IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 29, 2021, accepted November 30, 2021, date of publication December 3, 2021,
date of current version December 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3132293

SCA-Net: A Spatial and Channel Attention Network for Medical Image Segmentation

TONG SHAN[1] AND JIAYONG YAN[2,3]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2]School of Medical Instruments, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China
[3]Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China

Corresponding author: Jiayong Yan (yanjy@sumhs.edu.cn)

This work was supported in part by the Jiangsu Provincial Key Research and Development Fund Project under Grant BE2017601.

ABSTRACT Automatic medical image segmentation is a critical tool for medical image analysis and disease treatment. In recent years, convolutional neural networks (CNNs) have played an important role in this field, and U-Net is one of the most famous fully convolutional network architectures among many kinds of CNNs for medical segmentation tasks. However, the CNNs based on U-Net used for medical image segmentation rely only on simple concatenation operation of multiscale features. The spatial and channel context information is easily missed. To capture the spatial and channel context information and improve the segmentation performance, in this paper, a spatial and channel attention network (SCA-Net) is proposed. SCA-Net presents two novel blocks: a spatial attention block and a channel attention block. The spatial attention block (SAB) combines the multiscale information from high-level and low-level stages to learn more representative spatial features, and the channel attention block (CAB) redistributes the channel feature responses to strengthen the most critical channel information while restraining the irrelevant channels. Compared with other state-of-the-art networks, our proposed framework obtained better segmentation performance in each of the three public datasets. The average Dice score improved from 88.79% to 92.92% for skin lesion segmentation, 94.02% to 98.25% for thyroid gland segmentation and 87.98% to 91.37% for pancreas segmentation compared with U-Net. Additionally, the Bland–Altman analysis showed that our network had better agreement between automatic and manually calculated areas in each task.

INDEX TERMS Deep learning, multiscale contextual information, attention, medical image segmentation.

I. INTRODUCTION

Medical image segmentation is an essential tool for current clinical applications, such as computer-aided diagnosis/detection (CAD) or therapy plan systems (TPSs) [1], [2]. Automation of medical segmentation can increase the speed and efficiency and greatly reduce tedious and time-consuming work for doctors. In brief, the main target of medical image segmentation is to distinguish the target region of interest from the background effectively. However, it is a challenging task due to several factors. First, medical images are collected by different acquisition facilities and usually have low imaging quality, leading to incomplete segmentation or excessive segmentation. Second, some segmentation targets usually have a wide variety of shapes and scales from patient to patient, making it difficult to construct excellent performance. Additionally, some targets of interest to be segmented have a wide range of orientations and positions in the context of medical images, such as the pancreas in magnetic resonance imaging (MRI) [3]–[5].

In recent years, deep learning has become the mainstream research method in many fields, and deep convolutional neural networks (CNNs) have attracted much attention from researchers in the field of medical image segmentation because of their good performance. Compared with traditional medical image segmentation methods, the ability to extract the features automatically helps CNNs learn from the obtained dataset. Many state-of-the-art works have achieved noticeable performance in medical image segmentation tasks. However, there are still some problems with CNNs. First, the weight-sharing design of CNNs between the same input feature layer and output feature layer easily weakens the learning ability of CNNs for complex textures and shapes. At the same time, the increased number of channels causes

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

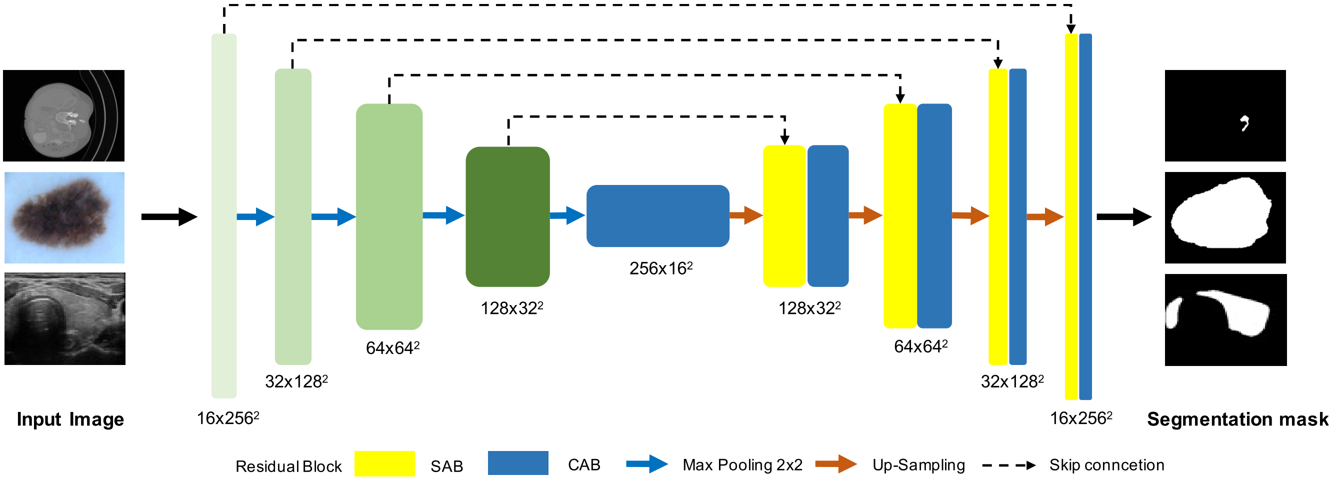


FIGURE 1. An overview of our proposed SCA-Net.

