Spatial Enhanced Rotation Aware Network for Breast Mass Segmentation in Digital Mammogram

YULIN CHENG, YING GAO, LINSEN XIE, XINYAN XIE, AND WENGEN LIN

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

Corresponding author: Ying Gao (gaoying@scut.edu.cn)

This work was supported in part by the Guangzhou People’s Livelihood Science and Technology Project under Grant 201803010097, and in part by the Guangzhou City Industrial Technology Major Research Project under Grant 201802010035.

ABSTRACT Breast cancer is the most common cancer with highest mortality risk among the female worldwide and breast mass is the most effective sign for cancer identification. Thus, accurate segmentation of breast mass is regarded as a key step to reduce the death rate. Traditional segmentation methods require prior knowledge and manually set parameters, while recent studies prefer to construct neural networks based on feature reuse. However, breast mass can display in different orientations and the spatial context is complex, which makes the segmentation remain a challenging task. For these concerns, we propose a Spatial Enhanced Rotation Aware Network (SERAN) for automatic breast mass segmentation. SERAN consists of two critical components: 1) a residual attention encoder with spatial enhancement mechanism for effective feature extraction, and 2) a decoder constructed by multi-stream rotation aware blocks for feature fusion and prediction refinement. To optimize SERAN better and avoid misclassification in background area, a regulation item named Inside-outside Loss (IOL) is used in training procedure. The experimental results tested on a representative subset of Digital Database for Screening Mammography (DDSM) dataset show that SERAN outperforms state-of-the-art methods among most adopted evaluation metrics.

INDEX TERMS Breast mass segmentation, digital mammogram, attention mechanism, rotation aware, inside-outside loss.

I. INTRODUCTION

Breast cancer is one of the most harmful diseases among the female worldwide. According to 2018 Global Cancer Statistics [1], breast cancer sufferers account for a quarter of all female cancer patients. Even more shocking, about 3% of early-stage patients die from it [2]. Therefore, early diagnosis is highly suggested for reducing death rate of breast cancer. As is known to all, breast mass is one of the most effective signs for cancer identification [3]. Thus, breast mass segmentation on medical image is regarded as the first step of early diagnosis and the key step prior to classification of benign and malignant. Traditional approaches for breast mass segmentation are manually, time-consuming and heavily dependent on radiologist’s experience. To reduce processing time and improve the accuracy of segmentation result, computer-aided detection (CADe) technology has been rapidly developed since the late 1980s [4] and digital mammogram is the most reliable technique which widely used in breast mass segmentation [5]. However, breast masses are varied in a wide range in shape, size and texture, which make the segmentation remain a challenging task [6].

Various machine learning algorithms were utilized to establish traditional CADe systems for disease detection and segmentation [7], [8], especially for breast mass segmentation [9]. Region growing and thresholding were the two most widely used methods. With respect to region growing, Mencattini et al. [10] introduced an effective region growing algorithm for breast mass segmentation on Digital Database for Screening Mammography (DDSM) [11] dataset. They followed the typical processing flow which consists of artifacts removal, contrast enhancement and segmentation. Moreover, an iterative post processing step was designed to remove peninsulas on the boundary of mass. A precision
of 93.38% and a recall of 88.34% were achieved on a subset of DDSM with 200 images. Regarding to thresholding technique, Kom et al. [12] reported promising results for detecting mass using a local adaptive thresholding method combined with linear transformation enhancement. The adaptive threshold value was calculated based on two context windows of distinct sizes. Their method achieved great performance on a dataset of 61 mammograms images and got a sensitivity of 95.91%. Besides, energy function-based methods [13] and clustering based methods [14] were also adapted as the detectors. To integrate the advantages of different techniques, some studies combined several techniques in a processing flow and achieved great improvement in breast mass segmentation. Min et al. [6] used multi-scale morphological filtering and simple linear iterative clustering segmentation to detect suspicious areas of different sizes. Multiple cascaded random forests were employed to make the final decision based on the features extracted from suspicious areas. Two sensitives of 94% and 77% were achieved by their approach for INbreast and DDSM BCRP respectively. Chakraborty et al. [15] presented a multilevel method controlled by thresholding and region growing in a coarse-to-fine manner. At each step, a thresholding method was used to find the focal region of a mass, and then, a region growing algorithm was employed to detect the accurate mass using gradient and intensity information. With a mixed dataset of 107 images from the mini-MIAS database and 158 digital radiography images from a local database, two sensitives of 95% and 98.8% were achieved by their approach respectively. Sharma et al. [16] combined watershed segmentation and k-means clustering to detect breast mass. The texture features extracted from suspicious areas were used to identify the true breast mass.

Although these methods achieved good performances, they were relatively simple and sensitive to some initialization parameters, such as seed, threshold value and number of cluster centroids. In most cases, the parameters or the initialization strategies were decided by experience. Otherwise, these methods highly rely on image enhancement and lack the abilities to extract required pattern from raw image data directly. This made them may ineffective in complex scenes and large-scale dataset.

In recent years, convolutional neural network (CNN) has developed rapidly and has become the most commonly selected deep learning structure in image processing domain. CNN can extract features by abundant kernels automatically and fuse them in a non-linear manner. All the parameters of CNN can be initialized randomly, adjusted by back propagation and trained in an end-to-end manner. Based on these characteristics, CNN has achieved great improvement in most computer vision tasks, such as image classification [17], object detection [18] and semantic segmentation [19].

