IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 23, 2021, accepted September 29, 2021, date of publication October 4, 2021, date of current version October 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3117578

Multi-Scale Fusion U-Net for the Segmentation of Breast Lesions

JINGYAO LI[1], LIANGLUN CHENG[2], TINGJIAN XIA[2], HAOMIN NI[1], AND JIAO LI[3]

[1]School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[2]School of Computers, Guangdong University of Technology, Guangzhou 510006, China
[3]State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Department of Medical Imaging, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

Corresponding author: Jiao Li (lijiao@sysucc.org.cn)

This work was supported in part by the Research and Development Project in Key Areas of Guangdong Province of China under Grant 2018B010109007 and Grant 2019B010109001, and in part by Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2020B1212060069.

ABSTRACT Breast lesion is a malignant tumor that occurs in the epithelial tissue of the breast. The early detection of breast lesions can make patients for treatment and improve survival rate. Thus, the accurate and automatic segmentation of breast lesions from ultrasound images is a fundamental task. However, the effectively segmentation of breast lesions is still faced with two challenges. One is the characteristics of breast lesions' multi-scale and the other one is blurred edges making segmentation difficult. To solve these problems, we propose a deep learning architecture, named Multi-scale Fusion U-Net (MF U-Net), which extracts the texture features and edge features of the image. It includes two novel modules and a new focal loss: 1) the Fusion Module (WFM) which segmenting irregular and fuzzy breast lesions, 2) the Multi-Scale Dilated Convolutions Module (MDCM) which overcoming the segmentation difficulties caused by large-scale changes in breast lesions, and 3) focal-DSC loss is proposed to solve the class imbalance problems in breast lesions segmentation. Moreover, there are some convolutional layers with different receptive fields in MDCM, which improves the network's ability to extract multi-scale features. Comparative experiments reveal that the MF U-Net proposed in this paper outperforms other segmentation methods, and the proposed MF U-Net achieves state-of-the-art breast lesions segmentation results with 0.9421 Recall, 0.9345 Precision, 0.0694 FPs/image, 0.9535 DSC and 0.9112 IOU on Benchmark for Breast Ultrasound Image Segmentation (BUSIS) dataset.

INDEX TERMS Breast cancer, deep learning, image segmentation, multi-scale feature, wavelet transform.

I. INTRODUCTION

Breast lesion is a malignant tumor that occurs in the epithelial tissue of the breast, and its incidence ranks the first among Chinese women [1]. It has the highest mortality rate compared to other types of cancer. By 2020, more than 1.9 million women will die from breast cancer each year. Breast cancer is not only the most frequently diagnosed cancer in most countries, but also the leading cause of cancer deaths in more than 100 countries [2]. Studies have shown that early detection of breast lesions can prompt patients to be treated and improve survival rates [3], [4]. However, the detection of breast lesions requires an experienced and well-trained radiologist. Even a trained specialist may have a high inter-observer variation rate on detection of breast lesions. Therefore, automatic detection or segmentation of breast lesion is very important. Currently, the methods used in breast lesions screening include X-ray, Ultrasound, and Magnetic Resonance Imaging. In this paper, ultrasound images are selected as the research object because of their versatility, safety and high sensitivity [5].

Many traditional methods have been proposed to detect breast lesions. Among them, Drukker *et al.* used Radial Gradient Index (RGI) filtering to automatically detect breast lesions and segmented it by maximizing an average radial gradient [6]. Yap *et al.* first performed histogram equalization on the image, and then used threshing segmentation and a rule-based approach to detect breast lesions [7]. Shan *et al.* proposed a combination of phase in max-energy orientation and radial distance with a traditional intensity-and-texture feature to distinguish breast lesions [8]. The Deform-able Part Models method proposed by Felzenszwalb *et al.* defines the low-resolution root filter template of the detection window and a set of high-resolution partial filter templates to capture

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan.

details [9]. Then the performance of the object was modeled based on the directional gradient histogram. However, these traditional methods are based on hand-crafted features, such as texture and Gab-or filters, which are not robust to pathological regions and are susceptible to image quality.

Since some methods based on Convolutional Neural Network (CNN) have shown good performance in image classification, scholars have begun to introduce deep learning to breast lesions detection and segmentation. There are three kinds of deep learning methods for breast lesions detection and segmentation: Patch-based CNNs approaches [10], [11], fully convolutional network based approaches [12] and transfer learning approaches [13], [14]. Ciresan *et al.* mapped each pixel-centered window to a neuron, and then extracted features with increasing levels of abstraction by a series of convolutional layers and maximum pooling layer [10]. Kooi *et al.* first applied the candidate detection pipeline to determine the five CNN seed points [11]. Then the convolutional layer and the maximum pooling layer were also used to extract features, and finally the category of the region was output. However, the patch-based CNNs method divides the ultrasound image into multiple patches, which will destroy the integrity of the spatial information of breast lesions. To ensure that the integrity of the spatial information of the lesion is not compromised, the U-Net was introduced into the segmentation of breast lesions by Yap et al, which is composed of a down-sampling path and an up-sampling path [12], [15]. The down-sampling path gradually reduces the size of the feature map and extracts low-resolution information, which provides a basis for lesion identification. The symmetrical up-sampling path gradually restores the smaller features to the same size as the original image and obtains high-resolution information, providing a basis for accurate lesion segmentation. However, the local receptive field and the efficiency of feature re-use are limited by the fixed convolution size and single down-sampling path in U-Net, which may not be conducive to dealing with the problems of large changes in breast lesion scale and blurred boundaries [16]. In addition, due to the scarcity of medical image data, transfer learning has been applied to the field of medical image processing by many scholars. The non-medical image data is used for the pre-training of the model, and then the pre-trained model is applied to the medical image. However, this method may cause the model to overlearn some data distributions that are not related to the target dataset, resulting in a decrease in model performance [13], [14].

In the breast lesion dataset, there was a significant difference in the size of the lesion area for breast cancer in it. In addition, the irregularity of the breast lesion areas and the blurred boundaries will bring severe challenges to segmentation. The samples of breast ultrasound images and ground truth images in the Benchmark for Breast Ultrasound Image Segmentation (BUSIS) dataset are shown in Fig. 1. However, none of the above methods specifically extract the multi-scale features and edge features of breast lesions, which will cause segmentation difficulties. In order to handle with the above problems, Multi-scale Fusion U-Net (MF U-Net) is proposed in this paper. On the basis of U-Net, the Wavelet Fusion Module (WFM) and the Multi-Scale Dilated Convolutions Module (MDCM) are integrated into MF U-Net. In WFM, the segmentation capability of U-Net for fuzzy edges is enhanced by multi-dimensional information fusion. In addition, the MDCM makes the network scale-invariant by designing a convolutional layer with multiple receptive fields, while keeping the model parameters constant. The Focal-DSC loss is proposed to learn class distribution to alleviate the problem of unbalanced voxels. In summary, the main contributions in the proposed method include:

1) In order to effectively segment breast lesions, we propose a fully automated method called Multi-scale Fusion U-Net (MF U-Net). MF U-Net is an end-to-end deep learning network, which extracts the texture features and edge features of the image, and finally assigns a category to each pixel.;
2) For improving the segmentation ability of irregular and fuzzy breast lesions, Wavelet Fusion Module (WFM) is proposed. In WFM, the image is decomposed into some detailed images which carry a lot of edge information. Then outputs of WFM are fused with pooling layers of U-Net, making the network more sensitive to the edge.
3) Moreover, to overcoming the segmentation difficulties caused by large-scale changes in breast lesions, a Multi-Scale Dilated Convolutions Module (MDCM) is proposed. The MDCM makes the network scale-invariant by designing a convolutional layer with multiple receptive fields, while keeping the model parameters constant.
4) Taking into account the class imbalance problems in breast lesions segmentation, we propose a new mixed loss function, Focal-DSC loss, which can effectively improve the contribution of difficult-to-divide targets in model optimization while suppressing the contribution of easily-divided targets.
5) On the BUSIS dataset, we conducted ablation experiments on the modules proposed in this paper, and the experiments proved the effectiveness of the two modules. Our MF U-Net achieves the best segmentation performance. The experimental results show that our MF U-Net is sensitive to edges and scale-invariant, so it can effectively extract breast lesion features and segment breast lesions accurately. At the same time, we prove that this method is superior to the current mainstream methods through comparative experiments.

