IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 25, 2020, accepted July 1, 2020, date of publication July 6, 2020, date of current version July 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007512

Skin Lesion Segmentation Based on Multi-Scale Attention Convolutional Neural Network

YUN JIANG, SIMIN CAO, SHENGXIN TAO, AND HAI ZHANG
College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China

Corresponding author: Simin Cao (466503912@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61962054 and Grant 61163036, in part by the 2016 Gansu Provincial Science and Technology Plan funded by the Natural Science Foundation of China under Grant 1606RJZA047, in part by the 2012 Gansu Provincial University Fundamental Research Fund for Special Research Funds, in part by the Gansu Province Postgraduate Supervisor Program in Colleges and Universities under Grant 1201-16, and in part by the Northwest Normal University's Third Phase of Knowledge and Innovation Engineering Research Backbone Project under Grant nwnu-kjcxgc-03-67.

ABSTRACT The incidence of skin cancer around the world is increasing year by year. However, early diagnosis and treatment can greatly improve the survival rate of patients. Skin lesion boundary segmentation is essential to accurately locate lesion areas in dermatoscopic images. It is true that accurate segmentation of skin lesions is still challenging dues to problems such as blurred borders, which requires an accurate and automatic skin lesion segmentation method. In this paper, we propose an end-to-end framework which can perform skin lesion segmentation automatically and efficiently, called the CSARM-CNN (Channel & Spatial Attention Residual Module) model. Each CSARM block of the model combines channel attention and spatial attention to form a new attention module to enhance segmentation results. The multi-scale input images are obtained by the spatial pyramid pooling. Finally, a weighted cross-entropy loss function is used at each side of the output layer to sum the total loss of the model. We evaluated in two published standard datasets, ISIC 2017 and PH2, and achieved competitive results in terms of specificity and accuracy, with 99.03% and 99.45% specificity, 94.96% and 95.23% accuracy, respectively.

INDEX TERMS Deep convolutional neural network, multi-scale, attention mechanism, skin lesion segmentation.

I. INTRODUCTION

As the largest organ of the human body, skin is usually directly exposed to the air, which lead to skin diseases one of the most common diseases in humans [1]. According to statistics, there are 5.4 million new cases of skin cancer every year [2]. Melanoma, as one of the most lethal malignant skin tumors, causes more than 10,000 deaths each year [3]. However, melanoma can be cured by simple resection if it can be detected early. Early diagnostic survival rate exceeds 95% and late detection is below 20% [4]. Therefore, the early diagnosis and early treatment of dermatoses is very important.

Dermatoscopy is widely used in the non-invasive early detection of melanoma, and has a higher accuracy than the naked eye assessment. Nonetheless, it is not possible to rely solely on their perception and vision to detect melanomas correctly, even if an experienced dermatologist performs a dermoscopy [5]. Computer-aided analysis avoids many of these problems and is increasingly being studied to help dermatologists to improve the efficiency and objectivity of dermoscopy image analysis [6]. Automatic segmentation of skin lesions is an important step in computer-assisted dermoscopy image analysis [7], [8]. However, due to insufficient training samples and blurred boundaries of skin lesions, the lesions of different subjects shows significant differences in location, shape and color of interference with the segmentation task. In addition, a large number of artifacts including inherent skin characteristics (such as hair, blood vessels) and artificial artifacts (Such as air bubbles, ruler marks, uneven lighting, incomplete lesions, etc.) make the task of skin lesion segmentation extremely difficult [9].

In the early days, edge-detection, thresholding [10], active contouring [11], [12] or region-based techniques [13] were

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

used to segment lesions. But these methods usually cannot obtain accurate segmentation results. In recent years, convolutional neural networks have been widely used in medical image processing [14], [15], especially for medical image segmentation. These CNN-based methods can be classified by pixels to distinguish background objects from foreground objects to achieve the final segmentation. The U-net network proposed in [16] is specifically designed for biomedical image segmentation based on the concept of a fully convolutional network (FCN) [17]. Yu *et al.* [5] proposed a deep residual network with more than 50 layers for automatic skin lesion segmentation, in which several residual blocks [19] were stacked on top of each other to increase the representativeness of the model. In [20], Bi *et al.* proposed a multi-stage method to combine the output of a fully convolutional network cascaded at each stage to achieve the final skin lesion segmentation. Compared with earlier methods, convolutional neural networks have greatly improved the performance of image segmentation, but accurate segmentation of skin lesion boundaries still poses huge challenges. Therefore, in addition to these factors, we looked at another aspect of architectural design, the attention mechanism. Many works [21], [22] have proved that the attention mechanism can highlight the distinguished areas in the image, tell us the focus on attention, and have excellent positioning capabilities. Therefore, we propose to use the attention mechanism to improve presentation ability, focus on important features and suppress unnecessary features.

Inspired by the above semantic segmentation depth model and the latest advances in attention mechanism, we propose a new framework based on deep CNN, Channel & Spatial Attention Residual Module (CSARM) for automatic segmentation of skin lesions in dermoscopy images. Our specific work is as follows:

1) A new CSARM-CNN model is proposed to accurate skin lesion segmentation in dermoscopy images. The model adds a multi-scale input module, uses the convolutional attention module to extract image features, and updates parameters for training by using a multi-label loss function Model to generate the final segmentation map.
2) We have designed a novel attention learning module - CSARM. This module embeds both the convolution module and the attention module, which further improves the feature representation and can be widely used in the network to improve the performance of lesion segmentation. We verified the effectiveness of the CSARM block through ablation experiments.
3) On the ISIC-2017 dataset [43], the performance of the CSARM segmentation method is compared with other latest algorithms, with significant improvements in accuracy and specificity. The robustness of the algorithm trained by ISIC-2017 was tested for another publicly available dataset called PH2.

II. RELATED WORKS

A. SKIN LESION SEGMENTATION

The skin lesion segmentation task is used to detect the location and boundary of the lesion. In traditional image segmentation algorithms, skin lesion segmentation methods include threshold-based [10] region growth [13] segmentation methods and active contour-based [11], [12] methods. The advantages and disadvantages of each method have been discussed and compared with many papers such as [11], [12], [23]. In recent years, with the continuous development of deep learning, the segmentation method based on CNN[14] was first applied in the field of image segmentation and achieved significant [7], [9], [23]–[25] results in skin lesion segmentation.

Bi *et al.* [20] proposed a multi-stage fully convolutional network (FCN) method which uses parallel integration methods to combine the outputs of each stage to accurately segment skin lesions. Yuan *et al.* [7] developed an end-to-end DCNN with a loss based on Jaccard distance for skin lesion segmentation without prior knowledge and sample re-weighting. Li *et al.* [9] proposed a dense deconvolution network based on residual learning to segment skin lesions. The DSNet proposed by Hasan *et al.* [27] is an automatic dermatological semantic segmentation network, which uses depth-wise separable convolutions instead of standard convolutions to project the learned distinguishing features onto the pixel spatial at different stages of the encoder. These methods have achieved some results in skin lesion segmentation. However, there are still challenges to the task of boundary segmentation for low-contrast images.

B. OVERVIEW OF U-NET ARCHITECTURE

In recent years, CNN has shown broad prospects of medical image segmentation [26], [28], and most of the credit goes to U-Net [16]. U-Net is a neural network specifically designed for biomedical image segmentation based on the concept of a fully convolutional network. The U-net model learns deep features through different levels of convolution during the downsampling process. Then recovers the image size by deconvolution in the upsampling process. And finally outputs a feature map of the number of categories. After that, a large number of models are proposed based on the U-net architecture to improve the performance of biomedical image segmentation.

Fu *et al.* [29] proposed a multi-label deep network composed of multi-scale input layer, U-shaped convolutional network, side output layer and multi-label loss function, called M-Net.In the field of skin lesion segmentation, Hasan *et al.* [27] proposed an u-net-based automatic depth SLS model. In their model, the encoder network consists of extended residual layers, and a pyramid-merging network of three convolutional layers is used for decoder to enhance the ability of learning features. Ibtehaz *et al.* [30] analyzed the U-Net model architecture in depth, and proposed an enhanced version of the U-Net architecture - MultiResUNet.

C. ATTENTION MECHANISM

It is well known that attention plays an important role in human perception [31]–[33]. Human vision quickly scans the global image to obtain the target area that needs attention, and then invests more attention resources in this area to obtain more detailed information about the target that needs attention, thereby suppressing other unnecessary information. The attention mechanism in deep learning is similar to the selective visual attention mechanism of human beings in essence. The core goal is to select information which is more beneficial to the current task goal from a lot of information.