Regarding to semantic segmentation using traditional CNN, the input image is divided into small patches using sliding window technique and the label of the central pixel of each patch is predicted by CNN [20]. However, small image patch only contains local context information about the patch itself and losses global context information about the whole image. Because of these, the performance of traditional CNN is limited. To address these problems, Long et al. [21] proposed fully convolutional networks (FCN) for segmentation. FCN replaces all the fully connected layers by convolution layers. For this reason, the input of FCN is allowed in arbitrarily size. As the size of segmentation result should be the same as the input, FCN upsamples the feature maps from high level layers and introduces skip connection to fuse the outputs from different layers. The experimental results showed that the spatial information supplied by skip connection and gradually upsampling can improve the performance of FCN. However, the segmentation details are still far from expectation. Inspired by FCN, many successful neural networks were proposed for image segmentation task, such as U-Net [19], DeepLab [22] and PSPNet [23]. Among these works, the encoder-decoder structure of U-Net is the easiest to implement and extend. The encoder part of U-Net employs multi-stage convolutions connected by maxpooling layers to extract features and expand receptive fields, while the decoder part contains multi-stage convolutions connected by upsampling layers to fuse the features and expand the size. Skip connection is applied in every decoder stage to concatenate the feature map from upsampling layers with the feature maps from the corresponding stage in encoder. Benefit from small stride upsampling to expand the size gradually and skip connection to make up for spatial information, U-Net achieves great improvement in image segmentation.

Afterwards, various U-Net based architectures have been proposed to handle different application scenes [24]–[27], especially in breast mass segmentation [28]–[30]. Li et al. [28] proposed an Attention Dense-U-net for breast mass segmentation on a subset of DDSM. They used dense connected block to remould the encoder and applied Attention Gates [31] at skip connection step for spatial information enhancement. An Area under the Receiver Operating Characteristic Curve (AUC) of 0.8605 was reported. Hai et al. [29] remoulded both the encoder and decoder by dense connected block. Moreover, they held the opinion that the multi-scale context was useful in segmentation task. Thus, they used atrous spatial pyramid pooling (ASPP) mechanism for multi-scale context caption. Their method achieved the best performance on a self-collected dataset with 380 mammograms in total. Different from the approaches mentioned above, Sun et al. [30] proposed an attention-guided dense-upsampling block to decrease the information loss in upsampling operation. The original upsampling operation was replaced by a dense-upsampling block and a channel-wise attention was equipped for information enhancement. An average Dice coefficient of 81.8% was achieved on CBIS-DDSM dataset.

To improve the encoder part of U-Net, most of the studies mentioned above held the opinion that feature reuse, which can be modeled by residual connection [17], is the most important factor. Unlike them, we think that the
feature extraction can be further improved by spatial attention. Spatial attention, which is mentioned in [32], can adjust the focus to the region of interest and filter out interference information. Thus, it is possible to improve the feature extraction by explicitly modeling the spatial attention. For this concern, spatial attention block (SA-Block) with a mask attention branch is designed in this paper to provide soft spatial attention about suspicious regions. Moreover, breast masses can display in every orientations in complex scenes, which can be summarized as rotation. Most studies employed data augment method to improve the perception to mass rotation. Different from them, we design multi-stream rotation aware block (MSRA-Block), which contains two extra asymmetric convolutions in parallel, to enhance the rotation invariance and refine the prediction in decoder part.

As a summary of the foregoing, a U-type Spatial Enhanced Rotation Aware Network (SERAN) is designed for breast mass segmentation in digital mammogram in this paper. The overview of SERAN is illustrated as Fig. 1. The materials used in this study and the details of SERAN are described in following sections.

### A. DATA PREPARATION

DDSM database from the University of Florida is used in this study. Approximately 2,500 cases are collected in DDSM and each of them contains two X-ray mammograms about each breast. The database can be downloaded online and all the image sizes are larger than 2000 × 3000 pixels. The location and type of suspicious area in each image has been annotated by experienced radiologists as ground truth. The images are encoded in ''.LJPEG'' format and the corresponding ground truths are saved in ''.OVERLAY'' format. To transform the images and the ground truths to ''.PNG'' format, which can be easily used during programing, a public tool named DDSM-LJPEG-Converter is used.

In this study, 400 representative images are selected from DDSM by an experience radiologist. The standard of the selection can be described as following: 1) the mass is relative clear in the image, and 2) the annotated area can cover the mass but not much bigger than the mass. The images are randomly divided into three nonoverlapping parts in a ratio of 4:1:1, which are used as training set, validation set and test set.

The rest of this paper is organized as follows. In Section II, we describe the details about dataset and pre-processing as well as the architecture of proposed SERAN. To verify the ability of proposed method, extensive experiments are summarized in Section III. Finally, a conclusion is drawn in Section IV.

### II. MATERIALS AND METHODOLOGY

This paper aims to build a Spatial Enhanced Rotation Aware Network (SERAN) for breast mass segmentation in digital mammogram. The overview of SERAN is illustrated as Fig. 1. The materials used in this study and the details of SERAN are described in following sections.

#### A. DATA PREPARATION

DDSM database from the University of Florida is used in this study. Approximately 2,500 cases are collected in DDSM and each of them contains two X-ray mammograms about each breast. The database can be downloaded online and all the image sizes are larger than 2000 × 3000 pixels. The location and type of suspicious area in each image has been annotated by experienced radiologists as ground truth. The images are encoded in ''.LJPEG'' format and the corresponding ground truths are saved in ''.OVERLAY'' format. To transform the images and the ground truths to ''.PNG'' format, which can be easily used during programming, a public tool named DDSM-LJPEG-Converter is used.