II. RELATED WORKS

A. EDGE FEATURE REPRESENTATION IN CONVOLUTIONAL NEURAL NETWORK

It is difficult to segment some breast lesions due to their blurred and irregular edges. At the same time, Texture features in medical images are easily extracted by convolutional neural networks(CNN), but edge features are also easily

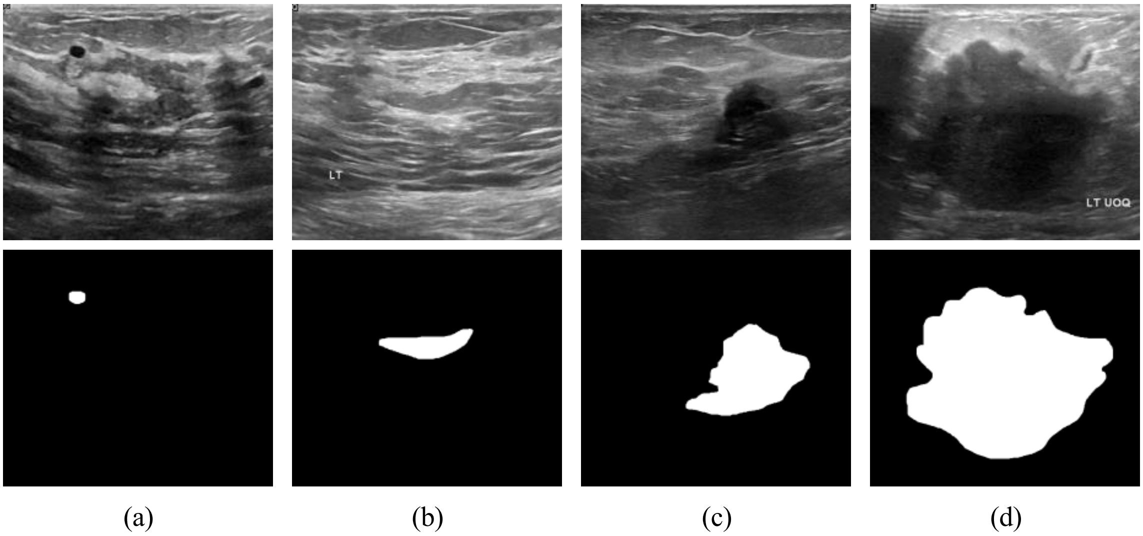


FIGURE 1. Samples of ultrasound breast images and ground truth images in the BUSIS dataset. Among them, both (a) and (b) are representative images of patients with early stage breast cancer, while both (c) and (d) are representative images of patients with advanced breast cancer.

ignored by them, which will lead to the poor segmentation effect of the model. [17]. Leon *et al.* found that texture features led to a substantial advance in image synthesis and manipulation in computer vision using CNNs [18]. Their work proved the dependence of CNN on texture features. In addition, Brendel *et al.* proposed a variant of Res-Net, which is based on the occurrences of small local image features without taking into account their spatial ordering [19]. This experiment shows that texture features are more concerned with CNN than edge features. Moreover, Gatys *et al.* found that the linear classifier on the CNN texture representation does not have much classification performance loss compared with the original network [20]. Besides, Hosseini *et al.* designed experiments to verify that different CNNs achieve similar accuracy on original images, but perform significantly different on negative images [21]. This showed that CNN prefers to classify objects based on the color rather than the shape. To sum up, the above literature shows that CNN is relatively inadequate in the recognition of edge features compared to texture feature extraction.

Inspired by the above-mentioned articles that study feature fusion [22], [23], we design a Wavelet Fusion Module (WFM) to improve the segmentation ability of irregular and fuzzy breast lesions. The original image is decomposed in WFM to get detailed images which fuse with U-Net after convolution, so that the network is more sensitive to the edge.

B. MULTI-SCALE FEATURE EXTRACTION

The large-scale changes and morphological problems of breast lesions in medical images will cause great difficulties in segmentation of breast lesions. Therefore, it is particularly important to propose a semantic segmentation algorithm with scale invariance. In order to achieve accurate detection and segmentation of multi-scale targets, many solutions have been proposed. Yang *et al.* make independent predictions on different resolution layers to ensure that small objects are trained on the high-resolution layer, while large objects are trained on the low-resolution layer [24]. Lin *et al.* proposed a top-down architecture with lateral connections for building high-level semantic feature maps at all scales [25]. He *et al.* also used pyramid representation, combining shallow and deep features to detect targets of different sizes [26]. Zhao *et al.* proposed an image cascade network, which incorporates multi-resolution branches and introduces the cascade feature fusion unit to quickly achieve high-quality segmentation [27]. He *et al.* proposed a Dynamic Multi-scale Network (DMNet) to adaptively capture multi-scale content to predict segmentation [28]. DMNet is composed of multiple parallel dynamic convolution modules, and context-aware filters are used in each module to estimate the semantic representation at a specific scale.

However, in the above-mentioned methods, the multi-scale features are extracted in layer-wisly, leading to a more complex computational process. Li *et al.* introduced a Dilated-inception Net (DIN) to extract and aggregate multi-scaigle features for right ventricular segmentation [29]. In the benchmark database of right ventricular segmentation challenges, DIN outperformed most advanced models. Liu *et al.* designed a deep neural network architecture called multi-scale deep fusion network (MSDF-Net), which uses Atrous Spatial Pyramid Pooling (ASPP) for feature extraction at different scales, and adds a capsule for processing complex relative entities [30]. Qi *et al.* proposed a network model called X-Net. X-Net used depthwise separable convolution instead of U-Net convolution operations, considering the effectiveness of it in reducing the parameters of the convolution kernel [31]. Inspired by these articles, we design a multi-scale feature module named Multi-Scale Dilated Convolutions Module (MDCM). The MDCM contains convolution kernels of different receptive fields, which can extract multi-scale features.

C. CLASS IMBALANCE IN MEDICAL IMAGE SEGMENTATION

Small object segmentation is always a challenge in semantic segmentation [32]. From a learning point of view, the challenge is caused by unbalanced data distribution, because image semantic segmentation requires pixel-by-pixel labeling, and small-volume organs contribute less to the loss. In this case, careful selection of the loss function is crucial. Havaei *et al.* used a sampling rule to make the foreground or background pixels have equal probability in the center of the patch, and used cross-entropy loss optimization [33]. Recently, A introduced log-cosh Dice loss, and compared 15 loss functions using the NBFS Skull-stripping dataset(brain CT segmentation) [34], found that Focal Tversky loss and Tversky loss are generally optimal [35]. Michael *et al.* experiment with seven different Dice-based and cross entropy-based loss functions on the Kidney Tumour Segmentation 2019 (KiTS19) dataset [36] and propose a Mixed Focal loss function, which is robust to significant class imbalance [37]. Zhang *et al.* state the Effective Example Mining(EEM) problem and propose a regression version of focal loss to make the regression process focus on high-quality sample [38].

We summarize the knowledge provided by previous research and propose a new mixed loss function, Focal-DSC loss, which can control the contribution of positive and negative targets. The DSC-part of Focal-DSC loss learning class distribution alleviates the problem of unbalanced voxels, while the Focal-part forces the model to better learn poorly classified voxels.

III. METHODS

To make the most of the characteristics of the experimental dataset, we performed a pixel-level statistical analysis of the size of the region of interest (ROI) of the lesions in the breast lesion dataset and concluded that the shape and size of breast lesions in different periods vary greatly. In order to reduce the segmentation difficulties caused by the large variation in lesion size, we propose the Multi-Scale Dilated Convolutions Module (MDCM) to achieve multi-scale feature extraction of breast lesions. Also, to better segment irregular edges and blurred breast lesions, we propose the Wavelet Fusion Module (WFM) to extract edge features. In addition, to alleviate the class imbalance problem in lesion segmentation, we propose a new loss function (Focal-DSC loss). In the subsequent sections, we will perform quantitative ablation experiments and comparison experiments on the different modules.

A. MULTI-SCALE FUSION U-NET FOR THE SEGMENTATION OF BREAST LESIONS

Inspired by U-Net [15] and X-Net [31], we design an Encoder-Decoder network as the main body for our MF U-Net. Different from the traditional structure of U-Net, we propose a Multi-Scale Dilated Convolutions Module (MDCM) to replace general convolution. On the breast lesions dataset, the shape and size of breast lesions in different periods are different, so there is a large difference between the segmentation target sizes. MDCM, which can extract multi-scale features, is more suitable for addressing these issues. Moreover, for better segmenting irregular edges and blurry breast lesions, we propose the Wavelet Fusion Module (WFM) to extract edge features.MDCM and WFM will be introduced in detail in the next two sections.

In MF U-Net, the input image is fed into the main encoder path with three MDCMs for multi-scale feature extraction. MDCM contains many 3×3 convolutional layers and 1×1 convolutional layers to generate a series of encoder feature maps to achieve multi-scale feature extraction and fusion. After each convolutional layer, the Rectified Linear Unit (ReLU) [39] is used as an activation function to improve the non-linearity of the network. Subsequently, the maximum pool layer is used to downsample in the encoder path, reducing the computation of the upper layers and the spatial complexity of the model by eliminating non-maximum values. At the same time, the input image is fed into the sub-encoder path with three WFMs for wavelet transform and convolution operations. The generated feature maps will be fused with the output feature maps of the main encoder paths to enhance the edge features of the image, as shown on the left side of Fig. 2. Correspondingly, the decoder path uses deconvolution to sample the feature maps, aiming to gradually restore the smaller size feature maps to a predicted map of the same size as the original input image. Importantly, the feature maps output by the same level of convolution and deconvolution layers are concatenated by skip connections. This operation enables the model to capture global features and local features at the same time and alleviate the degradation of the neural network, thus improving the accuracy of the model in segmenting breast lesions [40]. The network structure of MF U-Net is shown in Fig. 2.

B. WAVELET FUSION MODULE

In breast lesions segmentation tasks with fuzzy edges, it is important for the model to capture edge information to improve accuracy. In order to solve the problem that the breast lesions edge is difficult to be accurately identified in the segmentation, Wavelet Fusion Module (WFM) is proposed in this paper.