Earlier, the Google Deep Mind team was the first to use the attention mechanism on RNN models for image description problems [34]. Subsequently, they proposed a model for recognition of multiple objects in images based on the attention mechanism [35]. Recently, attention mechanisms have been successfully applied to various fields of computer vision and natural language processing to improve the performance of DCNN, and considerable progress has been made. And it has made great progress. Wang *et al.* [36] proposed a residual attention network using an encoder-decoder style attention module. By refining feature maps, the network not only performs well, but it is also robust to noisy inputs. Abraham *et al.* [37] Used improved attention U-Net for skin lesion segmentation and proposed a generalized focus loss function based on Tversky index to solve the problem of data imbalance in medical image segmentation. Kaul *et al.* [38] proposed an attention-based full convolutional network, FocusNet, which uses feature maps generated by separate convolutional autoencoders to focus attention on convolutional neural networks for medical image segmentation. Chen *et al.* [39] shared the attention of spatial and channel in the convolutional network. Spatial and channel attention weights are generated by neural networks followed by softmax layers, respectively. Although the above attention mechanism can effectively improve the performance of deep learning models in large-scale image segmentation tasks, the attention weight of these methods is learned by using other learnable layers with many additional parameters, which may not only lead to the computational costs but also produce overfitting for small training datasets. To alleviate this problem, Hu *et al.* [40] introduced a compact module to leverage relationships between channels. In their ''Squeeze and Excitation'' module, they use the global average pooling function to calculate channel attention. Subsequently, in addition to the channel, the CBAM module proposed by Woo *et al.* [41] introduced spatial attention in a manner similar to SE-Net [40]. They think that the importance of considering the pixels of different channels must also consider the importance of pixels in different positions of the same channel. Therefore, the attention map is inferred along two independent dimensions (channel and space), and then the attention map is multiplied by the input feature map for adaptive feature improvement. Although the lightweight model effectively alleviates the problem of heavy computation, these methods only use rescaling for feature fusion, which is not effective enough for global context modeling.

In response to the above problems, our proposed CSARM-CNN can effectively improve the balance between data redundancy caused by extra parameters and weak context. A large number of ablation experiments and comparative experiments were conducted to demonstrate the effectiveness of the proposed model in the skin segmentation task.

III. PROPOSED CSARM-CNN ARCHITECTURE

In this paper, the proposed CSARM-CNN model is an end-to-end multi-label depth network composed of multiple CSARM blocks, multi-scale input layer, U-shaped convolutional network and side output layers. In each CSARM block, the residual learning mechanism is used to solve the degradation problem, and the channel attention module and the spatial attention module are combined to design a new attention mechanism to enhance the ability of discriminative expression. Multi-scale input layer constructs image pyramid to realize multi-level receptive field. U-Net is used as the main network structure to learn rich hierarchical representations. The side output layer is used as an early classifier to generate the accompanying local prediction maps of different scale layers. Finally, a multi-label loss function is added to update the parameters to train the model and generate the final segmentation map. The architecture of this mode is shown in Fig.1.

A. CONVOLUTIONAL ATTENTION MODULE

In order to obtain good segmentation performance, we propose the CSARM-CNN model, which embeds residual learning and attention learning mechanisms. From an implementation perspective, both residual learning and attention learning can be embedded in each CSARM block. We can build a CSARM-CNN model with an arbitrary depth by stacking multiple CSARM blocks and train it in an end-to-end manner. Therefore, the architecture of the CSARM block is the basic module of the model, where each CSARM block is stacked by a fixed-mode of convolutional layers, channel and spatial attention module, residual learning, and attention learning mechanism. Convolution extracts image information. Each attention module uses bottom-up and top-down feed-forward structure trainable layers to learn weights, and then multiplies the weights with the convolutional features. The residual mechanism effectively alleviates the model degradation problem. The CSARM block is shown in Fig.2. Next, we will introduce the internal structure of the CSARM block in detail.

1) CHANNEL & SPATIAL ATTENTION LEARNING

Given an input image, the channel attention module aims to focus more on the meaningful parts by establishing the association between channels, enhancing the channel's specific semantic response capability, and emphasizing it. As a supplement to channel attention, the spatial attention module

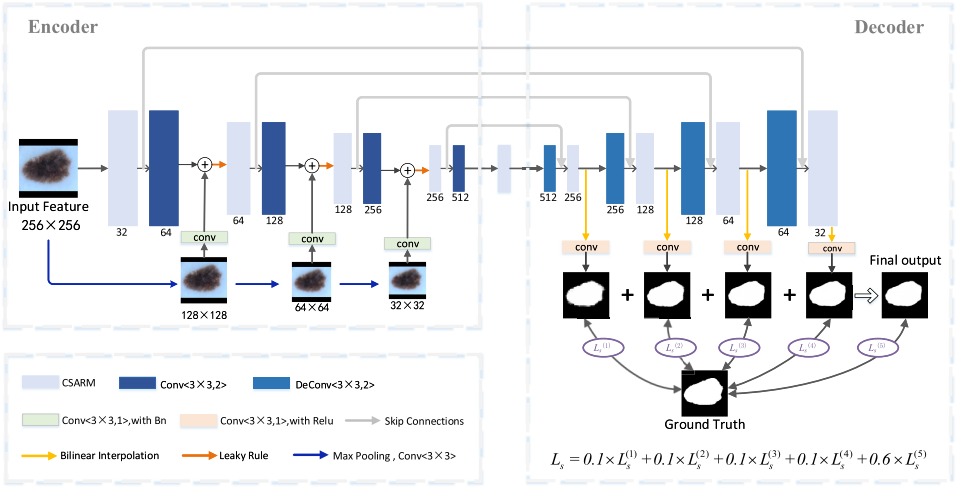

FIGURE 1. The architecture of the CSARM-CNN model.

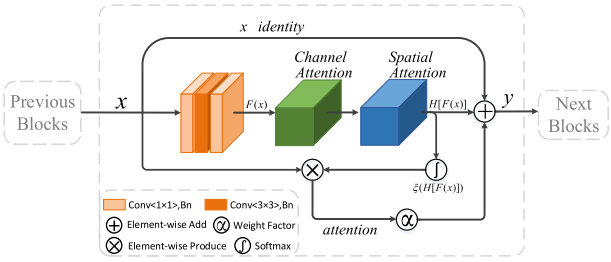

FIGURE 2. CSARM module structure.

aims to use the association between any two point features to mutually enhance the expression of their respective features, so it pays more attention to the features of the spatial position. Two attention sub-modules are shown in Fig.3. In order to further to obtain the characteristics of the global dependency relationship, the output results of the two modules are added and fused to obtain the final features for pixel classification.

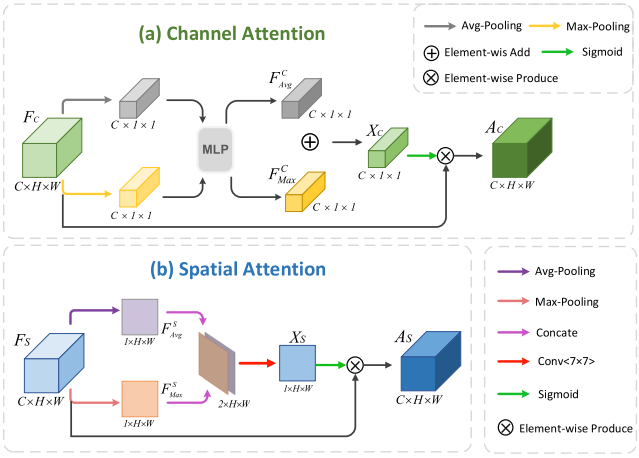

FIGURE 3. The Attention module. (a) Channel attention module. MLP: Multi-layer perceptron. (b) Spatial attention module.

In the channel attention module, since each channel of the feature map is considered as a feature detector [42]. Therefore, we use the channel attention module to build the interdependence between channels. Using the dependency relationship between feature channels, the feature representation of specific semantics can be improved to generate channel attention maps. The input feature $F_C \in R^{\times C\times H\times W}$ is used to aggregate the spatial information of the feature maps by using average pooling and maximum pooling operations to generate average pooling features and maximum pooling features. The multi-layer perceptron (*MLP*) consists of the Fc1 layer, the *Relu* activation function and the Fc2 layer. It can learn from the given training data and make accurate predictions based on the new data given. When two pooling operations are used at the same time, both of aggregated channel features are located in the same semantic embedding space. Therefore, we use a shared *MLP* to perform attention inference to save parameters and obtain the correlation strength of the two channels, and output feature maps $F^C_{\text{Avg}} \in R^{\times 1\times 1\times 1}$ and $F^C_{\text{Max}} \in R^{\times 1\times 1\times 1}$ respectively. Then use the element-by-element summation through the attention graph between the channels, so that each channel can produce a global correlation to further enhance the feature representation, and merge the output feature map X_C, as shown in formula (1), and then go through the sigmoid operation obtain the final output channel attention feature map A_C, as shown in formula (2).

$$X_C = F^C_{\text{Arg}} \oplus F^C_{\text{Max}}, \quad X_C \in R^{C\times 1\times 1} \tag{1}$$

where, $\oplus$ means adding element by element.

$$A_C = \bigcup_{i=0}^{c} \frac{1}{1+\exp\left(-X_{i,1,1}\right)}, \quad A_C \in R^{C\times 1\times 1} \tag{2}$$

where C represents the number of channels, $X_{i,1,1}$ represents the element whose coordinate is $C\times 1\times 1$, and $\bigcup$ represents the contact element by element.