In this study, 400 representative images are selected from DDSM by an experience radiologist. The standard of the selection can be described as following: 1) the mass is relative clear in the image, and 2) the annotated area can cover the mass but not much bigger than the mass. The images are randomly divided into three nonoverlapping parts in a ratio of 4:1:1, which are used as training set, validation set and test set.
B. PRE-PROCESSING AND DATA AUGMENTATION

Each image in DDSM is composed of breast region and background. However, some artifacts exist in the background. These artifacts can be regarded as noise and affect the ability of neural network badly. Thus, we design several steps to remove the artifacts in the background for denoising. Firstly, we employ gamma transformation to correct image and enhance the details about the breast region. Secondly, the image is binarized by Otsu [33]. Then, the maximum connected area in the binary image is obtained as the breast region. With these, the pixel value of the breast region is recorded as 1 and the background is recorded as 0. Finally, the original image is multiplied by the breast region binary image to get the unlabeled image. The results are shown as Fig.2.

After that, all the images are resized to 224 × 224 based on bilinear interpolation for reducing the computation cost and all the pixel values in an image are scaled into a range from 0 to 255 by Min-Max normalization [34]. The sizes of masses are varied in a wide range and the distribution in each data subset is similar to the others (as shown in Fig.3).

For training a deep neural network for segmentation, a large number of annotated images are required to avoid overfitting. To enlarge the training set, several data augmentation methods are applied. Firstly, each image is mirrored in horizontal direction. Then, all the images are rotated according to following principles: 1) take the central point of the image as the origin of coordinate system, 2) generate an angle in a range from -10 to 10 degree randomly, 3) rotate the image according to the origin point and the degree, 4) fill the extra area by 0 to form a rectangle image, and 5) crop the image to 224 × 224 according to origin point. With these, the size of training set is quadruple and the data diversity is increased.

C. SPATIAL ENHANCED ROTATION AWARE NETWORK

The proposed SERAN consists of four parts: a residual spatial attention encoder, a multi-stream rotation aware decoder, skip connections and a final prediction layer. The descriptions of all the modules are summarized in following subsections.

1) RESIDUAL SPATIAL ATTENTION ENCODER

As shown in Fig.3, the shapes and sizes of different breast masses are varied in a wide range. Thus, it is required to build a powerful encoder for effective feature extraction.

Typical encoder is made up of convolution layers for feature extraction and pooling layers for receptive field expanding. All the layers are formed in a series connection style and the input of layer \( l \) is only the output of layer \( l - 1 \). However, the ability of encoder may degrade with gradient vanishing. To address this problem, He et al. [17] proposed ResNet with residual connection. Using residual connection, the input of convolution layer \( l \) is summed with the convolution result to form the output, which can be expressed as:

\[
 x_{l+1} = \sigma ( F(x_l) + H(x_l) )
\]

where \( x_{l+1} \) and \( x_l \) are inputs of layer \( l \) and \( l + 1 \) respectively. \( F(\cdot, \omega) \) is a learning function and \( H(\cdot) \) is a mapping function. When the number of channels of \( x_l \) is equal to the channel number of \( F(\cdot, \omega) \), \( H(\cdot) \) is an identity mapping function (shown as Fig.4 (a)). Otherwise, \( 1 \times 1 \) convolution is used in \( H(\cdot) \) to adjust the channel number (shown as Fig.4 (b)). \( \sigma \) refers to the ReLu [35] function for activation.

Residual connection can avoid gradient vanishing and provide feature reuse for better convergence. For these reasons, we employ the residual learning structure as a basic module in encoder part.

In addition, in our view, spatial information is the most important factor to boost the performance in segmentation task. Thus, it will be helpful to model the attention on spatial information explicitly.

Inspired by [32], a spatial attention block (SA-Block) is designed to enhance the spatial context information at each pixel position. As shown in Fig.5, the SA-Block consists of a feature extraction branch and a mask attention branch. The feature extraction branch utilizes two \( 3 \times 3 \) convolution layers to capture features from complex image context. The mask attention branch is implemented as a U-Net liked structure. A single-channel attention map is generated by the mask.
FIGURE 4. Two variants of residual connected block: (a) basic residual block, (b) residual block with channel number adjustment. The numbers in the boxes are kernel size, filter number, stride, and padding respectively.

FIGURE 5. The structure of SA-Block.

FIGURE 6. Structures of Mask Attention Branches in SA-Block of different depths.

FIGURE 7. The top row shows the comparative prediction result (c) of original image (a) compared to ground truth (b). The bottom row shows the attention maps from each encoder stages.

attention branch. Each value in the attention map represents the weight of the corresponding pixel position in spatial dimension. Because the values in the attention map range from 0 to 1, the output of the feature extraction branch is used twice to avoid the degradation of feature values in the deep layers. The output of the feature extraction branch is weighted by the attention map firstly and then summed with the weighted result. The summed result is then activated by ReLu. The formulation of SA-Block can be summarized as:

\[ H_{i,j,c}(x) = \sigma \left( (1 + M_{i,j}(x)) \ast F_{i,j,c}(x) \right) \]  

(2)

where \( i, j \) and \( c \) are pixel locations in width, height and channel dimension respectively. \( M(\cdot) \) and \( F(\cdot) \) are the learning functions of mask attention branch and feature extraction branch respectively. \( \sigma \) is the ReLu activation function and Batch Normalization (BN) [36] is equipped after each convolution operation.

As well known, the low level layers in a deep CNN model capture shallow features, such as shape, size and texture, while the high level layers contain more abstract semantic information. From this point of view, the low level layers need stronger spatial attention. Thus, two similar mask attention branches with different depths are proposed, which are shown in Fig.6. The deeper mask attention branch is equipped in the shallow layers of the encoder and the shallower one is embedded into the deeper layers.