Our WFM is implemented based on Wavelet Packet Transform (WPT) and the modulus maximum edge detection method, which decomposes an image into a sequence of wavelet coefficients of the same size. The edge information of an image is represented in the image as a discontinuity of the signal, corresponding to the high frequency part of the image. According to the renowned scholar Mallat *et al.* [41], the local modulus maximum of the wavelet transform corresponds to the abrupt change points of the signal, i.e. the edge information of the image. Haar wavelet is selected in this paper, because it is enough to depict different-frequency breast lesions information. In 2D Haar wavelet, four filters

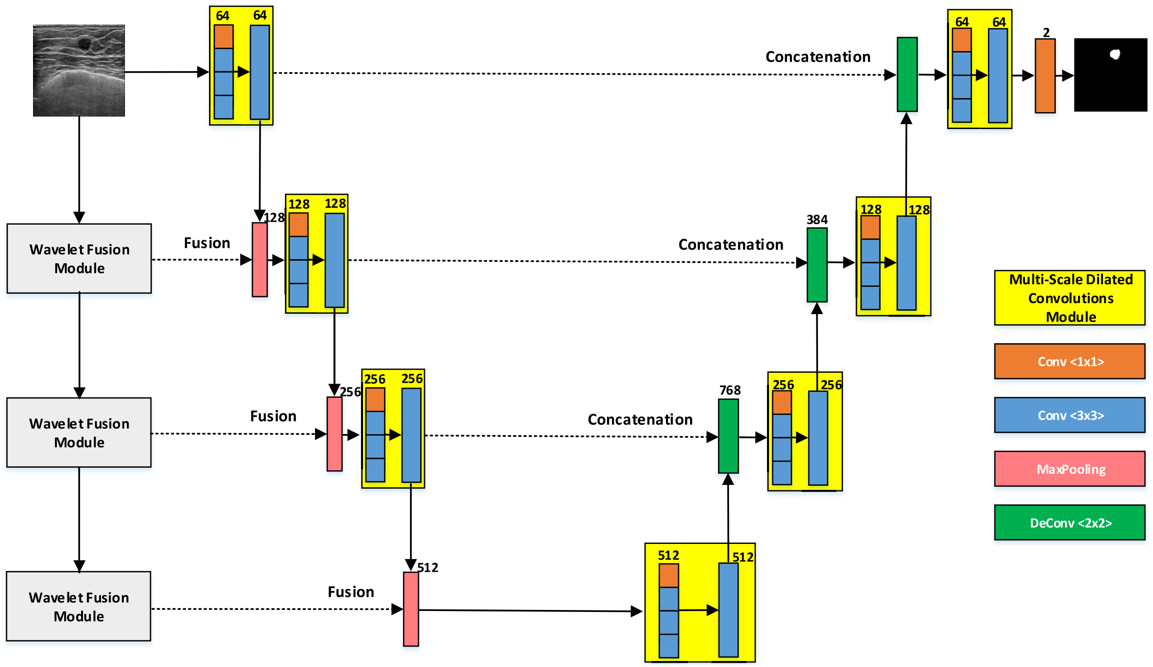


FIGURE 2. Illustration of our MF U-Net: The structure of MF U-Net is an Encoder-Decoder network, in which we use a Multi-Scale Dilated Convolutions Module instead of a general convolution block. The input image is processed by the Wavelet Fusion Module and concatenated with the pooling layer of each stage.

are defined as follows:

$$f_{LL} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} f_{LH} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} \tag{1}$$

$$f_{HL} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} f_{HH} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{2}$$

Suppose $I(x, y)$ is the pixel of the ultrasound image at (i, j), the value of approximation image after 2D Haar transform can be calculated as:

$$\begin{aligned} I_{LL}(i, j) = I(2i-1, 2j-1) + I(2i-1, 2j) \\ + I(2i, 2j-1) + I(2i, 2j) \end{aligned} \tag{3}$$

Analogously, $I_{LH}(i, j)$, $I_{HL}(i, j)$, $I_{HH}(i, j)$ can be calculated. After completing the wavelet transform, the modulus and magnitude of each pixel in the image will be calculated by the following equations (3) and (4) respectively:

$$Mf(x, y) = \sqrt{|W^x f(x, y)|^2 + |W^y f(x, y)|^2} \tag{4}$$

$$Af(x, y) = arctg\left[\frac{W^x f(x, y)}{W^y f(x, y)}\right] \tag{5}$$

where $W^x f(x, y)$ and $W^y f(x, y)$ denote the components of the 2D wavelet transform functions $\psi^x(x, y)$ and $\psi^y(x, y)$, respectively, which are defined as follows:

$$\begin{cases} W^x f(x, y) = f * \psi_s^x(x, y) \\ W^y f(x, y) = f * \psi_s^y(x, y) \end{cases} \tag{6}$$

where s is the scale factor and in this paper we take $s = 2^j j \in Z$.Correspondingly, the 2D wavelet transform functions $\psi^x(x, y)$ and $\psi^y(x, y)$ are obtained by taking the partial derivatives of the smoothing function $\theta(x, y)$ in the x and y directions, calculated as follows:

$$\begin{cases} \psi^x = \dfrac{\partial \theta(x, y)}{\partial x} \\ \psi^y = \dfrac{\partial \theta(x, y)}{\partial y} \end{cases} \tag{7}$$

where the smoothing function $\theta(x, y)$ should satisfy the following conditions:

$$\iint_R \theta(x, y)\, dxdy = 1 \tag{8}$$

For visualization purposes, we normalize the wavelet coefficients of each sub-image to the range of [0, 255] and the results of edge detection are shown in Fig. 3. In Fig. 3, the approximation image is the low-frequency information of the entire pathology map, including the rough texture of the breast. $I_{LL}(i, j)$, $I_{HL}(i, j)$ and $I_{HH}(i, j)$ are detailed images, which contain high-frequency mammary gland information after orthogonal decomposition. This high frequency information represents exactly the edge features of the image, which is very important for CNN that loses accurate edge information.

In WFM, we hope that the feature map with the edge information and it generated by U-Net will be fused. There are usually two ways to fuse convolutional features: 1) concatenate the convolutional features of multiple inputs along the channel dimension, and then fuse the merged features with the next convolutional layer, 2) fuse convolutional features of multiple inputs directly by element fusion rules.

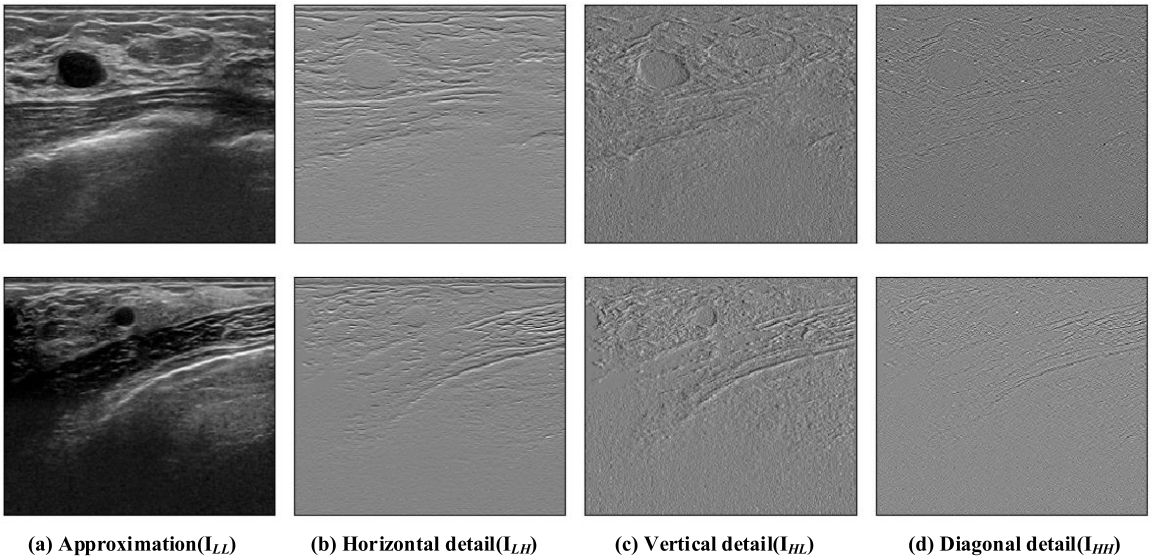

(a) Approximation(I_{LL})
(b) Horizontal detail(I_{LH})
(c) Vertical detail(I_{HL})
(d) Diagonal detail(I_{HH})

FIGURE 3. Wavelet decomposition results.

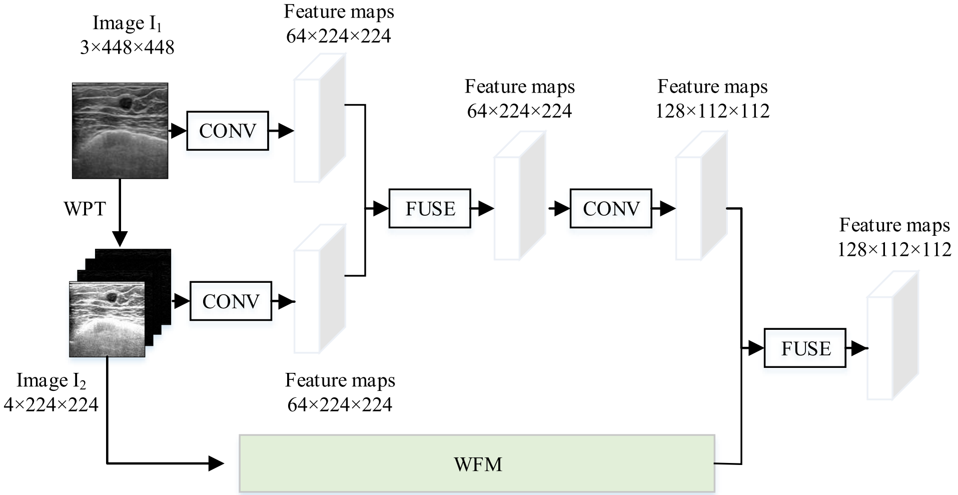


FIGURE 4. The structure of WFMs and their fusion with MF U-Net.