The spatial attention module encodes extensive contextual information into local features, thus enhancing their expressive power. We use the spatial relationships among the elements to generate the spatial attention maps. For features at a certain location, they can be updated by using weighted summation to aggregate the features at all locations, so regardless of their distance in the spatial dimension, features at any two locations with similar features can promote each other. In order to calculate spatial attention, we use the average pooling and maximum pooling operations of the original feature $F_S \in R^{\times C\times H\times W}$ along the channel axis to summarize the channel information of the feature map to generate two 2-dimensional feature maps F^S_{Avg} and F^S_{Max} (the specific process is shown in formula (4) and (5)).And then perform weighted fusion on the features, so that for the points at various positions, it fuses similar features in the global space through the attention map. Then the convolutional layer is used to generate 2D space attention map X_S as:

$$X_S = \mathrm{Conv}\left(\mathrm{Cat}\left(F^s_{\mathrm{Avg}}; F^S_{\mathrm{Max}}\right) W + b\right), \quad A_S \in R^{\times 1\times H\times W} \tag{3}$$

Among them, W, b respectively represent *MLP* weight, *MLP* biase; *Cat*(.) Represents concatenate; *Conv*(.) Represents convolution operation.

$$F^s_{\mathrm{Avg}}(x,y) = \frac{1}{C}\sum_{i=1}^{c} X_{i,x,y}, \quad F^S_{Avg} \in R^{\times 1\times H\times W} \tag{4}$$

$$F^s_{\mathrm{Max}}(x,y) = \max_{i=0}^{c}\left(X_{i,x,y}\right), \quad F^S_{Max} \in R^{\times 1\times H\times W} \tag{5}$$

Finally, after the sigmoid operation, the spatial attention feature map A_S obtained by the final output as follows:

$$A_s = \bigcup_{m}^{H}\bigcup_{n}^{W} \frac{1}{1+\exp\left(-X_{Sm,n}\right)}, \quad A_S \in R^{\times 1\times H\times W} \tag{6}$$

where, m, n represent m^{th} position and n^{th} position respectively. $\bigcup$ represents the contact element by element.

It should be noted that for feature images, average pooling is usually used to summarize spatial information, and maximum pooling can be used to infer more detailed channel attention. Therefore, in both modules, we have chosen to use both average pooling and maximum pooling functions.The channel sub-module uses the maximum pool output and the average pool output of the shared network. The spatial submodule uses similar two outputs, which are pooled along the channel axis and transmitted to the convolutional layer.

2) CSARM BLOCK

The traditional attention mechanism strengthens attention learning by using additional learnable layers, such as the convolutional layers used in [36] or the fully connected layers used in [40]. The CBAM proposed in [41] divides the attention process into two independent parts, the channel attention module (look what) and the spatial attention module (look where). These lightweight models all use the attention mechanism to enhance their own feature extraction ability to a certain extent, but they do not fully utilize the global context information. Different from these solutions, we propose that the proposed CSARM block can effectively model the global context by addition fusion while having the properties of the lightweight model. In addition, the experimental part of IV.B.2) and IV.B.3) also showed that CSARM block is superior to other attention modules in the skin lesion segmentation task.

The structure of the CSARM module is shown in Fig.2. Firstly, three consecutive convolutional modules are used to extract image information, and the channel & spatial attention module is used to recalibrates the importance of different spatial positions and channels for feature fusion.

Secondly, the designed attention learning mechanism generates attention through the network itself to strengthen the discriminative representation of the network. We believe that the higher layers of the network have stronger effective information than the lower layers. Therefore, we use the higher layer to generate abstract feature maps as the lower-level attention features. In this way, the discriminability of the network is enhanced by generating attention by the network itself, without introducing additional learnable layers. Suppose a set of stacked layers of $\{L_1 \ldots L_n\}$, where L_1 represents the lower layer, L_n represents the upper layer, and $E \in R^{\times 1\times H\times W}$ is the input feature of L_n,To avoid overfitting, we use the Softmax function to batch normalization E, The normalization function is defined as:

$$\xi(E) = \left\{ i|i^C_{m,n} = \frac{e^{E^C_{m,n}}}{\sum_{m',n'} \cdot e^{E^C_{m',n'}}} \right\} \tag{7}$$

where m, n represent the spatial position, and C represents the channel index of the feature map.

In a CSARM module, we used the identity mapping designed in the residual block. The input feature is x and the feature map $F(x)$ is obtained through residual mapping. For channels and spatial modules, it was pointed out in [40] that sequential permutation can provide better results than parallelism. Therefore, we also adopted the method of channel permutation followed by spatial attention, as shown in Fig.3. $F(x)$ obtains the feature maps $H[F(x)]$ through the channel and spatial attention modules. We use s to normalize the feature maps $H[F(x)]$, and then obtain the attention feature map through the element-wise production as:

$$A = \xi(H[F(x)]) \cdot x \tag{8}$$

In order to further obtain the characteristics of global dependency, the output results of the three modules are added and fused. The final output y of CSARM block is the sum of the elements of the identity map, the residual feature map and the attention feature map, and its calculation formula expressed is as follows:

$$y = x + F(x) + \alpha \cdot A \tag{9}$$

where α is a learnable weighting factor, indicating the relationship between the attention feature map and the other two maps.

B. CSARM-CNN

In this paper, the proposed CSARM-CNN model is shown in Fig.1, including the encoder path (left), decoder path (right). The skip connections transfer the corresponding feature map from encoder path and concatenate them to up-sampled decoder feature maps. Table1 shows the architecture of the CSARM-CNN model. The model uses CSARM blocks to extract image features. Each CSARM block is stacked by a fixed-mode 1×1, 3×3, 1×1 convolutional layer, channel attention module, and spatial attention module.

1) THE ENCODER

The encoder path uses CSARM blocks to extract feature information, and uses a 3×3 convolution with a step size of 2 to replace the pooling operation in downsampling. To avoid overfitting, the batch normalization layer is used to normalize the feature map of each layer after the convolution operation of each layer, and then it is activated using the Leaky Relu activation function. In the encoder path, we build a multi-scale input layer. Multi-scale image technology, also known as Multi-Resolution Analysis (MRA), refers to the use of multi-scale representation of images and processing at different scales. But the multi-scale input method we use is different from the traditional method of using multi-scale image input and then fusing the results. We establish multi-scale input in the encoding path of the image for the down-sampling process, and then input the image on the feature map in a multi-scale manner, and encode the multi-scale context information. As shown in Fig.1. Given an original image with an input size of 256×256, three different sizes of 128×128, 64×64, and 32×32 images were obtained after three down-sampling processes, and combined with the original image to construct an image pyramid input and achieve multi-level reception field fusion. Adding multi-scale input to the coding layer can ensure the transmission of the original image features, improve the quality of the segmentation, and at the same time, increase the network width of the decoder path to avoid a large increase in parameters [31].

2) THE DECODER

The decoder path is the exact opposite of the encoder path. As shown in Fig.1, in the process of up-sampling, we first used the 3×3 deconvolution layer with a step size of 2, batch standardization and a CSARM block output decoder feature map for each layer. After that, the output feature maps of each layer of CSARM blocks are extracted, and the feature maps are expanded to the size of the original input image using bilinear interpolation, and then input them into the classifier for classification. The classifier consists of a 3×3 convolution with a step size of 2 and then a softmax function. Due to the skin lesion segmentation task, the two-channel probability map of lesion and background is output. Therefore, the designed classifier is to convert each layer of multi-channel feature maps into two-channel feature maps. Finally, the probability maps obtained by different classifiers are fused into the final classification result to complete multi-scale feature fusion.

In this process, in order to alleviate the problem of gradient disappearance and enhance the training of the early layer, the decoder path receives the output loss from the back-propagation of the output layer and updates the parameters. We use the cross-entropy loss function to calculate the output loss for each layer of output images. For the sample (x, y), $x = \{x_i, i = 1, \ldots, N\}$ represents the training data and $y = \{y_i, i = 1, \ldots, N\}$ is the corresponding ground truth.Among them $y_i = \{0, 1\}$, the probability that the i-th sample is predicted as 1 is y_p. N represents the total number of samples. M represents the number of multi-scale output layers. At this time $M = 5$, the corresponding loss weight of each multi-output layer is expressed as $a_i = \{y_i, i = 1, \ldots, M\}$, And $ai = \{0.1, 0.1, 0.1, 0.1, 0.6\}$. For each output image, the loss L. For each output image, the loss L is defined as (10):

$$L_{log}(Y, P) = -\frac{a_i}{N} \sum_{x} \left[y_{i.} \log \left(y_p \right) + (1 - y_i) \cdot \log \left(1 - y_p \right) \right] \tag{10}$$

We overlay the $L(N)$ of each output layer. The final output loss function L is:

$$L = \sum_{i=0}^{N \cdot I} L^{(i)}(Y, P) \tag{11}$$

IV. EXPERIMENTS

A. EXPERIMENTAL SETUPS

1) DATASET

We used two public dermoscopy datasets to evaluate the proposed segmentation network, the ISIC-2017 challenge dataset [43] and the PH2 dataset [44]. The ISIC-2017 challenge dataset is provided by the International Skin Image Collaboration (ISIC) archive. The challenge dataset contains 8-bit RGB dermoscopy images with image sizes ranging from 540×722 to 4499×6748 pixels. It provides 2,000 training images and separate datasets of 150 and 600 images for validation and testing, respectively. All these dermoscopy images were marked as Benign Nevus, Melanoma or Seborrheic Keratosis, respectively. The PH2 dataset was collected by the dermatology department of the Pedro Hispano Hospital and the research team of the University of Porto in TÍecnico Lisboa, Matosinhos, Portugal. It contains 200 images, of which 160 are Nevus (ie ordinary Nevus and atypical Nevus), and the rest Forty images are melanoma.These images are 8-bit RGB images with a fixed size of 768×560 pixels, and are acquired using 20x magnification under the same conditions. Table 2 summarizes the distribution of the two datasets. Besides, both datasets provide original images paired with lesion segmentation boundaries, which were annotated by a professional dermatologist.