Composed of residual convolution blocks and SA-Blocks, the encoder can focus on extracting features from suspicious places (as shown in Fig.7). Moreover, the responses of regions with similar features to the target in one image can be restrained gradually. The architecture of encoder is shown as the left part of Fig.1.

2) MULTI-STREAM ROTATION AWARE DECODER

In complex scenes, the masses can display in every orientations, which can be summarized as rotation. In order to assist doctors for precise diagnosis, the neural network must be aware to mass rotation and predict robust results in different rotations. The goal of each decoder stage is to fuse the features and refine the prediction of mass location and contour.
of MSRA-Block can be described as:

$$x_{l+1} = \sigma \left( F_{3 \times 3} (x_l) + F_{1 \times 3} (x_l) + F_{3 \times 1} (x_l) \right) \quad (3)$$

where $x_{l+1}$ and $x_l$ are inputs of layer $l$ and $l+1$ respectively. $F_{3 \times 3}(\cdot)$, $F_{1 \times 3}(\cdot)$, and $F_{3 \times 1}(\cdot)$ are the learning functions of three feature extraction streams with $3 \times 3$, $1 \times 3$ and $3 \times 1$ kernels respectively. $\sigma$ is the ReLu function used for activation.

The decoder consists of four stages and each stage is composed of two MSRA-Blocks (shown as the right part of Fig.1). The input of each stage is the concatenate feature maps which is formed by the corresponding skip connection and the deconvolution result of previous stage.

3) SKIP CONNECTION AND FINAL PREDICTION
As a common view, the detail information lost in pooling-oversampling structure can be supplied through skip connection. Thus, skip connection is used to connect each encoder stage with corresponding decoder stage. The output of each SA-Block is treated as a part of the input of corresponding MSRA-Block through skip connection.

The output of the last MSRA-Block is used for final prediction. A $1 \times 1$ convolution is employed to fuse the features, reduce the number of channels and keep the spatial size. Two feature maps are generated and a channel-wise softmax operation is used to transform the pixel values to probabilities. One of the feature maps is regarded as background and another one predicts the mass.

D. INSIDE-OUTSIDE LOSS
To train a segmentation network, binary cross-entropy (BCE) loss was the most common choice due to its smooth derivative and the stable optimization procedure. The formulation of BCE can be denoted as:

$$l^{bce} = -\frac{1}{N} \sum_{i} \sum_{j} l^{ce}_{ij}$$

$$l^{ce} = - \left( y^{true} \log y^{predict} + (1 - y^{true}) \log (1 - y^{predict}) \right) \quad (4)$$

where $l^{ce}$ is the cross entropy loss of one pixel. $y^{predict}$ is the predicted label, and $y^{true}$ denotes the actual label. BCE is the average cross entropy loss of all the pixels.

It is obvious that BCE regards segmentation task as pixel-wise classification task and treats each pixel equally. However, the number of pixels representing breast mass is only a small proportion of the entire image. In other words, there exists heavy data imbalance when applying BCE as the objective function. To address the problem, Dice loss was presented to measure the similarity between the predicted area and the ground truth, which can be denoted as:

$$l^{dice} = 1 - \frac{2 \times |A \cap B|}{|A| + |B|} \quad (5)$$

where $A$ is the predicted probability map and $B$ is the ground truth. $| \cdot |$ is used to calculate the summation of pixel values. $|A \cap B|$ represents the intersection area of prediction and ground truth, which is named True Positive (TP).

Dice loss offers a hidden linked constraint about the intersection area within mass region and the area outside mass region, which is indicated by $|A|$. The change of TP is the main factor affecting Dice loss. In our point of view, explicitly constrain the two parts can be more efficient for model optimization. Thus, a novel regulation item, named Inside-outside Loss (IOL), is designed to explicitly constrain both the intersection area and the area outside the mass for better optimization. The goal of IOL is to maximize the probabilities inside the ground truth area and minimize the outside probabilities. The formulation of IOL is shown details. Thus, it is possible to boost the rotation robustness of the neural network by building a rotation aware decoder.

According to [37], different asymmetric convolutions are robust to different rotation scenes. As shown in Fig.8, $3 \times 1$ kernel is much more robust to horizontal flipping than $3 \times 3$ kernel.

In order to extract more effective features and boost the rotation robustness, we design a multi-stream rotation aware block (MSRA-Block) to fuse features with different receptive fields (as shown in Fig.9). The MSRA-Block contains a $3 \times 3$ convolution and two asymmetric convolutions whose kernels are $1 \times 3$ and $3 \times 1$ respectively. The three convolutions are used to get different purposes. The $3 \times 3$ convolution extracts features in a relatively larger receptive field. The $1 \times 3$ and $3 \times 1$ convolutions are used to boost the rotation robustness to vertical and horizontal flipping respectively. Moreover, the width of the decoder is expanded resulting in multi-scale feature fusion. The feature maps from the three streams are added up as a fusion result. BN and ReLu are used to activate the result in a nonlinear manner. The formulation of MSRA-Block can be described as:

$$x_{l+1} = \sigma \left( F_{3 \times 3} (x_l) + F_{1 \times 3} (x_l) + F_{3 \times 1} (x_l) \right) \quad (3)$$

where $x_{l+1}$ and $x_l$ are inputs of layer $l$ and $l+1$ respectively. $F_{3 \times 3}(\cdot)$, $F_{1 \times 3}(\cdot)$, and $F_{3 \times 1}(\cdot)$ are the learning functions of three feature extraction streams with $3 \times 3$, $1 \times 3$ and $3 \times 1$ kernels respectively. $\sigma$ is the ReLu function used for activation.
Intersection Over Union (IOU) and Dice coefficient. All of the metrics are denoted as following:

\[
\text{InsideLoss} = 1 - \frac{|A \cap B|}{|B|}
\]

\[
\text{OutsideLoss} = \frac{|A| - |A \cap B|}{H \times W - |B|}
\]

\[
\text{IOU} = \text{InsideLoss} + \text{OutsideLoss}
\]

where \( A \) is the predicted probability map and \( B \) is the ground truth. \( H \) and \( W \) are the height and width of image respectively. \(| \cdot |\) is used to calculate the summation of pixel values. \(|A \cap B|\) represents the intersection area of prediction and ground truth, which is named True Positive (TP). And \(|A| - |A \cap B|\) represents the misclassification in background area, which is named False Positive (FP). The InsideLoss indicates the probability loss inside the ground truth mass while the OutsideLoss indicates the loss outside the mass. In other words, the Insideloss focus on increasing TP indicator and the Outsideloss aims to reduce the FP indicator.

To avoid overfitting, we treat IOL as a regulation item and set a small weight for IOL. The final objective function is composed by Dice loss and IOL, which is defined as following:

\[
\ell_{\text{total}} = \ell_{\text{dice}} + \lambda \times \text{IOL}
\]

where \( \lambda \) is the weight of IOL. Using IOL, the network can avoid misclassification in background area and make a better prediction (as shown in Fig.10).

### III. EXPERIMENTS AND ANALYSIS

#### A. EVALUATION METRICS

To estimate the performance of proposed approach, five commonly used evaluation metrics in image segmentation task are adopted, namely Sensitive, Specificity, Accuracy, Intersection Over Union (IOU) and Dice coefficient. All of the metrics are shown in Table 1.

| PREDICTION | LABEL   | Mass | Background |
|------------|---------|------|------------|
| Mass       | TP      | FP   |            |
| Background | FN      | TN   |            |

where \( TP \) is the number of pixels correctly predicted in mass region and \( TN \) is the number of correct predicted pixels in background. \( FP \) are the pixels predicted wrong as mass and \( FN \) are the wrong pixels as background. The confusion matrix is showed in Table 1 to make a clear explanation about \( TP \), \( TN \), \( FP \) and \( FN \).

Moreover, an average value (MEAN) of the results of 5 independent experiments is recorded as the final result to preclude statistical error and result bias. Furthermore, standard deviation (STD) is computed to reflect performance stability.

#### B. IMPLEMENTATION DETAILS

Adam [38] method is employed as an optimizer to optimize the model. The parameters of Adam are set as: \( \beta_1 = 0.9 \), \( \beta_2 = 0.999 \), and \( \varepsilon = 1e^{-8} \). The learning rate is initialized to 0.001 and halved every 20 epochs after the 70\textsuperscript{th} epoch. The batch size is set to 16 and the weight of IOL is set to 0.01 in training phase.

Several early stopping strategies are used to avoid overfitting. If the loss of validation set keeps increasing in 5 consecutive epochs or changes within a small range in 10 consecutive epochs, the training procedure will stop and the model with the lowest validation loss will be saved.

All the preprocessing steps and data augment methods are implemented on Python interface of OpenCV library. The proposed neural network is implemented on Python interface of Pytorch platform, while related experimentations are tested using Intel(R) Core(TM) i7-6850K @ 3.60GHz CPU and two GeForce GTX 1080p GPUs on Ubuntu16.04.1. The versions of Python, OpenCV and Pytorch are 3.6.6, 3.4.3 and 1.0.1 respectively.
comprehensive comparison with 5 state-of-the-art segmentation methods is presented at the end of this section to show the advancement of proposed SERAN. U-Net is implemented as a baseline model. All the results are achieved in test set to show the generalization ability of each methods. Since the values of Specificity and Accuracy are close, we record them in more detail.

1) CONTRIBUTION ANALYSIS OF RESIDUAL SPATIAL ATTENTION ENCODER
To analyze the contribution of residual spatial attention encoder, a spatial enhanced network (SEN) is constructed after replacing the encoder of U-Net by proposed encoder. The performances of SEN and U-Net are showed in the top 2 rows in Table 2 for evaluation. In addition, both SEN and U-Net are trained by Dice loss.

In summary, SEN wins 9 out of 10 comparisons and most of the metrics are improved significantly. Due to the explicitly modeling of spatial attention, SEN can focus more on the region of interest and make the feature extraction more effective. As shown in Fig.7, the number of focuses is obviously decreased and the response values in mass region are relatively bigger in the attention map of deeper layer. That means the network owns the ability to identify the mass from multiple regions with high response values. Moreover, the performance of SEN is more stable than U-Net, which is indicated by STDs.

2) CONTRIBUTION ANALYSIS OF MULTI-STREAM ROTATION AWARE DECODER
As shown in the line 2 and line 3 of Table 2, the performances of SERAN and SEN are showed to investigate the effect of multi-stream rotation aware decoder mentioned in Section II.C.2. SERAN transforms the decoder part of SEN by the proposed one. For the sake of fairness, Dice loss is used to train SERAN.