The skip structure of U-Net belongs to the former. In WFM, since the sharp features (maximum values) are expected to be preserved, we modify the maximum fusion rule for the fusion of the output feature map and the pooling layer of U-Net. The output fusion feature can be expressed as:

$$Z = h\left(W^T \left[B_{WFM}, B_{CNN}\right] + b\right) \tag{9}$$

where $h()$ represents the ReLU activation function, and W^T represents the connection weight. B is the characteristic diagram of different modules, and b is the offset value.

In WFM, the convolution operation of CNN is used to compute the detailed image with edge information produced by the wavelet transform, which is described as follows. The pixels in feature images are obtained by convolving detailed images with a kernel size of 3×3. In order to increase the nonlinearity of WFM, ReLU is used as an activation function after convolution. The output size of WFM is half of the input, and its dimensions are the same as the pooling layer of each stage. Finally, the outputs fuse with the feature map generated by U-Net to help the network extract more edge features. The structure of the WFMs and the process of their integration with the MF U-Net are shown in Fig. 4.

WFMs are integrated into the endcoding path of MF U-Net, which has the following properties: 1) avoiding the large growth of parameters by using only one convolutional layer in WFM; 2) decomposing the image to get high-frequency features to make the network more sensitive to edges; 3) supplementing the information lost by using the maximum pooling layer in down-sampling.

C. MULTI-SCALE DILATED CONVOLUTIONS MODULE

Large scale variation across breast lesions, especially the early breast lesions (very small targets) is an important factor that makes detection difficult. On the newly collected dataset, the median scale of breast lesions instances relative to the image is 0.012. To make matters worse, the scale of the smallest and largest 10% of breast lesions instances is 0.026 and 0.465 (resulting in scale variations of almost 18 times). The relationship between the fraction of RoIs (regions of interest) and the scale of RoIs is shown in Fig. 5.

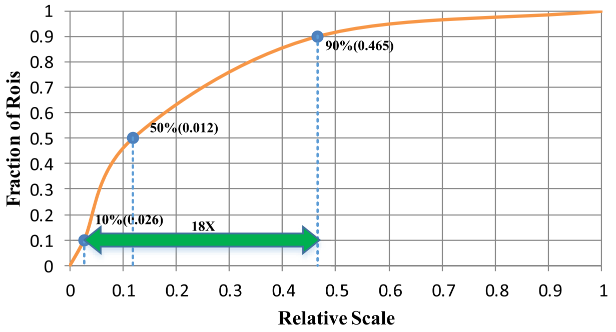


FIGURE 5. Fraction of RoIs vs scale of RoIs.

To address the problem of large scale variation in breast lesions, smaller and larger dilation rates can be used to capture information such as the texture of small and large targets respectively. Last but not least, the input features of the model are enhanced by fusing multiple feature maps at different scales to efficiently extract context information at different scales [42].

The receptive fields of ordinary convolution kernels are fixed, so they are not scale-invariant for features. When convolving the same morphological features at different scales, the similarity between the two features cannot be understood by the convolution kernel. In order for the network to segment targets of different scales in a network that is not scale-invariant, it is necessary to collect images of objects at different scales. But it is a difficult challenge for medical images, for which data is very scarce. At the same time, it is also very important to reduce model parameters in the training of small datasets to prevent over-fitting. To solve these problems, this paper proposes a convolutional module, called the Multi-Scale Diluted Convolutions Module (MDCM), which can still achieve multi-scale feature extraction without increasing the model parameters. The structure of MDCM is shown in Fig. 6.

After the feature map enters the MDCM, it is first convolved by the Compound Dilated Convolution Layer. The number of convolution kernels in the Compound Dilated Convolutional Layer is, and the convolution kernels are divided into four groups according to different convolution kernel forms. The first group is composed of convolution kernels with a size of 1×1, and the second to fourth groups are composed of 3×3 kernels at different dilation rates. The dilated convolution kernel is a special kind of convolution kernel, which can maintain relatively low parameters and calculations under the condition of large receptive field [43]. Besides, the size of the receptive field is controlled by different dilation rate, for a 3×3 convolution kernel, the calculation formula for its receptive field is as follows:

$$F = \left(2^{i+1} - 1\right) \times \left(2^{i+1} - 1\right) \tag{10}$$

where $i + 1$ is the dilation rate. If the dilation rate is 1, the dilated convolution is the same as the ordinary convolution. The dilation rates of the second to fourth groups in MDCM are set to 1, 2, and 3 respectively in this paper.

Four feature maps of the same size are generated after the Compound Dilated Convolution Layer. Due to the convolution of kernels with different receptive fields, features of different scales are retained in different feature maps, and they are converted into a feature map through the concatenation in the channel dimension. ReLU is used as the activation function of this layer. Then, the feature map through the activation function is convolved with convolution kernels with a size of 3×3, thereby fusing features of different scales. Finally, ReLU is used again to output the feature map of MDCM.

D. FOCAL-DSC LOSS FUNCTION

Firstly, both Dice and IOU are the most commonly used evaluation indicators in segmentation networks. The Sørense-Dice index is known as the Dice similarity coefficient (DSC). We can use True Positives (TP), False Positives (FP) and False Negatives (FN) to define DSC and IOU:

$$IOU = \frac{TP}{TP + FN + FP} \tag{11}$$

$$DSC = \frac{2TP}{2TP + FN + FP} \tag{12}$$

Combining the above two formulas, we can get:

$$IOU = \frac{DSC}{2 - DSC} \tag{13}$$

The Dice loss of segmentation can be written as follows:

$$L_{DSC} = 1 - DSC \tag{14}$$

Focal loss is a variant of the cross entropy loss function [44]. By controlling the weight of positive and negative samples and the weight of easy-classify and difficult-classify samples, it solves the problem of class imbalance. In order to obtain the focal loss function, we first abbreviate the binary cross entropy (BCE) used for binary classification as:

$$CE(p, y) = \begin{cases} -log(p), & \text{if } y = 1 \\ -log(1 - p), & \text{if otherwise} \end{cases} \tag{15}$$

We can simplify the BCE loss using the following formula.

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if otherwise} \end{cases} \tag{16}$$

Now, BCE loss is defined as follows:

$$L_{BCE_{(p,y)}} = CE(p_t) = -log(p_t) \tag{17}$$

To reduce the influence of negative samples, you can add a coefficient named α before the conventional loss function. Similar to Pt, when y=1, $\alpha_t = \alpha$; when y=otherwise,

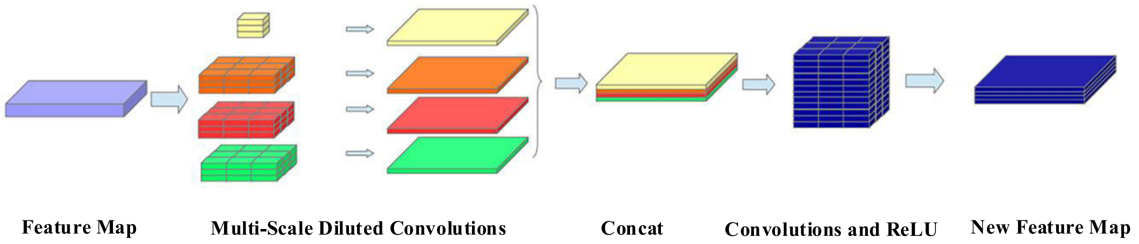


FIGURE 6. The structure of Multi-Scale Dilated Convolutions Module.

$\alpha_t = 1 - \alpha$, and the range of α is also 0 to 1. At this point, we can control the contribution of positive and negative samples to loss by setting α. When $\gamma = 0$, the Focal loss simplifies the BCE loss.

$$L_{Focal} = \alpha_t(1 - P_t)^{\gamma} \cdot L_{BCE_{(p,y)}} \tag{18}$$

Inspired by the Focal-EIoU loss [38] and the Enhanced mixing loss [45], we propose a new mixed loss function consisting of contributions from both Dice loss and Focal loss. To enable the Dice loss focus on high-quality examples, we use the value of IOU and L_{DSC} to replace α_t and $L_{BCE_{(p,y)}}$ in Eq.(8). The Focal-DSC loss can be formulated as:

$$L_{Focal-DSC} = IOU^{\gamma} \cdot L_{DSC} \tag{19}$$

where IOU and γ are a parameter to control the degree of inhibition of outliers. Using the Eq.(3) to replace IOU, we can define Focal-DSC loss as:

$$L_{Focal-DSC} = (\frac{DSC}{2 - DSC})^{\gamma} \cdot L_{DSC} \tag{20}$$

We also try different forms of reweighting process, like using DSC to replace IOU in Eq.(14). Now Focal-DSC* loss can be defined as:

$$L_{Focal-DSC*} = (DSC)^{\gamma} \cdot L_{DSC} \tag{21}$$

As shown in Fig. 7, we test different γ and forms of the Focal-DSC loss. We find that Eq.(15) with $\gamma = 0.5$ has the best performance.

IV. EVALUATION METRICS

In this section, we will describe the process of our experiments in detail and conduct an analysis according to the experiment results.

In the traditional methods, the seed point which is the location of breast lesions is detected first and then they perform segmentation accordingly. The detection accuracy of seed points is used to measure the quality of the algorithm. For the convenience of comparison, we also use the detection of seed points as the evaluation standard. Since our model is end-to-end, we define the segmented breast lesions center as the seed point. If a seed point is within the bounding box, it is called True Positive (TP); otherwise it is False Positive (FP). At the same time, if a seed point is not detected, then we call it False Negative (FN).