TABLE 1. CSARM_CNN model architecture and implementation details.

Layer Name	Down sample			Up Sample		
	Input Size	Output Size	Structure	Input Size	Output Size	Structure
Layer_1	$256 \times 256 \times 3$	$128 \times 128 \times 64$	CSARM[3,32,$\alpha = 0.5$] Conv[3×3,64,s=2] Bn,Leaky Relu CSARM[64,64,$\alpha = 0.5$]	$128 \times 128 \times 64$	$256 \times 256 \times 2$	DeConv[64,32,s=2,p=1] Bn,Leaky Relu CSARM[32,2,$\alpha = 0.5$]
Layer_2	$128 \times 128 \times 64$	$64 \times 64 \times 128$	Conv[3×3,128,s=2] Bn,Leaky Relu CSARM[128,128,$\alpha = 0.5$]	$64 \times 64 \times 128$	$128 \times 128 \times 64$	DeConv[128,64,s=2,p=1] Bn,Leaky Relu CSARM[64,64,$\alpha = 0.5$]
Layer_3	$64 \times 64 \times 128$	$32 \times 32 \times 256$	Conv[3×3,256,s=2] Bn,Leaky Relu CSARM[256,256,$\alpha = 0.5$]	$32 \times 32 \times 256$	$64 \times 64 \times 128$	DeConv[256,128,s=2,p=1] Bn,Leaky Relu CSARM[128,128,$\alpha = 0.5$]
Layer_4	$32 \times 32 \times 256$	$16 \times 16 \times 512$	Conv[3×3,512,s=2] Bn,Leaky Relu	$16 \times 16 \times 512$	$32 \times 32 \times 256$	DeConv[512,256,s=2,p=1] Bn,Leaky Relu CSARM[256,256,$\alpha = 0.5$]
Bottom Layer	$16 \times 16 \times 512$	$16 \times 16 \times 521$	CSARM	-	-	-

TABLE 2. Distribution of the ISIC Challenge 2017 [42] and PH2 datasets [43].

Datasets	ISIC 2017				PH2			
	Me	SK	Ne	Total	Me	SK	Ne	Total
Training data	404	296	1450	2150	-	-	-	-
Test data	117	90	393	600	40	-	160	200

$*$ Ne, Me and SK represent Benign Nevus, Melanoma and Seborrheic Keratosis, respectively.

2) PREPROCESSING

In deep learning methods, training usually requires a large amount of data. Therefore, we apply a data augmentation process to expand the training dataset. First, the 2000 training data on the ISIC-2017 dataset was combined with 150 validation data to generate 2150 training datasets. Secondly, a pre-processing program is provided to augment and facilitate the learning of our proposed segmentation method from the training dataset. In the ISIC-2017 dataset, the 2150 dermoscopy images were generated by the combination all exist in RGB form. In order to understand the different color space characteristics, we also added three channels of hue saturation value (HSV). 2,150 dermoscopy images in HSV format were generated, and the resulting HSV images are shown in Fig. 4(b). Samples were randomly generated by horizontal flipping, vertical flipping and horizontal vertical flipping, so that the training data set has 17,200 dermoscopy images. The generated samples are shown in Fig. 4(c), and the upper right of each sample is horizontal flipping, the lower left is vertical flipping, and the lower right is horizontal and vertical flipping. Since the size of the original image ranges from 540×722 to 4499×6748 pixels, in order to facilitate the training of the segmentation network, as shown in the upper right corner of Fig. 4(c), we adjusted the width of the image to 256px according to the aspect ratio of the original image. The image is filled with black edges up and down, and the height is increased to 256px to form a training image. Figure 4 shows examples of dermoscopy images from the ISIC 2017 Challenge dataset, ground truth, and pre-processed images. The first two columns are Melanoma, the middle two columns are Seborrheic Keratosis, and the last two columns are Nevus.

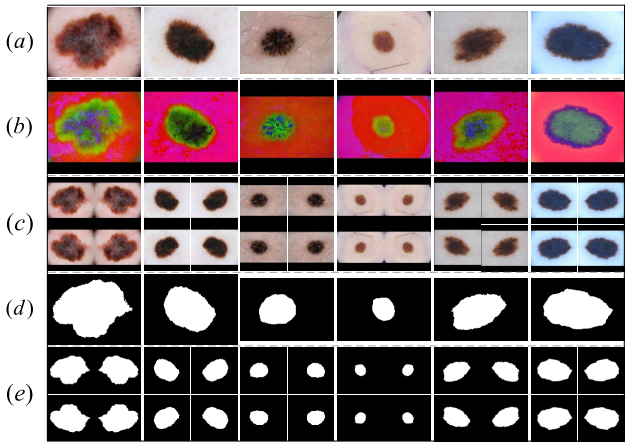

FIGURE 4. Preprocessing visualization results on the ISIC 2017 dataset. (a) The original image (b) Image generated after HSV image preprocessing (c) The image generated after the original image is preprocessed (d) the ground truth (e) the ground truth processed image.

3) EVALUATION METRICS

In order to quantitatively evaluate the segmentation capability of the proposed CSARM-CNN network, we used the following skin lesion segmentation evaluation indicators. Sensitivity (SEN) is defined in eq.(12), which represents the proportion of skin lesion pixels that are correctly segmented. And high sensitivity (close to 1.0) indicates that the segmentation effect is good. Specificity (SPE) (as in eq.(13)) indicates the proportion of non-lesioned skin pixels which are not correctly segmented. High specificity indicates the ability of this method to segment non-lesionable pixels. Jaccard index (JAC) and Dice coefficient (DIC) are used to measure the similarity between the segmented lesion and the annotated

ground truth, as in eq.(15) and (16) respectively. Accuracy (ACC) (as in eq.(14)) is also provided to show overall pixel-level segmentation performance. The Matthew correlation coefficient (MCC) is used to measure the correlation between annotated and segmented skin lesion pixels, as in eq.(17). All of these indicators are calculated based on the elements of the confusion matrix.

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Specificity = \frac{TN}{FP + TN} \tag{13}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{14}$$

$$JAC = \frac{TP}{TP + FN + FP} \tag{15}$$

$$DIC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{16}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{17}$$

Among them, TP indicates that the skin lesion pixels are correctly segmented, and FN indicates that the skin lesion pixels are incorrectly segmented. In contrast, if the segmentation of non-lesioned pixels is correctly classified as non-lesioned, it is considered as TN. Otherwise, they are FP.

4) IMPLEMENTATION DETAILS

The hardware equipment used in this experiment is Intel Xeon (R) CPU E5-2620 v3 2.40GHz, NVIDIA Tesla K80 (12G). All training and testing are performed in the same hardware environment. The operating system used in the experiment is Ubuntu 16.04, using Python 3.6 as the programming language, and using Facebook's open-source Pytorch 1.0.0 deep learning framework for algorithm design and coding. The network uses the Stochastic gradient descent method of Nesterov momentum for end-to-end training. The momentum parameter is set to 0.9, the weight attenuation coefficient is 5e-4, and the mini-batch size is 8. The multi-output loss function is used to calculate the loss value. The backpropagation algorithm is used to propagate the loss layer by layer and update the network parameters, and the Softmax function is used for final classification. We train for 200 epochs, the initial learning rate is set to 0.1, and the learning rate is reduced by 5 times in turn after 80, 120, and 160 epochs.

B. ABLATION EXPERIMENT

1) NETWORK ABLATION EXPERIMENT

The proposed CSARM-CNN model shows excellent performance in skin lesion segmentation. We believe that the improvement in performance is mainly due to the use of the CSARM attention module, which allows DCNN to focus more on the semantically meaningful part of the lesion, thereby enhancing the ability of the network to learn discriminative representation. To validate this, we performed an ablation experiment on the proposed model using 600 dermoscopy image tests on the ISIC-2017 dataset. We treat the model after removing the CSARM block as the baseline model, and compare it with the proposed segmentation performance of CSARM-CNN and the baseline model. At the same time, in order to verify the generality of the CSARM block, the three networks U-Net, FCN and SegNet are used as backbone networks. For ease of description, in this paper, we replace a convolutional module in each layer of the network with a CSARM block and the network is called its new attention network, which is respectively named as "U-Net+CSARM", "FCN+CSARM" and "SegNet+CSARM". Table 3 and 4 summarize the segmentation performance comparison of the three infrastructures with their new attention network and the proposed CSARM-CNN with the baseline model. The experimental results show that for different network architectures, the six evaluation metrics of the new attention network are much higher than those of the backbone network model. Therefore, it is proved that the CSARM block can effectively improve the performance of the model. At the same time, the CSARM block can be applied to different network architectures and has good robustness.