SERAN wins all the competitions about average performance. Due to the multi-stream structure, the width of the decoder is expanded and features from different aspects are fused. Thus, the prediction result can be refined to a certain extent. The Accuracy and Specificity of SERAN is better and more stable than U-Net and SEN. That means the location of suspicious area is easier to detect by using SERAN. Although SERAN seemed more unstable than U-Net and SEN in

### TABLE 2. Contribution analysis (MEAN±STD).

| METHOD     | Sensitive (%) | Specificity (%) | Accuracy (%) | IOU (%) | Dice (%) |
|------------|--------------|-----------------|--------------|---------|----------|
| U-Net      | 82.60 ± 1.63 | 99.904 ± 0.027  | 99.821 ± 0.018 | 70.38 ± 0.60 | 81.21 ± 0.32 |
| SEN        | 87.18 ± 1.15 | 99.901 ± 0.020  | 99.835 ± 0.010 | 72.71 ± 0.25 | 83.41 ± 0.15 |
| SERAN      | 87.69 ± 1.48 | 99.910 ± 0.012  | 99.839 ± 0.005 | 73.01 ± 0.69 | 83.71 ± 0.61 |
| SERAN-JOL  | 87.70 ± 1.33 | 99.906 ± 0.008  | 99.846 ± 0.005 | 73.95 ± 0.70 | 84.30 ± 0.61 |

### TABLE 3. Performance on expanded test set (MEAN±STD).

| METHOD     | Sensitive (%) | Specificity (%) | Accuracy (%) | IOU (%) | Dice (%) |
|------------|--------------|-----------------|--------------|---------|----------|
| U-Net      | 81.58 ± 1.80 | 99.891 ± 0.018  | 99.808 ± 0.016 | 68.95 ± 0.80 | 79.82 ± 0.72 |
| SEN        | 86.63 ± 1.19 | 99.901 ± 0.020  | 99.826 ± 0.008 | 71.66 ± 0.50 | 82.49 ± 0.39 |
| SERAN      | 86.47 ± 1.44 | 99.902 ± 0.017  | 99.831 ± 0.002 | 72.09 ± 0.43 | 82.70 ± 0.24 |

### TABLE 4. Performance of different weight of IOL (MEAN±STD).

| λ          | Sensitive (%) | Specificity (%) | Accuracy (%) | IOU (%) | Dice (%) |
|------------|--------------|-----------------|--------------|---------|----------|
| 1          | 94.94 ± 0.71 | 99.820 ± 0.009  | 99.793 ± 0.007 | 69.09 ± 0.98 | 80.90 ± 0.82 |
| 0.1        | 86.80 ± 1.23 | 99.907 ± 0.008  | 99.826 ± 0.027 | 72.29 ± 1.18 | 83.00 ± 1.21 |
| 0.01       | 87.70 ± 1.33 | 99.906 ± 0.008  | 99.846 ± 0.005 | 73.95 ± 0.70 | 84.30 ± 0.61 |
| 0.001      | 86.09 ± 1.29 | 99.910 ± 0.012  | 99.834 ± 0.018 | 72.53 ± 0.64 | 83.08 ± 0.42 |
the main factor affecting the loss function. The Outside loss in IOL aims to reduce the FP indicator. Thus, it can avoid the prediction of multiple regions. The Inside loss in IOL focus on the perceptron about the region of mass. When the $\lambda$ is small, such as 0.001, IOL has little effect on the loss function. Thus, the performance of SERAN trained by $\lambda = 0.001$ is similar to the one trained by Dice loss and multiple region prediction occurs (as shown in Fig.11 (d)). When the $\lambda$ gets larger, IOL affects the loss function a lot. Thus, SERAN tends to predict the mass with larger size than ground truth to get the highest TP when trained by $\lambda = 1$ (as shown in Fig.11 (a)). That is the reason to explain why it gets the highest Sensitive but falls in other metrics. To make a balance about constraints of the mass and the background, we determined the value of $\lambda$ to 0.01 in this work based on the experimental results.

**4) COMPREHENSIVE COMPARISON WITH STATE-OF-THE-ARTS**

For the purpose to show the advancement of proposed SERAN, 5 state-of-the-art methods are implemented for comprehensive comparison. Three of them are neural networks and the other two are traditional artificial intelligence methods. We implement the three neural networks on Pytorch platform. The two traditional artificial intelligence methods are implemented mainly based on OpenCV library and Sklearn library.

Table 5 shows the comparison between the 5 methods and SERAN. And Table 6 displays the computation cost and model complexity of each method. In this section, SERAN is trained by Dice loss with IOL. The method proposed in [16] has achieved the same results every time. Thus, the STDs of the method are 0. The visual results of these methods are showed in Fig.12. It can be concluded that SERAN outperform all the state-of-the-arts in a certain extent. The performance of SERAN is more stable than other neural networks and competitive to traditional artificial intelligence methods.

The comparison result of sensitive shows that all the methods can predict masses with accurate locations. Indicated by IOU and Dice coefficient, the mass predicted by SERAN owns the biggest overlapping area with ground truth. The Accuracy indicator shows that SERAN can make better prediction about both mass and background than other methods.

The most competitive performance is achieved by the method proposed by Sun et al. [30]. It can be attributed to the dense- upsampling and channel attention. The two mechanisms are effective to maintain most useful features. However, the channel attention is in a fully connected style, which is the leading cause for the high time cost. The method proposed by Li et al. [28] employs spatial attention to guide the prediction made by decoder and achieves competitive results in all the metrics. The method proposed in [29] owns the minimal complexity among all the neural networks and gets a good performance as well. The multi-scale context is effective to extract features under different receptive fields. All the neural networks run faster than the traditional artificial intelligence methods. This may be caused by the great ability of Pytorch framework and the end-to-end style of neural network. The method proposed in [6] has used a complex processing flow to detect and identify the masses in each scale, which consists of morphological filtering, simple linear iterative clustering segmentation, feature extraction and classification. The method proposed in [16] has used watershed segmentation twice to find out all the regions of interest. A k-means clustering algorithm is used to reduce the number of regions found out by the first watershed segmentation algorithm and provide the maker to the second watershed.