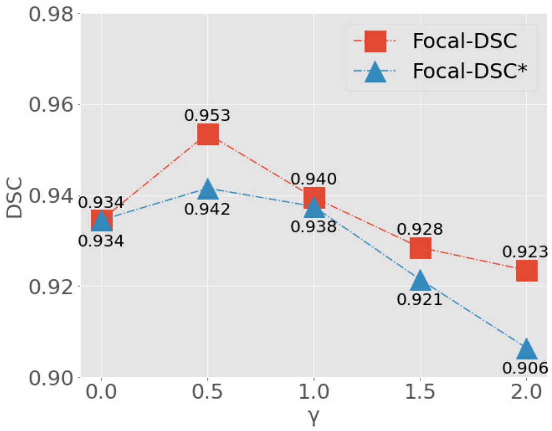


FIGURE 7. The Focal-DSC and Focal-DSC* loss with different γ.

To evaluate the performance of the algorithm, we use Recall, Precision, FPs/image (false positives per image), DSC (dice similariy coefficient) and IOU (intersection over union) to evaluate the performance of the network. We will use TP, FP and FN to calculate the above performance evaluation metrics, which are defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{22}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{23}$$

$$\text{FPs/image} = \frac{\text{number of FPs}}{\text{number of images}} \tag{24}$$

The sensitivity of the algorithm to seed points is measured by Recall. If all seed points can be detected, then Recall will be close to 1. And FPs/image measures the possibility of algorithm detection errors. Precision is used to measure the proportion of pixels that are actually positive out of those predicted to be positive. However, some algorithms that can detect multiple seed points will get a higher Recall, but the FPs/image also is higher. To better exploit the realistic significance of Recall and FPs/image, we will use the most commonly used performance evaluation metrics in segmentation networks - DSC and IOU. DSC and IOU both measure the similarity between two sets, and they are used to measure the similarity between network segmentation results and the gold standard mask, which is calculated as follows:

$$\mathrm{DSC} = \frac{2 \times \mathrm{TP}}{(2 \times \mathrm{TP}) + \mathrm{FP} + \mathrm{FN}} \tag{25}$$

$$\mathrm{IOU} = \frac{\mathrm{DSC}}{2 - \mathrm{DSC}} \tag{26}$$

The DSC and IOU are pixel-level and they take values between 0 and 1, with values closer to one indicating better performance of the model.

Furthermore, in order to be accurate to the extent that each pixel is correctly classified and each class is correctly segmented. PA(pixel accuracy), MPA(mean pixel accuracy), and MIOU(mean intersection over union) were used to evaluate the ratio of those to total pixels, the ratio of correctly classified pixels in each class, and the average IOU of each class, respectively. The calculation methods of PA, MPA and MIOU are as follows:

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{27}$$

$$\mathrm{MPA} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{28}$$

$$\mathrm{MIOU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{29}$$

where P_{ij} represents the pixel point in the i-th row and the j-th column of the image.

V. EXPERIMENTS

A. EXPERIMENTAL ENVIRONMENT AND IMPLEMENTATION DETAILS

The environment configuration in this paper is Ubuntu 16.04 operating system, 32GB DDR4 RAM and GeForce GTX 1080Ti graphics card. The programming language used for the algorithm is Python. training and testing were implemented on TensorFlow version 1.14.

In the training of the proposed model, we set the batchsize to 5, the epoch to 100, and the weight decay factor to 0.001. We also set the learning rate at 10-5. The loss function used is the Focal-DSC loss function, and stochastic gradient descent is used as an optimization algorithm during training.

B. DATASET AND DATA AUGMENTATION

The data used in this paper were collected from publicly available online datasets, ADEChallengeData 2016 and Benchmark for Breast Ultrasound Image Segmentation (BUSIS) respectively. Radiologists with extensive experience in breast ultrasound examined the collected data using an ACUSON S2000 (SIEMENS, Germany) system with a high transducer frequency (12-15 MHz). All our dataset is certified for pathology testing, which ensures accurate labelling.

The dataset contains over 1000 ultrasound images with an average image size of 495 × 412 pixels. In this paper it will be resized to 448 × 448. We perform data enhancement to increase the diversity of data available for training the model and to alleviate storage requirements. The data enhancement includes image rotation, brightness enhancement and contrast enhancement. After data enhancement, the number of images in the dataset grew to four times the original size. The dataset was then randomly divided into a training set and a test set in the ratio of 8:2.

C. ABLATION EXPERIMENT

In this section, we will conduct ablation experiments on the proposed wavelet fusion module (WFM), multi-scale dilution convolution module (MDCM) and loss function (Focal-DSC). The purpose of the experiment is to respectively prove the effectiveness of WFM, MDCM and Focal-DSC and the effectiveness of the combination of these three. Table 1 shows in detail the improvement effect after adding each function module in U-Net [15].

On the recently collected breast lesion dataset, the cross-entropy loss function is used to train U-Net. Subsequently, it was tested on the BUSIS dataset, and the segmentation performance of Recall (0.8913) and DSC (0.9304) were obtained respectively. In view of the good segmentation performance of U-Net, we use it as the baseline model for ablation experiments.

To demonstrate the effectiveness of the Focal-DSC loss function, it is used to train U-Net instead of the cross-entropy loss function. As expected, the model trained with the Focal-DSC loss function achieved better evaluation results on the test set, achieving Recall (0.9037) and DSC (0.9386), respectively. Compared to the baseline model (U-Net), Recall and DSC showed an improvement of 0.0124 and 0.0082 respectively, which would be attributed to the DSC penalty factor in the Focal-DSC loss function that enables the model to better measure the similarity between the predicted output and the mask.

In addition, in order to verify the effectiveness of WFM, WFM is embedded in U-Net to perform low-frequency and high-frequency filter decomposition and maximum edge detection on the image. As can be clearly seen in Table 1, the U-Net with WFM has significantly improved the Recall, Precision, DSC and IOU metrics obtained on the BUSIS dataset, which shows that WFM can help the model to obtain better segmentation performance. Because WFM can make the model decompose more high-frequency information and detailed information that helps extract the edge features of the image, and improve the sensitivity of the model to the edge of the lesion.

In order to validate the effectiveness of MDCM, MDCM was added to U-Net for multi-scale feature extraction. By using different dilation rates to increase the receptive field of the model, MDCM incorporates multiple feature maps of different scales to enhance the multi-scale feature extraction capability of the model. Compared with the baseline model, the U-Net with MDCM showed significant improvement in several evaluation metrics, achieving high scores of Recall (0.9216) and DSC (0.9467) respectively.

TABLE 1. Ablation experiments for quantitative analysis of the proposed WFM, MDCM and Focal-DSC.

Method	Recall	Precision	FPs/image	DSC	IOU	MIOU	PA	MPA
U-Net	0.8913	0.9015	0.1128	0.9304	0.8701	0.8813	0.9719	0.9553
U-Net + Focal-DSC	0.9037	0.9157	0.0921	0.9386	0.8844	0.8837	0.9804	0.9579
U-Net + WFM	0.9187	0.9018	0.0874	0.9421	0.8905	0.8905	0.9874	0.9617
U-Net + MDCM	0.9216	0.9158	0.0831	0.9467	0.8987	0.8987	0.9886	0.9602
U-Net + WFM + MDCM	0.9374	0.9268	0.0734	0.9526	0.9096	0.9096	0.9913	0.9623
U-Net + WFM + MDCM + Focal-DSC	**0.9421**	**0.9345**	**0.0694**	**0.9535**	**0.9112**	**0.9112**	**0.9927**	**0.9631**

TABLE 2. Performance comparisons of the different methods on the BUSIS collected dataset.

Method	Recall	Precision	FPs/image	DSC	IOU	MIOU	PA	MPA
RGI [6]	0.7528	0.7324	1.5027	0.4843	0.3256	0.3346	0.6236	0.5253
Multifractal [7]	0.5907	0.6254	0.5126	0.5674	0.3952	0.4075	0.6753	0.5471
RBRR [8]	0.6494	0.6325	0.5283	0.8106	0.4283	0.4448	0.8379	0.6784
DPM [9]	0.8117	0.7926	0.1925	0.8137	0.6819	0.7016	0.8621	0.7221
LeNet [46]	0.8458	0.8621	0.1348	0.8334	0.7184	0.7324	0.8973	0.8132
FCN [47]	0.8874	0.8764	0.1753	0.9027	0.8227	0.8487	0.9674	0.9627
LinkNet [48]	0.9053	0.8896	0.1324	0.9157	0.8444	0.8535	0.9692	0.9554
PSPNet [49]	0.9119	0.9002	0.1254	0.9254	0.8611	0.8611	0.9711	0.9511
U-Net [15]	0.8913	0.9015	0.1128	0.9304	0.8701	0.8813	0.9719	0.9553
SegNet [50]	0.9110	0.9145	0.1023	0.9323	0.8731	0.8731	0.9714	0.9484
DeepLabv3+ [51]	0.9215	0.9201	0.0846	0.9427	0.8916	0.8916	0.9825	**0.9694**
MF U-Net (ours)	**0.9421**	**0.9345**	**0.0694**	**0.9535**	**0.9112**	**0.9112**	**0.9927**	0.9631

Furthermore, we apply WFM and MDCM to U-Net at the same time to form the MF U-Net network structure proposed in this paper. The segmentation performance of MF U-Net in the test data has been further improved, and the performance of Recall (0.9374) and DSC (0.9526) have been obtained respectively. Last but not least, we use the most recommended combination in this article to experiment-use the Focal-DSC loss function instead of the cross-entropy loss function to train MF U-Net. It can be clearly seen from **Table 1** that the segmentation performance of the MF U-Net trained with the Focal-DSC loss function on the test data has been further improved, and the performance of Recall (0. 9421) and DSC (0. 9535) have been achieved respectively. Compared with the baseline model, its Recall and DSC accuracy are improved by 0.0508 and 0.0231 respectively.