The results of the experiments on U-Net, FCN and SegNet in this paper are slightly lower than those of Al-Masni *et al.* [48] and Goyal *et al.* [49], whose experimental results refer to Tables 9 and 10 for details. The reason for the difference may be that our training parameters are different from their settings, as follows: (1) Al-Masni M A *et al.* used the pre-training weights trained on the VGG-16 network layer of the large public ImageNet dataset as the initial weights of FCN and SegNet, and then retrained and fine-tuned on the ISIC2017 dataset, and we reproduced U-Net, FCN and SegNet only set random initial weights; (2) Al-Masni M A and others were batch set to 20, trained with NVIDIA GeForce GTX 1080 (16G) GPU, and implemented with Theano and Keras deep learning libraries and AdamOptimizer optimization algorithm in Python 2.7.14. However, we set the batch to 8, trained on NVIDIA Tesla K80 (12G)GPU, adopted Python3.6 as the programming language, designed and coded the algorithm using Pytorch 1.0.0 deep learning framework, and conducted end-to-end training using Nesterov momentum Stochastic gradient descent method.

In order to show the segmentation effect of the lesion more clearly, we compared the true segmentation profiles of Benign, Melanoma and Seborrheic Keratotic lesions in the ISIC-2017 test data set. As shown in Fig.5, the segmentation results of the three networks verified by the ablation experiment and its new attention network, CSARM-CNN and baseline model are visualized. As for Fig.5(f), we enlarged the lesion area on the image, in order to visually show the comparison between the segmentation results of the four models of CSARM module and the ground truth. In Fig.5, the first two lines are Melanoma lesions, the middle two lines are Seborrheic Keratotic lesions, and the last two lines are

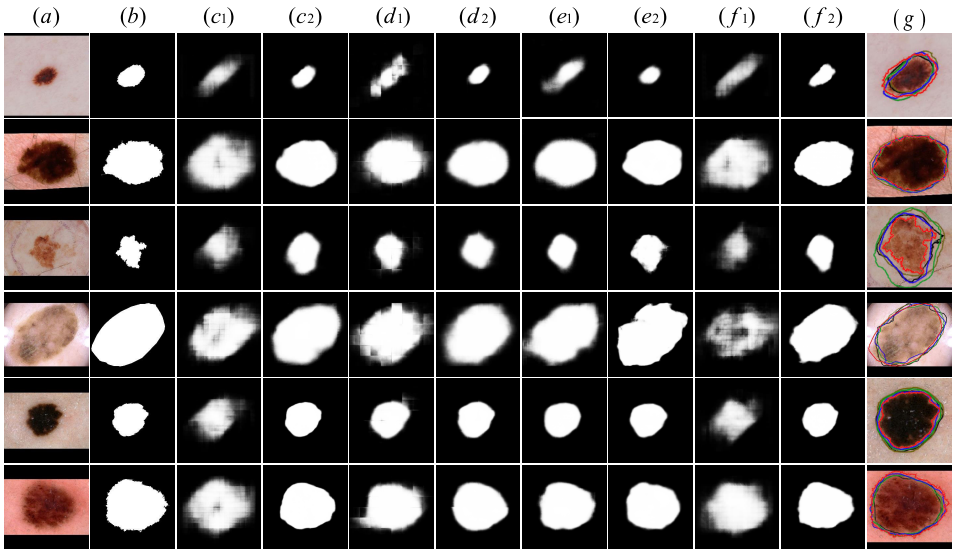

FIGURE 5. Visualization results of an ablation experiment. (a) The original image (b)The ground truth (c_1) U-Net (c_2) U-Net+CSARM (d_1) FCN (d_2) FCN+CSARM (e_1) SegNet (e_2) SegNet+CSARM (f_1) baseline model (f_2) CSARM-CNN model (g) Comparison of segmentation results of ground truth (red) and U-Net+CSARM (green), FCN+CSARM (yellow), SegNet+CSARM(black) and ours (blue). All images are pre-processed.

Nevus. It can be clearly seen in the figure that compared with the three backbone networks and baseline models, its new attention network and our model have a clearer boundary, and can identify the lesion area in the dermoscopy image. Therefore, adding the CSARM block we proposed to the network can greatly improve the performance of lesion segmentation.

2) ATTENTION MODULE ABLATION EXPERIMENT

The CSARM block is the main part that improves the segmentation performance. To illustrate the difference between the CSARM block and other attention methods, we designed a set of ablation experiments based on different attention modules. Using the baseline network as a quantitative model, we respectively selected two typical attention modules that can be embedded in other models and compared them with our CSARM attention module. Among them, the first is the Squeeze-and-Excitation attention module of SE-Net that is often used in classification tasks, which we call SE Block. It uses the interdependence between convolutional feature channels in the network to improve the network's presentation ability. Another Attention module is the Feature Pyramid Attention (FPA) module applied in the PAN of segmentation task, which uses the attention mechanism to introduce the global context information as the prior knowledge into the channel selection. Meanwhile, through the spatial pyramid attention structure, the multi-scale information is fused to produce better pixel-level attention.

In Table 5 and 6, we compared the segmentation performance of three skin lesions based on the Baseline model of SE Block and FPA with the CSARM-CNN model. Although the two attention modules have improved the performance of the model to a certain extent, from the six evaluation indicators, the overall segmentation result of CSARM-CNN is much higher than the results of the two attention modules embedded in the Baseline network. SE Block contains two operations. Squeeze and Excitation. The Squeeze operation first uses global average pooling to obtain channel-level global features. Then perform Excitation operation on the global features. Two Fully Connected layers form a Bottleneck structure to model the correlation between channels, and output the same number of weights as the input features. However, this method only focuses on which layers on the channel level will have stronger feedback capabilities. It does not reflect the region of interest in the spatial dimension, and the global context information is not fully utilized. The FPA module adopts the idea of PSPnet's global-pooling. The result of pooling is added to the result of the attention convolution, and the global context information is introduced into the channel selection as a priori knowledge. However, the use of corresponding channel attention vectors in such a structure is not enough to effectively extract features of multiple scales and lack pixel-level information. These two attention models construct the correlation between channels to the fusion of feature channels and ignore more spatial feature information.

However, for the skin lesion image, the gray scale changes little, and the boundary is relatively blurred. And through the stepwise pooling, the spatial resolution of the features continues to decrease, and the spatial position information is continuously lost. Ignoring the spatial feature information can easily lead to blurring of the edge of model segmentation results. Therefore, features with rich spatial location information are particularly important for restoring feature spatial resolution. The CSARM block we proposed jointly uses residual learning and channel and spatial attention learning

TABLE 3. On the ISIC2017 dataset. Comparison of sensitivity, specificity and accuracy performance of the three infrastructure networks and their new attention networks with the CSARM-CNN model and the baseline model.

Method	Nevus cases			Melanoma cases			SK cases			Overall		
	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC
U-Net[16]	81.80	97.53	93.83	70.31	98.08	89.62	57.84	98.34	89.07	74.53	97.59	92.32
U-Net+CSARM	87.64	99.12	96.49	73.03	98.69	91.84	85.61	99.14	93.12	86.33	98.97	95.32
FCN[17]	85.83	99.14	95.98	72.85	98.00	90.50	66.28	98.03	89.62	74.78	98.67	93.76
FCN+CSARM	**89.18**	99.06	96.32	77.68	99.14	93.09	**76.05**	99.35	94.11	83.59	99.20	95.29
SegNet[45]	82.83	98.65	95.60	73.24	99.02	91.69	65.62	98.98	92.64	80.19	99.00	95.46
SegNet+CSARM	85.75	98.88	96.71	72.57	99.36	92.20	72.07	99.05	92.65	**81.94**	**99.95**	95.74
Baseline	78.53	97.25	93.60	60.43	97.70	88.02	60.30	97.88	89.59	72.12	97.17	91.92
Ours	87.13	**99.36**	**97.30**	**75.70**	**99.39**	**94.30**	72.94	**99.55**	**94.81**	80.22	99.40	**95.85**

TABLE 4. On the ISIC2017 dataset. Comparison of Dice coefficient, Jaccard index, Matthew correlation coefficient performance of the three infrastructure networks and their new attention networks with the CSARM-CNN model and the baseline model.

Method	Nevus cases			Melanoma cases			SK cases			Overall		
	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC
U-Net[16]	78.71	64.89	75.12	73.16	57.68	66.73	64.39	47.49	58.55	74.32	59.13	69.64
U-Net+CSARM	86.80	76.68	84.86	78.71	64.89	74.32	77.52	63.29	73.48	84.08	72.83	81.57
FCN[17]	84.91	73.78	82.61	76.93	62.51	71.15	70.68	54.67	65.36	79.04	65.34	75.64
FCN+CSARM	86.94	76.90	84.83	82.05	69.56	78.06	**78.23**	**64.99**	**74.94**	83.54	71.74	80.89
SegNet[45]	83.66	71.91	81.28	77.84	63.73	73.28	73.67	58.31	70.15	83.07	71.05	80.70
SegNet+CSARM	87.15	77.23	85.29	79.80	66.39	75.60	75.38	60.49	71.56	84.06	72.80	81.71
Baseline	76.13	61.74	72.55	68.62	52.23	62.25	67.74	51.22	62.32	72.63	57.02	67.72
Ours	**89.72**	**81.35**	**88.19**	**82.29**	**69.91**	**79.09**	77.18	62.84	74.89	**84.62**	**73.35**	**82.32**

mechanisms to improve its ability to discern representations. The attention learning mechanism uses feature maps learned by high layers to generate low-level attention maps, which effectively improves segmentation performance while avoiding the computational burden caused by too many parameters.