**TABLE 5. Comparison with State-of-the-arts (MEAN±STD).**

| METHOD       | Sensitive (%) | Specificity (%) | Accuracy (%) | IOU (%) | Dice (%) |
|--------------|---------------|-----------------|--------------|---------|----------|
| SERAN        | 87.70±1.33    | 99.06±0.005     | 99.846±0.70  | 83.30±40.61 |
| Li et al. [28]| 83.41±1.92    | 99.80±0.013     | 99.805±66.90 | 78.55±11.81 |
| Hai et al. [29]| 81.16±4.13    | 99.891±0.027    | 99.782±5.61  | 77.17±2.65 |
| Sun et al. [30]| 84.57±2.72    | 99.911±0.022    | 99.813±6.21  | 80.60±4.21 |
| Min et al. [6]      | 79.33±1.15    | 99.786±0.010    | 99.694±5.86  | 73.25±3.33 |
| Sharma et al. [16]     | 78.07±0.67    | 99.773±0.011    | 99.632±5.05  | 69.77±6.36 |

**TABLE 6. Computation cost.**

| METHOD          | Testing time on one image (s) | Parameters | FLOPs   |
|-----------------|-------------------------------|------------|---------|
|                 | CPU                           | GPU        |         |
| SERAN           | 0.644                         | 0.014      | 64.5M   | 69.53G |
| Li et al. [28]  | 0.944                         | 0.021      | 11.8M   | 120.35G|
| Hai et al. [29] | 0.256                         | 0.013      | 0.57M   | 6.83G  |
| Sun et al. [30] | 0.877                         | 0.155      | 85.6M   | 83.89G |
| Min et al. [6]  | 125.482                       | /          | /       |
| Sharma et al. [16]| 17.123                      | /          | /       |
segmentation method. Finally, the features extracted from each region are used to identify the true mass. Most of the components are time-consuming which result in the high time costs. However, the performances of these methods are poorer than neural networks mentioned above.

In the real world, it is easy for specialists to discover the masses with large size while the small size masses are difficult to recognize. To evaluate the performance on small size masses, we select all the masses smaller than 200 pixels and recorded the Dice coefficient achieved by state-of-the-arts and proposed SERAN. As shown in Fig.13, SERAN has achieved much better performance than other methods in most cases. It can be concluded that SERAN owns the ability to work in complex conditions.

Moreover, the visual results of all the cases in test set have been reviewed by an experienced radiologist. In summary, our work in this paper has achieved great improvement in breast mass segmentation and can assist the radiologists to a certain extent.

IV. CONCLUSION

In this paper, a Spatial Enhanced Rotation Aware Network (SERAN) is developed for breast mass segmentation using digital mammogram. Two main critical components are proposed for effective feature extraction and prediction refinement. An encoder with spatial attention enhancement under residual learning paradigm is designed for effective feature extraction. Spatial attention maps are explicitly modeled to adjust the focus in every encoder stage. Moreover, residual connection is utilized to avoid gradient vanishing and achieve better convergence. To boost the robustness of SERAN to masses displayed in different orientations, a decoder using multi-stream rotation aware mechanism for feature fusion and prediction refinement is designed. A $3 \times 3$ convolution and two asymmetric convolutions whose kernels are $1 \times 3$ and $3 \times 1$ respectively are combined in parallel to boost the rotation robustness. To avoid misclassification in background area and optimize SERAN for better prediction, a novel regulation item named Inside-outside Loss is applied in training procedure. Comparing with state-of-the-arts, SERAN has achieved significant performance improvement for breast mass segmentation. A sensitive of 87.7%, an IOU of 73.95%
and a Dice coefficient of 84.3% are achieved by SERAN on a representative subset of DDSM dataset. In future work, we will focus on developing SERAN to fit different types of medical image, such as ultrasound, CT and MRI. Moreover, we will try to transform 2D convolution by 3D convolution to adapt 3D scenes, such as 3D CT and 3D MRI. Besides, we will update the current system to fit different application scenes, such as breast cancer segmentation and prostate cancer segmentation.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA, Cancer J. Clin., vol. 68, no. 6, pp. 394–424, Sep. 2018.

[2] D. Selvathi and A. A. Poornila, “Deep learning techniques for breast cancer detection using medical image analysis,” in Biologically Ratio- nalized Computing Techniques For Image Processing Applications, 2018, Springer, pp. 159–186.

[3] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du, “Approaches for automated detection and classification of masses in mammograms,” Pattern Recognit., vol. 39, no. 4, pp. 646–668, Apr. 2006.

[4] M. L. Giger, H.-P. Chan, and J. Boone, “Anniversary paper: History and status of CAD and quantitative image analysis: The role of medical physics and AAPM,” Med. Phys., vol. 35, no. 12, pp. 5799–5820, Nov. 2008.

[5] O. Akin, S. B. Brennan, D. D. Dershaw, M. S. Ginsberg, M. J. Gollub, H. Schöder, D. M. Panicek, and H. Hricak, “Advances in oncoplastic imaging: Update on 5 common cancers,” CA, Cancer J. Clin., vol. 62, no. 6, pp. 364–393, Oct. 2012.

[6] H. Min, S. S. Chandra, N. Dhungel, S. Crozier, and A. P. Bradley, “Multi-scale mass segmentation for mammograms via cascaded random forests,” in Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI), Apr. 2017, pp. 113–117.