In summary, the WFM, MDCM, and Focal-DSC proposed in this article all utilize model performance improvements. It should be pointed out that the combination of WFM, MDCM and Focal-DSC will bring a better improvement to the accuracy of the model.

D. COMPARATIVE EXPERIMENT

To verify the effectiveness of our method, we compare our MF U-Net with the several state-of-the-art breast lesions segmentation approaches: Radial Gradient Index (RGI) [6], Multifractal [7], Rule-Based Region Ranking (RBRR) [8], Deformable Part Models (DPM) [9], LeNet [46], Fully Convolutional Networks (FCN [47]), LinkNet [48], PSPNet [49], SegNet [50], DeepLabv3+ [51] and U-Shape Convolutional Network (U-Net) [28]. The comparison results of our method with other state-of-the-art are shown in Table 2.

As can be seen from Table 2, RGI, which uses an RGI filtering technique to filter lesions, achieves 0.7528 Recall, 0.7324 Precision, 1.5027 FPs/image and 0.4842 DSC. The FPs/image is higher than other methods because RGI produces more seed points. Multi fractal and RBRR perform poorly, because the manual features they extract are unsuitable for the dataset. DPM performed well, achieving high Recall and DSC. Neural networks can extract image features automatically and powerfully. LeNet achieves 0.8458 Recall, 0.8621 Precision, 0.1348 FPs/image and 0.8334 DSC. Deep learning can segment images correctly by learning abstract features of objects. In FCN, LinkNet, PSPNet, U-Net, SegNet and DeepLabv3+, the convolution layer and the deconvolution layer are connected through skip connection, so that global and local features can be captured at the same time. They all outperform traditional segmentation methods in terms of segmentation performance on the BUSIS dataset. Our MW-Net takes into account the scale and morphological issues of breast lesions, so that MW-Net can effectively segment breast lesions. MW-Net achieves 0.9421 Recall, 0.9421 Precision, 0.0694 FPs/image and 0.9535 DSC, surpassing all other methods.

Fig. 8 shows visual examples of the segmentation results of different algorithms. The first two rows show some easy-to-detect lesions, which have obvious features and are very different from the background, so all methods can correctly segment breast lesions. The third row shows a small breast lesion. Only RBRR and neural network can effectively detect lesions. The last row shows a challenging case for breast lesions segmentation, where the image is blurred and the texture and edges of the lesion are not obvious. For this case, none of the methods can segment the lesion correctly.

Fig. 9 shows the segmentation results of MF U-Net and other six segmentation algorithms for breast lesions. From top to bottom, the lesion area of the breast gradually becomes larger, and the scale varies widely. In addition, the boundary pixels of the breast lesion area are very

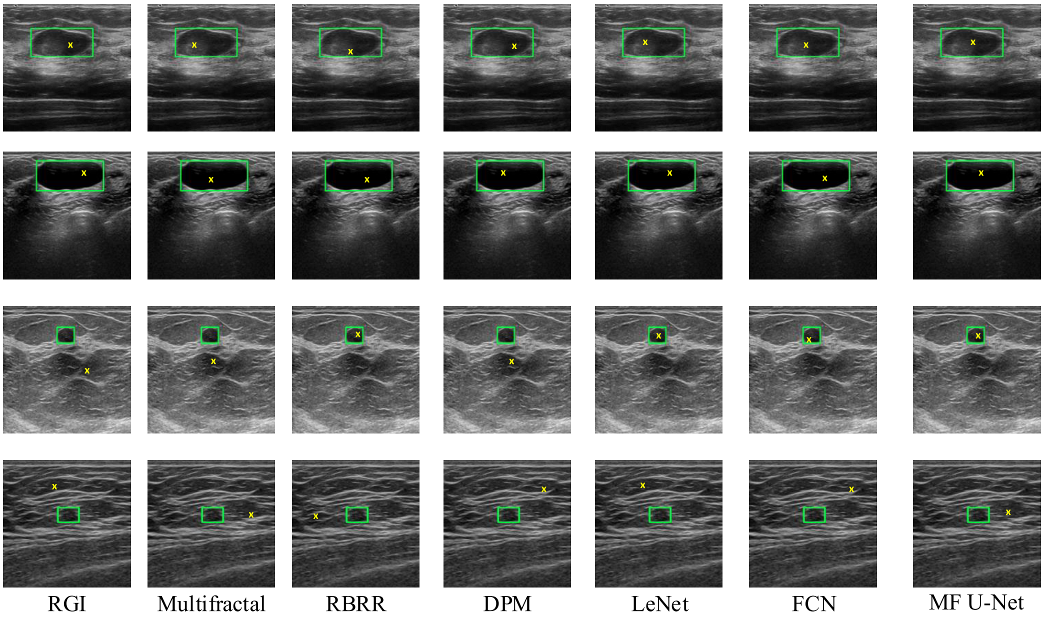


FIGURE 8. Comparison of MF U-Net results for breast lesions with the other six segmentation algorithms. In each image, the green border represents the breast lesion and the yellow dots represent the seed points.

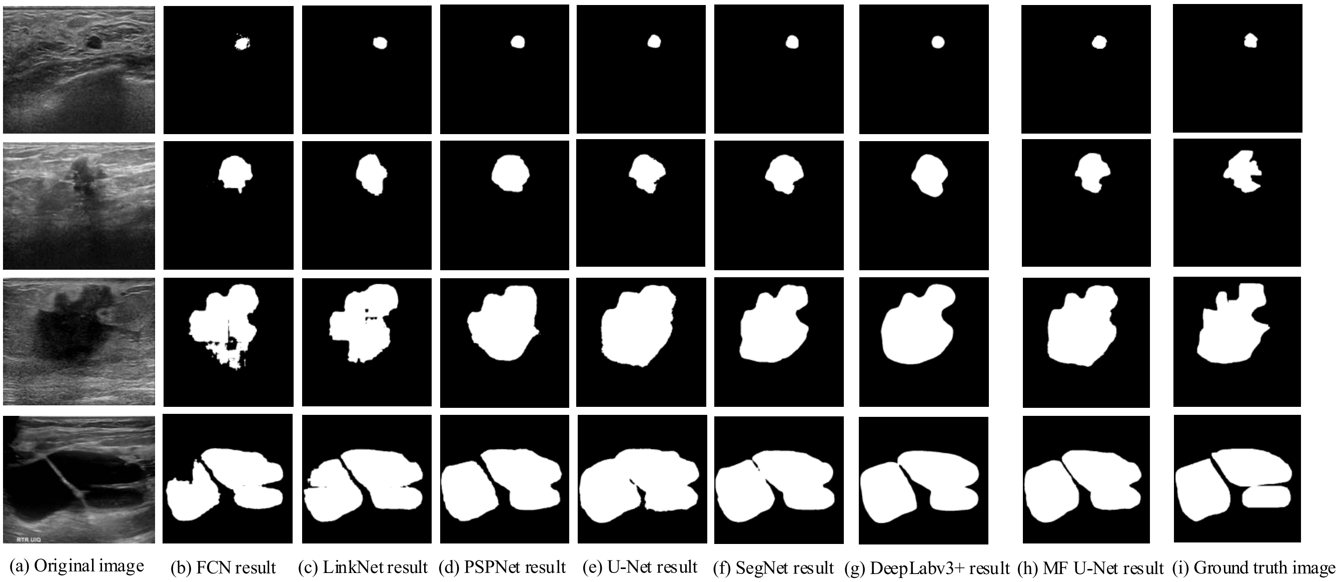
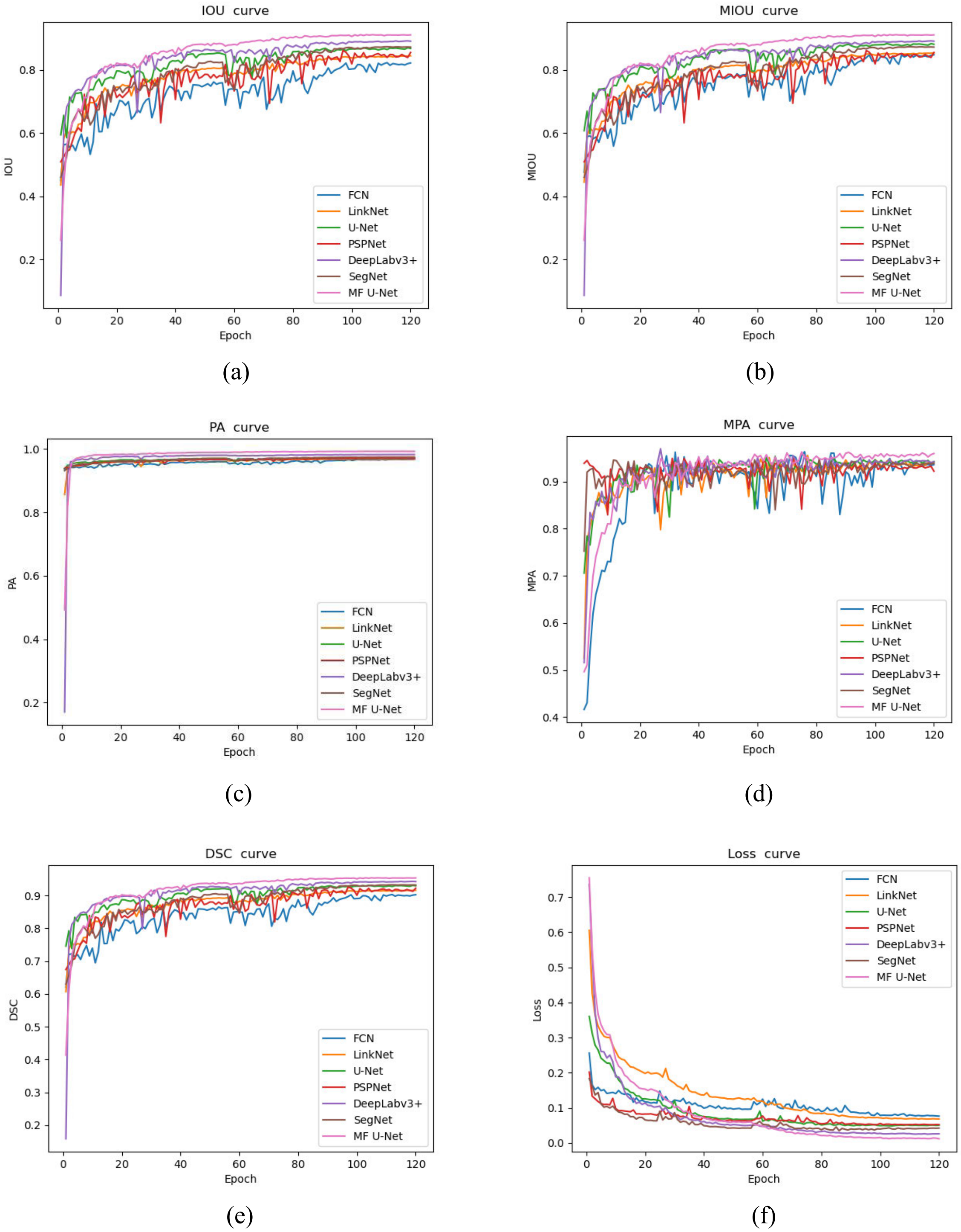