Fig. 6 shows six examples of dermoscopy images and corresponding segmentation masks to the different attention modules. The first two columns are Melanoma, the middle two columns are Seborrheic Keratosis,and the last two columns are Nevus. It can be clearly seen from the segmentation results in Fig.6 that our method has clearer segmentation boundary information, making the network more focused on the skin lesion area. In order to more intuitively show the segmentation results between attention modules and the comparison with the ground truth, we superimposed all segmentation results on the example dermoscopy images and enlarged the lesion areas. Obviously, compared with the other two modules, the blue line representing CSARM-CNN is closer to the red line representing the ground truth, proving that on the baseline model, CSARM Block has better feature presentation ability than SE Block and FPA.

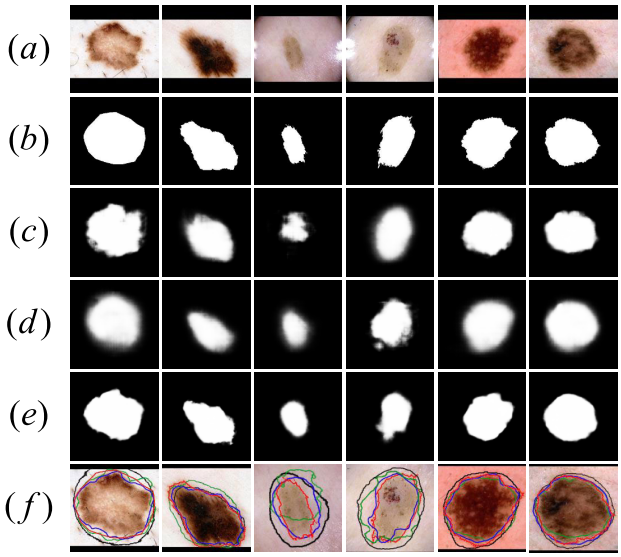

FIGURE 6. Visualization results of the attention module comparison. (a) Original image of skin lesions (b) The ground truth corresponding to the lesion (c) Lesion segmentation results of SE Block+Baseline (d) Lesion segmentation results of FPA+Baseline (e) Lesion segmentation results of CSARM-CNN model (f) Comparison of segmentation results of the ground truth (red) and SEnet+Baseline (green), FPA+Baseline (black) and ours (blue). All images are pre-processed.

3) STRUCTURAL ABLATION EXPERIMENT

Woo *et al.* [41] believed that SENet only focused on which layers at the channel level would have stronger feedback capability, but failed to reflect the region of interest in the spatial dimension. Therefore, the designed CBAM applies attention to both the channel and spatial dimensions, and obtains more spatial feature information than SE Block, which can effectively improve the representation of the model without significantly increasing the model parameters and computation (It can also be proved by the experimental results of the two models in Table.5-8). However, CBAM still has the problem that it cannot obtain effective global context information.

In order to solve the problems existing in CBAM module and fuse the context information more effectively, we have improved on the basis of CBAM module. We use CBAM Block as part of the internal structure of the CSARM model to extract attention features on spatial and channels. In order to

TABLE 5. On the ISIC2017 dataset. Comparison of Sensitivity, Specificity, and Accuracy performance between different attention modules and CSARM blocks using baseline as the infrastructure.

Method	Nevus cases			Melanoma cases			SK cases			Overall		
	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC
SE Block+Baseline	84.29	99.28	96.07	72.62	98.74	92.36	65.03	98.77	91.60	**80.89**	98.73	94.97
FPA+Baseline	90.57	99.29	96.58	74.19	99.06	93.52	71.19	98.65	92.32	80.64	99.07	95.06
Baseline	78.53	97.25	93.60	60.43	97.70	88.02	60.30	97.88	89.59	72.12	97.17	91.92
Ours	**87.13**	**99.36**	**97.30**	**75.70**	**99.39**	**94.30**	**72.94**	**99.55**	**94.81**	80.22	**99.40**	**95.85**

TABLE 6. On the ISIC2017 dataset. Comparison of Dice coefficient, Jaccard index, Matthew correlation coefficient performance between different attention modules and CSARM blocks using baseline as the infrastructure.

Method	Nevus cases			Melanoma cases			SK cases			Overall		
	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC
SE Block+Baseline	85.09	74.05	83.02	79.18	65.43	74.92	71.76	55.96	68.37	82.35	69.99	79.54
FPA+Baseline	88.00	78.57	86.08	81.27	68.45	77.97	74.27	59.07	69.85	82.61	70.38	80.14
Baseline	76.13	61.74	72.55	68.62	52.23	62.25	67.74	51.22	62.32	72.63	57.02	67.72
Ours	**89.72**	**81.35**	**88.19**	**82.29**	**69.91**	**79.09**	**77.18**	**62.84**	**74.89**	**84.62**	**73.35**	**82.32**

obtain global context information without introducing additional learnable layers, the lower layers of the network use the feature information generated by higher layers, and use the context information of the network itself to generate attention to enhance the feature representation. Use residual learning to further obtain external context information, and effectively model the global context through additive fusion. To verify the improvement of CSARM on the performance of CBAM, we designed a set of structural ablation experiments based on the baseline on the ISIC 2017 dataset, as shown in Table7 and 8. The experimental results show that among the three types of skin diseases, the CSARM model has lower sensitivity (SEN) than the CBAM module except for Nevus and overall, and all other evaluation Metrics are higher than CBAM. In order to compare the performance of the two models more intuitively, we visualize the structural ablation results, as shown in Fig.7, The first two columns are Melanoma,the middle two columns are Seborrheic Keratosis, and the last two columns are Nevus. Meanwhile, in Fig.7(e), we enlarged the lesion area on the image and placed the segmentation results of the two models and the ground truth on the same lesion image for comparison.It can be clearly seen from the segmentation results that the CSARM model can segment the lesion edge more clearly than the model using CBAM block, and better integrate characteristics of the global dependency, making the segmentation results closer to the ground truth.

To further powerfully prove the improved segmentation performance of the model, we perform statistical hypothesis testing on the segmentation performance of the proposed CSARM and CBAM. P-value is an important basis for testing decision. we propose a hypothesis: The performance of the CSARM model is better than the CBAM model. We conducted a P-value analysis of the three lesions -Melanoma, Seborrheic Keratosis, Nevus and overall on Dice and ACC. The results are shown in Table 9. In the table, except for the overall ACC, the P value of the CSARM model is less than 0.05, indicating that the CSARM model has statistical

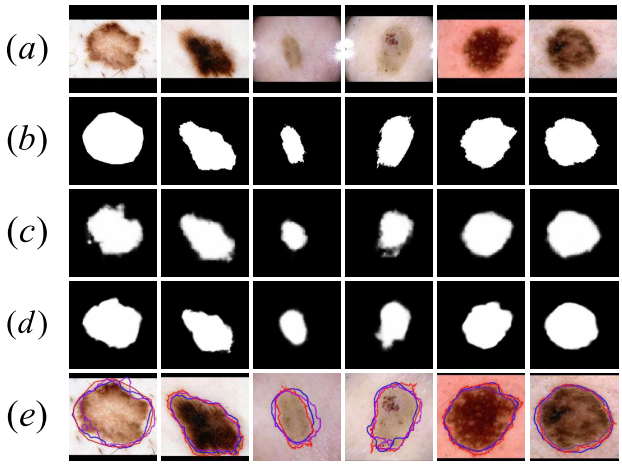

FIGURE 7. Visualization results of structural ablation. (a) Original image of skin lesions (b) The ground truth corresponding to the lesion (c) Lesion segmentation results of CBAM+Baseline (d) Lesion segmentation results of CSARM-CNN model (d) Comparison of segmentation results of the ground truth (red) and CBAM+Baseline(purple), ours(blue). All images are pre-processed.

significance. That is, CSARM model outperforms CBAM model in dermoscopy image segmentation, and has better performance than the CBAM model.

C. COMPARATIVE EXPERIMENT

1) COMPARISON WITH PRIOR ART BY LESION TYPE

This section further evaluates the proposed model for three types of skin lesions in different categories. Tables 10 and 11 summarize the comparison of our proposed CSARM-CNN method to the six best methods tested on the ISIC-2017 dataset. The results of U-Net, FCN, DeepLabv3+, SegNet, Mask-RCNN and FrCN in Table 10 and Table 11 are all from Al-Masni *et al.* [48], and Goyal *et al.* [49]. Compared with other algorithms, in this challenge, the CSARM-CNN model obtained the highest scores of 99.40% and 95.85% in specificity and accuracy respectively, which indicates that it can segment more skin lesion pixels correctly. Even though

TABLE 7. On the ISIC2017 dataset. Use baseline as the infrastructure. Comparison of Sensitivity, Specificity, and Accuracy performance between CBAM block and CSARM block.

Method	Nevus cases			Melanoma cases			SK cases			Overall		
	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC
Baseline	78.53	97.25	93.60	60.43	97.70	88.02	60.30	97.88	89.59	72.12	97.17	91.92
CBAM+Baseline	**91.08**	97.36	96.83	75.19	97.72	92.97	70.48	98.11	93.36	**82.20**	97.99	95.25
Ours	87.13	**99.36**	**97.30**	**75.70**	**99.39**	**94.30**	**72.94**	**99.55**	**94.81**	80.22	**99.40**	**95.85**

TABLE 8. On the ISIC2017 dataset. Use baseline as the infrastructure. Comparison of Jaccard index, Dice coefficient, Matthew correlation coefficient performance between CBAM block and CSARM block.

