidia Tortajada, Mariona Vilagran, Melcior Sentís, and Robert Martí. Breast density analysis using an automatic density segmentation algorithm. *Journal of digital imaging*, 28(5):604–612, 2015.
26. Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.
27. Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017.
28. S Punitha, A Amuthan, and K Suresh Joseph. Benign and malignant breast cancer segmentation using optimized region growing technique. *Future Computing and Informatics Journal*, 3(2):348–358, 2018.
29. Hairong Qi, Wesley E Snyder, Jonathan F Head, and Robert L Elliott. Detecting breast cancer from infrared images by asymmetry analysis. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No. 00CH37143)*, volume 2, pages 1227–1228. IEEE, 2000.
30. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
31. Wolfgang Arthur Schulz. Molecular biology of human cancers: an advanced student's textbook. 2005.
32. B Senthilkumar, G Umamaheswari, and J Karthik. A novel region growing segmentation algorithm for the detection of breast cancer. In *2010 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–4. IEEE, 2010.
33. Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.
34. Fan Wu, Haiqiong Yang, Linlin Peng, Zongkai Lian, Mingxin Li, Gang Qu, Shancheng Jiang, and Yu Han. Agnet: Automatic generation network for skin imaging reports. *Computers in Biology and Medicine*, 141:105037, 2022.
35. Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 327–331. IEEE, 2018.
36. Dawei Yang, Yan Du, Hongli Yao, and Liyan Bao. Image semantic segmentation with hierarchical feature fusion based on deep neural network. *Connection Science*, 34(1):1772–1784, 2022.
37. Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.
38. Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1857–1866, 2018.
39. Boxiang Zhang, Wenhui Li, Yuming Hui, Jiayun Liu, and Yuanyuan Guan. Mfenet: Multi-level feature enhancement network for real-time semantic segmentation. *Neurocomputing*, 393:54–65, 2020.

40. Zhijie Zhang and Yanwei Pang. Cgnet: cross-guidance network for semantic segmentation. *Science China Information Sciences*, 63(2):1–16, 2020.
41. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
42. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

致谢信

自从我开始学习各种编程语言，我便对计算机科学产生了浓厚的兴趣。随着年龄的增长，技术也有了些许积累。在 Nadine 老师的鼓励下，我开始阅读各类技术博客和书籍，并且发现了我对深度学习、医学影像的图像处理领域的兴趣。同时，我也关注到了目前乳腺癌对全球女性健康带来的严重危害，便准备基于深度学习相关知识，查阅相关文献，提出一种方法以提升乳腺癌分割的准确率。

本项目中，我独立完成了模型设计、论文编写和各种实验。感谢 Nadine 老师对我的积极鼓励和引导，给予了我在探索途中最重 要的信心和勇气。她帮助解答了在论文写作产生的种种疑惑，让我在学习中也增进了自身的技术水平，同时也让我对这门技术的兴趣 更加浓厚。感谢在我学习之路上提供帮助的家人和朋友们。

最后，我还要感谢丘成桐中学科学奖，给予了我一个展示自己技术和能力的平台，能够激励我一次又一次地实现自我进步。这次探索的过程，是一次科学研究，也是一次历练，让我对编程语言有了许多新的理解和感悟，同时也加深了我对深度学习这门技术的理解程度，给予了我很多收获。