[7] D. Poap, M. Wozniak, R. Damasevicius, and W. Wei, “Chest radiographs segmentation by the use of nature-inspired algorithm for lung disease detection,” in Proc. IEEE Symp. Ser. Comput. Intell. (SSCI), NOv. 2018, pp. 2298–2303.

[8] M. Wozniak, D. Polap, G. Capizzi, G. L. Sciuto, L. Kośmider, and K. Frankiewicz, “Small lung nodules detection based on local variance analysis and probabilistic neural network,” Comput. Methods Programs Biomed., vol. 161, pp. 173–180, Jul. 2018.

[9] I. A. Ibaichir, R. Es-salhi, I. Daudsi, S. Tallal, and H. Medromi, “A survey on segmentation techniques of mammogram images,” in Advances in Ubiquitous Networking (Lecture Notes in Electrical Engineering), vol. 397, Springer, 2017, pp. 545–556.

[10] A. Mencattini, G. Rabottino, M. Salmeri, R. Lojacono, and E. Colini, “Breast mass segmentation in mammographic images by an effective region growing algorithm,” in Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst. Berlin, Germany: Springer, 2008, pp. 948–957.

[11] J. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in Proc. Eur. Syst. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 21–37.

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

[14] D. Creusan, A. Giusti, M. Gambardella, and J. Schmidhubner, “Deep neural networks segment neuronal membranes in electron microscopy images,” in Proc. Adv. Neural Inf. Process. Syst. 2012, pp. 2843–2851.

[15] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, May 2014.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 801–818.

[17] 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. 2881–2890.

[18] T. Liu, Y. Tian, S. Zhao, X. Huang, and Q. Wang, “Automatic whole heart segmentation using a two-stage U-Net framework and an adaptive threshold window,” IEEE Access, vol. 7, pp. 83628–83636, 2019.

[19] T. Hassanazadeh, L. G. C. Hamey, and K. Ho-Shon, “Convolutional neural networks for prostate magnetic resonance image segmentation,” IEEE Access, vol. 7, pp. 36748–36760, 2019.

[20] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, “Joint optic disc and cup segmentation based on multi-label deep network and polar transformation,” IEEE Trans. Med. Imag., vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[21] Y. Ding, F. Chen, Y. Zhao, Z. Wu, C. Zhang, and D. Wu, “A stacked multi-connection simple reducing net for brain tumor segmentation,” IEEE Access, vol. 7, pp. 104011–104024, 2019.

[22] S. Li, M. Dong, G. Du, and X. Mu, “Attention dense-U-Net for automatic breast mass segmentation in digital mammogram,” IEEE Access, vol. 7, pp. 59037–59047, 2019.

[23] J. Bai, K. Qiao, J. Chen, H. Tan, J. Xu, L. Zeng, D. Shi, and B. Yan, “Fully convolutional DenseNet with multiscale context for automated breast tumor segmentation,” J. Healthcare Eng., vol. 2019, Jan. 2019, Art. no. 8415485.

[24] H. Sun, C. Li, B. Liu, H. Zheng, D. Deng, and S. Wang, “AUNET: Attention-guided dense upsampling networks for breast mass segmentation in whole mammograms,” 2018, arXiv:1810.10151. [Online]. Available: http://arxiv.org/abs/1810.10151

[25] O. Oktay, J. Schlemper, L. El Fegoloo, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention U-Net: Learning where to look for the pancreas,” 2018, arXiv:1804.03999. [Online]. Available: http://arxiv.org/abs/1804.03999

[26] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3156–3164.

[27] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Trans. Syst., Man, Cybern., vol. 39, no. 4, pp. 59037–59047, 2019.

[28] Y. K. Jain and S. K. Bhandare, “Min max normalization based data transformation,” in Proc. Int. Conf. Mach. Learn., 2011, pp. 45–50.

[29] Y. C. Jan and S. K. Bhandare, “Min max normalization based data perturbation method for privacy protection,” Int. J. Comput. Commun. Technol., vol. 2, no. 8, pp. 45–50, 2011.

[30] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in Proc. Int. Conf. Int. Conf. Mach. Learn., 2010, pp. 807–814.

[31] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1–11.

[32] X. Ding, Y. Guo, G. Ding, and J. Han, “ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1911–1920.

[33] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980
YULIN CHENG received the bachelor’s degree in network engineering from the South China University of Technology, Guangzhou, China, in 2017, where he is currently pursuing the M.A.Eng. degree. His research interests focus on computer vision, including semantic segmentation, medical imaging analysis, and so on.

YING GAO received the bachelor’s and master’s degrees in computer science from the Central South University of China, in 1997 and 2000, respectively, and the Ph.D. degree in computer science from the South China University of Technology, China, in 2006. She is currently a Professor with the School of Computer Science and Engineering, South China University of Technology. She has published more than 30 articles in international journals and conferences. Her current research interests include computer vision, software architecture, and network security.

LINSEN XIE received the B.S. degree in computer science and technology from the South China University of Technology, Guangzhou, China, in 2019, where he is currently pursuing the M.S. degree in computer science and technology. His research interests focus on computer vision, including image processing, object detection, semantic segmentation, and so on.

XINYAN XIE received the bachelor’s degree in computer science from the South China University of Technology, Guangzhou, China, in 2019, where he is currently pursuing the M.A.Eng. degree. His research interests focus on computer vision and so on.

WENGEN LIN received the bachelor’s degree in computer science from the South China University of Technology, Guangzhou, China, in 2018, where he is currently pursuing the M.S. degree. His research interests include multimodality learning, person reidentification, action recognition, face recognition, and semantic segmentation.