FIGURE 9. Comparison of segmentation results of MF U-Net and other segmentation algorithms for breast lesions. Each pixel in the label is assigned as either a background pixel or a lesion pixel. The lesion pixel is marked as 1 and the background pixel is marked as 0. Among them, (a) is the original image and (i) is the corresponding ground truth label. In addition, (b), (c), (d), (e), (f), (g) and (h) represent the segmentation results for FCN, LinkNet, PSPNet, U-Net, SegNet, DeepLabv3+ and MF U-Net, respectively.

blurred. These will cause difficulties for segmentation. From Fig.9(b), (c), (d), (e), (f), (g) and (h), it can be concluded that FCN and LinkNet are seriously insufficient in segmentation of the lesion, which is caused by insufficient up-sampling. In addition, due to the different scales of breast lesions, U-Net and SegNet cannot achieve accurate segmentation of large and small targets; on the contrary, thanks to MDCM, our method (MF U-Net) can accurately segment the lesions despite the different scales of the lesions. Although PSPNet and DeepLabv3+ can segment targets of different scales, they cannot achieve accurate segmentation due to their insufficient recognition of the edge of the lesion. Correspondingly, thanks to WFM, MF U-Net can accurately identify the edge information of the lesion and achieve precise segmentation. More importantly, in the method shown, only MF U-Net is able to efficiently learn the edge information of the lesion to achieve accurate segmentation of breast lesions, while the others are not.



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 10. Comparison of segmentation performance curves for FCN, LinkNet, U-Net and MF U-Net on the BUSIS dataset. Among them, (a), (b), (c), (d), (e) and (f) represent the IOU curve, MIOU curve, PA curve, MPA curve, DSC curve and Loss curve of the four segmentation algorithms respectively.

In order to compare the performance of the four segmentation algorithms more intuitively, the IOU curve, MIOU curve and DSC curve are used to measure the similarity between sets. In addition, PA and MPA are used to evaluate the ratio of correctly classified pixels to the total pixels, and the ratio of those in each category, which allows us to understand whether each pixel is correctly classified. Importantly, the Loss curve allows us to know more clearly when the model converges, which is beneficial to the training of the model. Therefore, the above curve is drawn to intuitively compare the segmentation performance between different algorithms.

Fig. 10 shows the comparison of the IOU curve, MIOU curve, PA curve, MPA curve, DSC curve and Loss curve of FCN, LinkNet, PSPNet, U-Net, SegNet, DeepLabv3+

and our proposed MF U-Net on the BUSIS dataset. From Fig. 10(a), (b) and (e), it can be concluded that the seven algorithms quickly achieved large values in about 20 iterations, and stabilized after about 100 iterations. Among them, MF U-Net achieved the maximum values in Fig.10(a), (b) and (e), with IOU, MIOU and DSC taking values of 0.9112, 0.9112, and 0.9535 respectively. It can be seen from Fig. 10(c) and (d) that compared with the other three algorithms, MF U-Net has achieved almost the best PA (0.9927) and MPA (0.9631) respectively, and its performance is stable. On the contrary, the PA and MPA indicators of the other three algorithms fluctuate widely. It can be seen from Fig. 10(f) that the loss value of the four algorithms decreases rapidly in the first 20 iterations, and stabilizes after 90 iterations, and the loss value of MF U-Net is the smallest.

Based on the above results, it can be concluded that MF U-Net performs better than the other six segmentation algorithms in terms of IOU, MIOU, PA, MPA, DSC and Loss, followed closely by DeepLabv3+, SegNet, U-Net. PSPNet and LinkNet respectively. FCN has the worst segmentation performance.

VI. DISCUSSION

In clinical diagnosis and treatment, the segmentation technology of medical images affects the reliability of diagnosis results to a great extent. Excellent segmentation algorithms can provide a reliable basis for clinical diagnosis and pathological research, and assist doctors make accurate diagnoses, thereby improving diagnosis efficiency [52]. Real-time ultrasound examinations mainly depend on the diagnostic experience of sonographers in most hospitals, which results in subjective interpretation and inter-observer variability. In addition, the large number of repetitive real-time ultrasound examinations would place a heavy workload on hospitals and doctors. In order to reduce the workload of doctors and improve the efficiency of breast ultrasound examinations, a sea of traditional segmentation algorithms have been proposed one after another.

However, most conventional segmentation algorithms are based on semi-automatic implementations that require interaction with the user to complete the segmentation of the lesion. Such methods usually require the setting of optimal ROIs/seeds for the image to improve the segmentation performance of the algorithm, which is an extremely tedious process. As shown in Table 2, traditional segmentation algorithms such as RGI [6] cannot accurately segment breast lesion images, and achieve poor results in FPs/image and DSC. Because this kind of algorithm will be affected by people's subjective willingness when interacting with users, resulting in the model not being able to adaptively segment the lesion accurately. In contrast, fully automatic segmentation methods have many good properties such as operator-independence and reproducibility. For this reason, this paper advocates the development of a fully automated segmentation algorithm which is based on a deep learning implementation.

There are currently a lot of networks in deep learning, and Fully Convolutional Networks (FCN) has become the mainstream segmentation network, because it integrates information from different layers and supplements spatial information. However, although it can be compensated by transmitting high-resolution information from the encoding side to the decoding side of the network, multi-scale features still cannot be extracted well by FCN. This is because segmentation of different scales requires different receiving fields, and the receptive field of the convolution kernel is single in FCN. In MF U-Net, we propose a Multi-ScaleDilated Convolutions Module (MDCM) to segment multi-scale targets. The 1×1 convolution kernel is responsible for extracting small target features, while some 3×3 convolution kernels at large dilation rates extract large target features. The ablation experiment of MDCM can verify the effectiveness of this module.

In addition, due to the characteristics of CNNs, ordinary deep learning networks always ignore edge information when learning features. In order to compensate for the disadvantage that neural networks do not learn edge information adequately, we propose the Wavelet Fusion Module (WFM) to improve the network's ability to segment breast lesions with irregular and blurred pixel boundaries. In WFM, the input image is decomposed into a number of detailes images after wavelet transform, which carries a large amount of edge information. The output of WFM is then fused with pooling layers of U-Net to make the network more sensitive to edges, thus improving the segmentation accuracy of the network.

Our fully automatic segmentation network can eliminate the dependence of the operator and the burden of the radiologist. Moreover, segmentation of the breast lesions could provide a priori information to improve ultrasound imaging in modes other than pulse echo by correction of aberrations in different tissues.

VII. CONCLUSION AND FUTURE WORK

In this paper, we present a deep learning framework for solving the problem of multi-scale variation in breast lesions and boundary pixel blurring, named MF U-Net, for fully automated segmentation of breast lesion regions. Wavelet Fusion Module (WFM) is proposed to decompose images to obtain detailed images that carry edge information, which improve the ability of the network to extract edge features. In addition, FM) is proposed. Multi-ScaleDilated Convolutions Module (MDCM) is proposed, where convolution kernels with different receptive fields are designed to simultaneously detect breast lesions of different scales.We also introduce a new com-pound loss function, Focal-DSC loss, which learn class distribution alleviates the problem of unbalanced voxels and forces the model to learn poorly segmented voxels better. The experimental results show that the proposed MF U-Net outperforms other segmentation methods, achieving the highest Recall(0.9421), DSC (0.9535), IOU (0.9112) and lowest FPs/image (0.0694) on the Breast Ultrasound

Image Segmentation Benchmark (BUSIS) dataset, which demonstrates that the proposed method achieves the state-of-the-art performance in breast lesion segmentation.