Method	Nevus cases			Melanoma cases			SK cases			Overall		
	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC
Baseline	78.53	97.25	93.60	60.43	97.70	88.02	60.30	97.88	89.59	72.12	97.17	91.92
CBAM+Baseline	88.94	80.09	87.10	81.68	69.89	78.13	74.97	60.26	70.44	84.01	72.98	81.21
Ours	**89.72**	**81.35**	**88.19**	**82.29**	**69.91**	**79.09**	**77.18**	**62.84**	**74.89**	**84.62**	**73.35**	**82.32**

TABLE 9. On the ISIC2017 dataset. Comparison of Pvalue analysis results based on Dice and Acc.

Types	Nevus cases		Melanoma cases		SK cases		Overall	
	DIC	ACC	DIC	ACC	DIC	ACC	DIC	ACC
Mean±Std(CBAM)	88.94±0.74	96.83±0.92	81.68±0.22	92.97±0.73	74.97±0.43	93.36±1.29	84.01±1.19	95.35±0.45
Mean±Std(ours)	89.72±0.66	97.30±1.02	82.29±0.18	94.30±0.94	77.18±0.34	94.81±1.07	84.62±1.10	95.85±0.47
P-value	0.01	0.03	0.01	0.003	0.01	0.03	0.02	0.14

TABLE 10. On the ISIC2017 dataset. Comparison of Sensitivity, Specificity, and Accuracy performance between CSARM-CNN model and the six best methods.

Method	Nevus cases			Melanoma cases			SK cases			Overall		
	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC
U-Net[47]	76.76	97.26	92.89	58.71	96.81	84.98	43.81	97.64	84.83	67.15	97.24	90.14
FCN-AlexNet [48]	82.44	97.58	94.84	72.35	96.23	87.82	71.70	97.92	89.35	78.86	97.37	92.65
FCN-32s [48]	83.67	96.69	94.59	74.36	96.32	88.94	75.80	96.41	89.45	80.67	96.72	92.72
FCN-16s [48]	84.23	96.91	94.67	75.14	96.27	89.94	75.48	96.25	88.83	81.14	96.68	92.74
FCN-8s [48]	83.91	97.22	94.55	78.37	95.96	89.63	69.85	96.57	87.40	80.72	96.87	92.52
DeepLabV3+ [47]	88.54	97.21	95.67	77.71	96.37	89.65	74.59	98.55	90.06	84.34	97.25	93.66
Mask-RCNN [47]	87.25	96.38	95.32	78.63	95.63	89.31	**82.41**	94.88	90.85	84.84	96.01	93.48
SegNet [47]	85.19	96.30	93.93	73.78	94.26	87.90	70.58	92.50	87.29	80.05	95.37	91.76
FrCN [47]	**88.95**	97.44	95.62	**78.91**	96.04	90.78	82.37	94.08	91.29	**85.40**	96.69	94.03
Baseline	78.53	97.25	93.60	60.43	97.70	88.02	60.30	97.88	89.59	72.12	97.17	91.92
Ours	87.13	**99.36**	**97.30**	75.70	**99.39**	**94.30**	72.94	**99.55**	**94.81**	80.22	**99.40**	**95.85**

TABLE 11. On the ISIC2017 dataset. Comparison of Dice coefficient, Jaccard index, Matthew correlation coefficient performance between CSARM-CNN model and the six best methods.

Method	Nevus cases			Melanoma cases			SK cases			Overall		
	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC
U-Net[47]	82.16	69.71	78.05	70.82	54.83	63.71	57.88	40.73	63.89	76.27	61.64	71.23
FCN-AlexNet [48]	85.61	77.01	82.91	75.94	64.32	70.35	75.09	63.76	71.51	82.15	72.55	78.75
FCN-32s [48]	85.08	76.39	82.29	78.39	67.23	72.70	76.18	64.78	72.10	82.44	72.86	78.89
FCN-16s [48]	85.60	77.39	82.92	79.22	68.41	73.26	75.23	64.11	71.42	82.80	73.65	79.31
FCN-8s [48]	84.33	76.07	81.73	80.08	69.58	74.39	68.01	56.54	65.14	81.06	71.87	77.81
DeepLabV3+ [47]	88.29	81.09	85.90	80.86	71.30	76.01	77.05	67.55	74.62	85.16	77.15	82.28
Mask-RCNN [47]	88.83	80.91	85.38	80.28	70.69	74.95	80.48	**70.74**	**76.31**	85.58	**77.39**	81.99
SegNet [47]	85.69	74.97	81.84	79.11	65.45	71.03	72.54	56.91	64.32	82.09	69.63	76.79
FrCN [47]	89.68	81.28	86.90	**84.02**	**72.44**	77.90	**81.83**	69.25	76.11	**87.08**	77.11	**83.22**
Baseline	76.13	61.74	72.55	68.62	52.23	62.25	67.74	51.22	62.32	72.63	57.02	67.72
Ours	**89.72**	**81.35**	**88.19**	82.29	69.91	**79.09**	77.18	62.84	74.89	84.62	73.35	82.32

the training images are uneven for different lesion categories, the high-quality segmentation results obtained for all categories still prove the success of our proposed lesion segmentation network.

In Fig. 8, six dermoscopy images and corresponding visualization results for segmentation of three types of skin lesions trained by CSARM-CNN are given, and their CAM saliency maps are shown. The CAM saliency map shows that the area of interest learned by the model, that is, the highlights in the CAM, has different positions and concentrations, which shows the segmentation effect more intuitively. At the same time, for qualitative evaluation, Figure 8(d) shows the

TABLE 12. Performance evaluation of Sensitivity, Specificity, and Accuracy of different segmentation algorithms of the proposed CSARM-CNN model on the PH2 dataset.

Method	Benign cases			Melanoma cases			Overall		
	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC
FCN [17]	**95.35**	94.09	94.44	90.30	94.02	92.82	90.30	94.02	92.82
SegNet [45]	91.57	96.57	95.19	75.50	96.83	86.04	86.53	96.61	93.36
U-Net [16]	86.68	97.63	94.60	70.58	**98.47**	84.36	81.63	97.76	92.55
FrCn [47]	94.48	95.46	95.20	**91.57**	96.55	**94.64**	**93.72**	95.65	95.08
Ours	94.84	**99.67**	**96.72**	84.20	96.12	91.33	88.54	**99.45**	**95.23**

TABLE 13. Performance evaluation of Dice coefficient, Jaccard index, and Matthew correlation coefficient of different segmentation algorithms of the proposed CSARM-CNN model on the PH2 dataset.

Method	Benign cases			Melanoma cases			Overall		
	DIC	JAC	MCC	DIC	JAC	MCC	DIC	JAC	MCC
FCN [17]	90.46	82.59	86.78	89.03	80.22	83.71	89.03	80.22	83.71
SegNet [45]	91.32	84.03	87.99	84.55	73.23	73.89	89.36	80.77	84.64
U-Net[16]	89.88	81.63	86.32	82.04	69.55	71.73	87.61	77.95	82.78
FrCN [47]	**91.38**	**84.13**	88.15	**92.92**	**86.77**	**88.62**	**91.77**	**84.79**	**88.30**
Ours	90.55	82.74	**88.60**	88.65	79.62	81.97	88.32	79.09	85.53

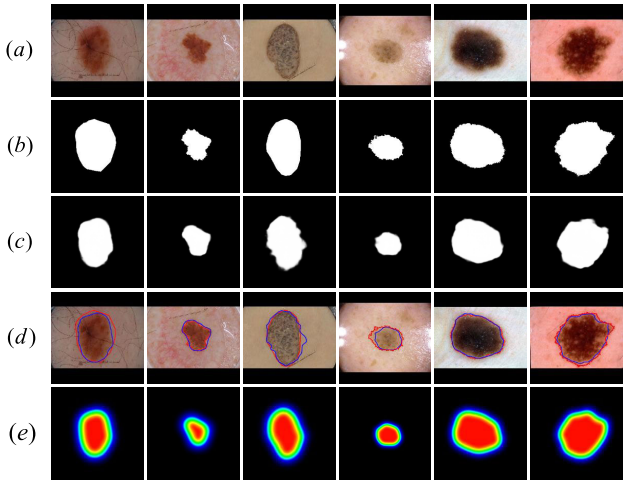

FIGURE 8. Visualization results for ISIC2017 dataset. (a) Original image of skin lesions (b) The ground truth corresponding to the lesion (c) Lesion segmentation results of CSARM-CNN model (d) Comparison of the ground truth (red) and segmentation results of ours (blue) (e) CAM saliency map visualization of lesion segmentation of CSARM-CNN model.