Our network has achieved success in ultrasound images, but for images of different modes (such as X-ray images), its segmentation effect will be reduced. In future work, we will try to make better improvements to MF U-Net to improve the spatial complexity, robustness and segmentation accuracy of the model. You can start from the following three aspects: in the first place, use the convolutional layer to compress the feature map with edge information after wavelet transformation to reduce the space complexity of the model; there is one more point, I should touch on, that seek to use two MF U-Net pairs of different Modal medical image feature extraction and multi-modal feature fusion, so that the model maintains good robustness under multi-modality; the last but not the least, design a bidirectional constraint top-level loss function for model loss calculation to improve model segmentation accuracy.

REFERENCES

[1] X. Jiang, H. Tang, and T. Chen, "Epidemiology of gynecologic cancers in China," *J. Gynecologic Oncol.*, vol. 29, no. 1, pp. 1–7, 2018.

[2] A. Brunßen, J. Hübner, A. Katalinic, M. R. Noftz, and A. Waldmann, "Breast cancer epidemiology," in *Management of Breast Diseases*. Cham, Switzerland: Springer, 2016, pp. 125–137.

[3] E. S. Hwang, D. Y. Lichtensztajn, S. L. Gomez, B. Fowble, and C. A. Clarke, "Survival after lumpectomy and mastectomy for early stage invasive breast cancer: The effect of age and hormone receptor status," *Cancer*, vol. 119, no. 7, pp. 1402–1411, 2013.

[4] Y. Zhou, J. Xu, Q. Liu, C. Li, Z. Liu, M. Wang, H. Zheng, and S. Wang, "A radiomics approach with CNN for shear-wave elastography breast tumor classification," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1935–1942, Sep. 2018.

[5] A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions," *Radiology*, vol. 196, no. 1, pp. 123–134, 1995.

[6] K. Drukker, M. L. Giger, K. Horsch, M. A. Kupinski, C. J. Vyborny, and E. B. Mendelson, "Computerized lesion detection on breast ultrasound," *Med. Phys.*, vol. 29, no. 7, pp. 1438–1446, Jun. 2002.

[7] M. H. Yap, E. A. Edirisinghe, and H. E. Bez, "A novel algorithm for initial lesion detection in ultrasound breast images," *J. Appl. Clin. Med. Phys.*, vol. 9, no. 4, pp. 181–199, 2008.

[8] J. Shan, H. D. Cheng, and Y. Wang, "Completely automated segmentation approach for breast ultrasound images using multiple-domain features," *Ultrasound Med. Biol.*, vol. 38, no. 2, pp. 262–275, Feb. 2012.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2009.

[10] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 2843–2851.

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

[12] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218–1226, Jul. 2017.

[13] W. K. Moon, Y.-W. Lee, H.-H. Ke, S. H. Lee, C.-S. Huang, and R.-F. Chang, "Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 190, Jul. 2020, Art. no. 105361.

[14] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, and V. Vaidya, "Understanding the mechanisms of deep transfer learning for medical images," in *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer, 2016, pp. 188–196.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[16] H. Yang, W. Huang, K. Qi, C. Li, X. Liu, M. Wang, H. Zheng, and S. Wang, "Clci-net: Cross-level fusion and context inference networks for lesion segmentation of chronic stroke," in *Medical Image Computing and Computer Assisted Intervention*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham, Switzerland: Springer, 2019, pp. 266–274.

[17] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," 2018, *arXiv:1811.12231*. [Online]. Available: https://arxiv.org/abs/1811.12231

[18] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture and art with deep neural networks," *Current Opinion Neurobiol.*, vol. 46, pp. 178–186, Oct. 2017.

[19] W. Brendel and M. Bethge, "Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet," 2019, *arXiv:1904.00760*. [Online]. Available: https://arxiv.org/abs/1904.00760

[20] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," 2015, *arXiv:1505.07376*. [Online]. Available: https://arxiv.org/abs/1505.07376

[21] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran, "Assessing shape bias property of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1923–1931.

[22] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.

[23] F. Ye, X. Li, and X. Zhang, "FusionCNN: A remote sensing image fusion algorithm based on deep convolutional neural networks," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 14683–14703, Jun. 2019.

[24] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2129–2137.

[25] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[26] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[27] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 405–420.

[28] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3562–3572.

[29] J. Li, Z. L. Yu, Z. Gu, H. Liu, and Y. Li, "Dilated-inception net: Multi-scale feature aggregation for cardiac right ventricle segmentation," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 12, pp. 3499–3508, Dec. 2019.

[30] X. Liu, H. Yang, K. Qi, P. Dong, Q. Liu, X. Liu, R. Wang, and S. Wang, "MSDF-Net: Multi-scale deep fusion network for stroke lesion segmentation," *IEEE Access*, vol. 7, pp. 178486–178495, 2019.

[31] K. Qi, H. Yang, C. Li, Z. Liu, M. Wang, Q. Liu, and S. Wang, "X-Net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies," in *Medical Image Computing and Computer Assisted Intervention*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham, Switzerland: Springer, 2019, pp. 247–255.

[32] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Med. Phys.*, vol. 46, no. 2, pp. 576–589, Feb. 2018.

[33] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.

[34] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, and D. L. Collins, "BEaST: Brain extraction based on nonlocal segmentation technique," *NeuroImage*, vol. 59, no. 3, pp. 2362–2373, 2012.

[35] S. Jadon, "A survey of loss functions for semantic segmentation," 2020, *arXiv:2006.14822*. [Online]. Available: https://arxiv.org/abs/2006.14822

[36] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich, J. Dean, M. Tradewell, A. Shah, R. Tejpaul, Z. Edgerton, M. Peterson, S. Raza, S. Regmi, N. Papanikolopoulos, and C. Weight, "The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," 2019, *arXiv:1904.00445*. [Online]. Available: https://arxiv.org/abs/1904.00445

[37] M. Yeung, E. Sala, C. B. Schnlieb, and L. Rundo, "A mixed focal loss function for handling class imbalanced medical image segmentation," 2021, *arXiv:2102.04525*. [Online]. Available: https://arxiv.org/abs/2102.04525

[38] Y. F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient iou loss for accurate bounding box regression," 2021, *arXiv:2101.08158*. [Online]. Available: https://arxiv.org/abs/2101.08158

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[40] F. Qi, C. Lin, G. Shi, and H. Li, "A convolutional encoder-decoder network with skip connections for saliency prediction," *IEEE Access*, vol. 7, pp. 60428–60438, 2019.

[41] S. Mallat, *A Wavelet Tour of Signal Processing—The Sparse Way*, 3rd ed. New York, NY, USA: Academic, 2009. [Online]. Available: https://www.elsevier.com/books/a-wavelet-tour-of-signal-processing/mall% at/978-0-12-374370-1

[42] X. Fu, N. Cai, K. Huang, H. Wang, P. Wang, C. Liu, and H. Wang, "M-Net: A novel U-Net with multi-stream feature fusion and multi-scale dilated convolutions for bile ducts and hepatolith segmentation," *IEEE Access*, vol. 7, pp. 148645–148657, 2019.

[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[44] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 99, pp. 2999–3007, Jul. 2017.

[45] Y. Zhou, W. Huang, P. Dong, Y. Xia, and S. Wang, "D-UNet: A dimension-fusion u shape network for chronic stroke lesion segmentation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 940–950, May 2021.

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

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

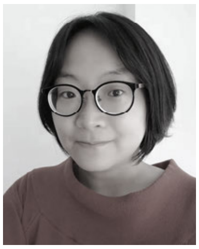
[48] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4, doi: 10.1109/VCIP.2017.8305148.

[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[50] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[51] D. Naik and C. D. Jaidhar, "Image segmentation using encoder-decoder architecture and region consistency activation," in *Proc. 11th Int. Conf. Ind. Inf. Syst. (ICIIS)*, Dec. 2016, pp. 724–729.

[52] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-Unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44247–44257, 2019.



JINGYAO LI received the bachelor's degree from Henan University of Technology, in 2012, and the master's degree from Guangdong University of Technology, in 2014, where she is currently pursuing the Ph.D. degree with the School of Automation. Her current research interests include computer vision, medical image processing, and cyber-physical systems.

LIANGLUN CHENG received the B.E. and M.S. degrees in automation from Huazhong University of Science and Technology, Wuhan, China, in 1988 and 1992, respectively, and the Ph.D. degree in automation from the Chinese Academy of Sciences, in 1999.

Since 2003, he has been a Professor with Guangdong University of Technology, Guangzhou, China. He is currently the Dean of the School of Computer, Guangdong University of Technology. His research interests include terahertz detection technology, visual image processing, and cyber-physical systems.

TINGJIAN XIA received the bachelor's degree from Guangdong University of Technology, in 2020, where he is currently pursuing the M.S. degree with the School of Computer Science. His research interest includes precise segmentation of medical images.

HAOMIN NI is currently pursuing the bachelor's degree with the School of Automation, Guangdong University of Technology. His research interests include deep learning and medical image segmentation.

JIAO LI received the Bachelor of Medical Imaging degree from Tianjin Medical University, in 2014, and the M.D. degree in medical imaging and nuclear medicine from Sun Yat-sen University, in 2019.

She is currently a Postdoctoral Researcher with Sun Yat-sen University Cancer Center, Guangzhou, China. She has participated in a few national or provincial-level scientific research projects, including the National Key Research and Development Plan and Guangzhou Science and Technology Project. Her research interests include breast imaging, radiomics, and artificial intelligence.