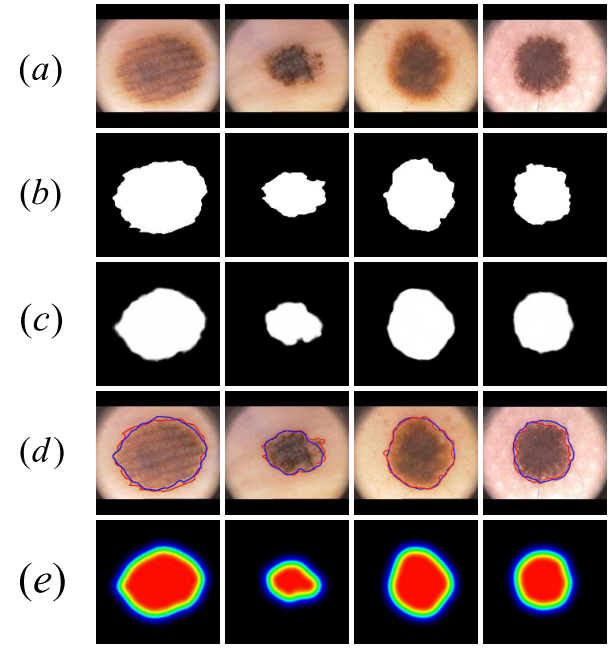

FIGURE 9. Visualization results for PH2 dataset. (a) Original image of skin lesions (b) The ground truth corresponding to the lesion (c) Lesion segmentation results of CSARM-CNN model (d) Comparison of the ground truth (red) and segmentation results of ours (blue) (e) CAM saliency maps visualization of lesion segmentation of CSARM-CNN model.

comparison between some typical segmentation results of CSARM-CNN and the ground truth contour of benign Nevus, Melanoma and SK lesions in ISIC-2017 test dataset. The first two columns are Melanoma, the middle two columns are Seborrheic Keratosis, and the last two columns are Nevus.

2) COMPARISON ON THE PH2 DATASET

To test the robustness and cross-dataset performance of our method, we also evaluated our proposed model using the PH2 test data set (including 200 dermoscopy images). As shown in Tables 12 and 13, our method performs slightly better in melanoma cases, while it shows a significant improvement in benign cases.

The proposed CSARM-CNN achieves significantly better specificity and accuracy than previous work, which means that our method has higher non-lesional segmentation ability and overall pixel-level segmentation performance compared to other methods. In addition, the experimental results also show that our method is equivalent to the existing technology in Dice coefficient and Jac Index, but in the previous work, before segmenting the lesion, in many cases, the hair removal problem was solved by additional pretreatment steps.This step involves applying different filters (such as a directional Gaussian filter [45]) to the original image. In contrast, our strategy requires no additional steps. Figure 9 visually shows the segmentation result map (d) and CAM saliency map (e) of our proposed CSARM-CNN compared with ground truth.

The first two columns are Melanoma, the middle two columns are Seborrheic Keratosis, and the last two columns are Nevus.

V. CONCLUSION

In this paper, an end-to-end skin lesion segmentation model CSARM-CNN is proposed. The model is based on the CSARM block, which uses a combination of residual learning, channel attention mechanism and spatial attention mechanism to improve the discriminant and representational ability of CNN. The model uses U-net as the basic structure. An image pyramid is built in the encoder path to feed multi-scale inputs. A local prediction map corresponding to the multi-scale input image is generated in the decoder path. At the same time, the multi-output cross entropy loss function is used to promote the training of the model.To verify its effectiveness, we evaluated the model using two publicly available datasets (ISIC-2017 Challenge and PH2 dataset). The results show that the proposed CSARM-CNN is superior to some of the latest algorithms for skin lesion segmentation, and it is verified that the CSARM block can be applied to different network models and improve the segmentation performance of the model on skin lesions. Compared with two common attention modules (SE Block and FPA), CSARM also has certain competitiveness in skin lesions. For future work, we believe that combining appropriate pre-processing and post-processing stages with the proposed model will further improve model performance and apply the model to other medical applications to demonstrate its versatility.

REFERENCES

[1] R. J. Hay, N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf, C. Michaud, C. J. L. Murray, and M. Naghavi, "The global burden of skin disease in 2010: An analysis of the prevalence and impact of skin conditions," *J. Investigative Dermatology*, vol. 134, no. 6, pp. 1527–1534, Jun. 2014.

[2] R. HW, W. MA, F. SR, and C. BM, "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012," *JAMA Dermatol.*, vol. 151, no. 10, pp. 1081–1086, 2015.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[4] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinge, "Skin lesion classification using hybrid deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1229–1233.

[5] L. Yu *et al.*, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, 2016.

[6] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," 2019, *arXiv:1903.03313*. [Online]. Available: http://arxiv.org/abs/1903.03313

[7] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[8] E. Ahn, J. Kim, L. Bi, A. Kumar, C. Li, M. Fulham, and D. D. Feng, "Saliency-based lesion segmentation via background detection in dermoscopic images," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 6, pp. 1685–1693, Nov. 2017.

[9] H. Li, X. He, F. Zhou, Z. Yu, D. Ni, S. Chen, T. Wang, and B. Lei, "Dense deconvolutional network for skin lesion segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 527–537, Mar. 2019.

[10] B. Erkol, R. H. Moss, R. Joe Stanley, W. V. Stoecker, and E. Hvatum, "Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes," *Skin Res. Technol.*, vol. 11, no. 1, pp. 17–26, Feb. 2005.

[11] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artif. Intell. Med.*, vol. 56, no. 2, pp. 69–90, Oct. 2012.

[12] M. Silveira, J. C. Nascimento, J. S. Marques, A. R. S. Marcal, T. Mendonca, S. Yamauchi, J. Maeda, and J. Rozeira, "Comparison of segmentation methods for melanoma diagnosis in dermoscopy images," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 1, pp. 35–45, Feb. 2009.

[13] Z. Ma and J. M. R. S. Tavares, "A novel approach to segment skin lesions in dermoscopic images based on a deformable model," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 2, pp. 615–623, Mar. 2016.

[14] C. Kang, X. Yu, S.-H. Wang, D. Guttery, H. Pandey, Y. Tian, and Y. Zhang, "A heuristic neural network structure relying on fuzzy logic for images scoring," *IEEE Trans. Fuzzy Syst.*, early access, Jan. 13, 2020, doi: 10.1109/TFUZZ.2020.2966163.

[15] S. Wang, J. Sun, I. Mehmood, and C. Pan, "Cerebral micro-bleeding identification based on a nine-layer convolutional neural network with stochastic pooling," *Concurrency Comput., Pract. Exper.*, vol. 32, no. 1, p. e5130, 2020.

[16] 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.

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

[18] M. Prathiba, D. Jose, R. Saranya, and Nandhinidevi, "Automated melanoma recognition in dermoscopy images via very deep residual networks," in *Proc. IOP Conf., Mater. Sci. Eng.*, vol. 561, Nov. 2019, Art. no. 012107.

[19] 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.

[20] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017.

[21] K. Xu, J. Ba, R. Kiros, K. Cho, and A. Courville, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[23] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, "Lesion border detection in dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 33, no. 2, pp. 148–153, Mar. 2009.

[24] Y. Yuan and Y.-C. Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 519–526, Mar. 2019.

[25] M. M. K. Sarker, H. A. Rashwan, F. Akram, and S. F. Banu, "SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 21–29.

[26] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: A review," *J. Med. Syst.*, vol. 42, no. 11, p. 226, Nov. 2018.

[27] M. Kamrul Hasan, L. Dahal, P. N. Samarakoon, F. Islam Tushar, and R. Marti Marly, "DSNet: Automatic dermoscopic skin lesion segmentation," 2019, *arXiv:1907.04305*. [Online]. Available: http://arxiv.org/abs/1907.04305

[28] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[29] 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.

[30] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.

[31] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[32] R. A. Rensink, "The dynamic representation of scenes. Visual Cognition7: 1742.[aZWP](2000b) Visual search for change: A probe into the nature of attentional processing," *Vis. Cognition*, vol. 7, p. 34576, Apr. 2000.

[33] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, Mar. 2002.

[34] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[35] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," 2014, *arXiv:1412.7755*. [Online]. Available: http://arxiv.org/abs/1412.7755

[36] 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.

[37] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 683–687.

[38] C. Kaul, S. Manandhar, and N. Pears, "FocusNet: An attention-based fully convolutional network for medical image segmentation," 2019, *arXiv:1902.03091*. [Online]. Available: https://arxiv.org/abs/1902.03091

[39] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2018, pp. 3–19.

[42] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.

[43] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[44] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH^2-A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5437–5440.

[45] 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.

[46] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," 2015, *arXiv:1505.07293*. [Online]. Available: http://arxiv.org/abs/1505.07293

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

[48] M. A. Al-masni, M. A. Al-antari, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Comput. Methods Programs Biomed.*, vol. 162, pp. 221–231, Aug. 2018.

[49] M. Goyal, A. Oakley, P. Bansal, D. Dancey, and M. H. Yap, "Skin lesion segmentation in dermoscopic images with ensemble deep learning methods," *IEEE Access*, vol. 8, pp. 4171–4181, 2020.

YUN JIANG was born in Zhejiang, China, in 1970. She received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2007. She is currently a Professor with the College of Computer Science and Engineering, Northwest Normal University. Her research interests include data mining, rough set theory and application, and medical image processing.

SIMIN CAO was born in Shanxi, China, in 1995. She is currently pursuing the M.S. degree with the College of Computer Science and Engineering, Northwest Normal University. Her research interests include deep learning and medical image processing.

SHENGXIN TAO was born in Gansu, China, in 1996. He is currently pursuing the M.S. degree with the College of Computer Science and Engineering, Northwest Normal University. His research interests include deep learning and medical image processing.

HAI ZHANG was born in Jiangxi, China, in 1995. He is currently pursuing the M.S. degree with the College of Computer Science and Engineering, Northwest Normal University. His research interests include deep learning and medical image processing.