redundant computation and memory consumption. Second, with the depth growth of CNNs, the network becomes hard to train, and the risk of gradient disappearance is increased. Third, continuous pooling operations cause important local and global context information to be lost. To efficiently enhance the network segmentation performance for the network constructed with convolution operation, some ideas have been fused in CNNs and showed signs of progress in the medical image segmentation field. For instance, U-Net [6], one of the most popular architectures in the medical image segmentation field, employed a symmetrical U-shaped structure with skip connections to concatenate multiscale feature maps from low-level and high-level layers. [7] employed a dilated convolution operation with multiple different dilation rates to extract contextual feature maps. [8] applied a fully connected CRF to maximize labeling similar pixel points and modeling the spatial contextual feature relationships in object classes. Although these methods enhanced the CNNs' feature extraction ability, the descriptive information for spatial features and channel features, which are very useful for medical image segmentation, is still limited.

To learn more local related features and overlook irrelevant details from the feature maps, several variants of attention mechanisms have been proposed and have achieved better performances in computer vision tasks [9]–[13]. Attention U-Net [13] employed the attention gate (AG), which fuses multistage contextual information from the encoder and the decoder. AGs learn to suppress the irrelevant characteristic response in the background while focusing on target regions. SE-Net [14] employed the SE-block, which is a kind of channel attention mechanism. It recalibrates the channel feature maps, assigns more weights to important feature channels and restrains irrelevant channels. The semantic segmentation methods proposed in [15] and [16] utilized similar ideas to enhance the network segmentation performance. [17] and [18] introduced an attention mechanism into the deep adversarial learning framework for capturing more contextual information. The results obtained by these works demonstrate the effectiveness of attention modules for segmentation tasks.

Inspired by previous works of CNNs on medical image segmentation, this paper introduces multiscale spatial and channel information to achieve better segmentation performance for medical images. Based on the encoder-decoder architecture and the attention mechanism, we proposed a spatial and channel attention network (SCA-Net) for medical image segmentation tasks, which is shown in Fig. 1. In SCA-Net, two novel attention blocks are constructed for capturing the spatialwise and channelwise relationships. One is the spatial attention block (SAB), and the other is the channel attention block (CAB). The two blocks are integrated into the decoder. The SAB learns to focus on the target spatial regions and ignores the irrelevant background by resigning each pixel weight. The CAB emphasizes the relativity of different channels, which redistributes the critical channel information and overgoes unrelated channel information.

In summary, the main contributions of our work are organized as follows:

1) We propose two attention blocks: a spatial attention block (SAB) and a channel attention block (CAB). The SAB is supported to recalibrate the features of spatial context information, and the CAB is supported to highlight the relevant channels and restrain the irrelevant channels.

2) The proposed blocks SAB and CAB are integrated in a novel network named SCA-Net. The ablation study shows that the proposed blocks can effectively capture the features for the targets of interest to be segmented.

3) Our proposed method was verified on three different medical image segmentation tasks. The experimental results show that SCA-Net has superior performance.

II. RELATED WORK

A. CNNs FOR IMAGE SEGMENTATION

Convolution is the core operation of CNNs. Without manually selecting features or prior knowledge, CNNs express the ability to learn features from acquired datasets automatically.

In recent research, CNNs have been widely applied in different tasks [19]–[21]. By deepening the CNN layers and using ReLU+dropout, AlexNet achieved the best classification results at that time [22].

By replacing the last fully connected layers of classification CNNs with convolution layers, fully convolutional network (FCN) architectures have made significant progress for natural semantic segmentation, such as DeepLab for semantic image segmentation [23]. Subsequently, SegNet [24] proposed the encoder and decoder architecture, which employed CNN as the base unit and achieved state-of-the-art performance for semantic image segmentation. However, the CNN performance is still limited by position-invariant convolutional kernels, without attending to spatial and channel information, which are very important for segmenting objects.

B. MULTI-SCALE INFORMATION FUSION

In computer vision tasks, rich contextual features extracted from multiscale information help the network achieve better segmentation performance. Many methods using multiscale information have been proposed and applied to 2D and 3D medical image segmentation. Similar to [24], the structure of U-Net [6] adopts a symmetrical encoder and decoder architecture with a skip connection to perform 2D medical image segmentation. To date, many models have been proposed based on U-Net, including U-Net++ [25], DoubleU-Net [26], and DUNet [27]. They have been successfully applied to different 2D medical image segmentation tasks. At the same time, 3D U-Net [28] and V-Net [29] were proposed for 3D medical image segmentation tasks.

To compensate for lost feature details during the downsampling operation, dilation convolution [30] with different rates enlarges the receptive field to capture more contextual information [31]–[34]. For instance, CE-Net designed a context extractor module to learn contextual semantic information [31]. It generated more presentive feature maps. [33] learned local geometric details using the cascaded pyramid architecture, which was fused in dilation convolution with different dilation rates. However, the scan area of dilation convolution is not continuous. For small targets, the gain is not worth the loss. Less attention has been paid to the interrelationship between spatial and channel characteristics.

C. ATTENTION MECHANISM

The attention mechanism has proven to be an efficient method to enhance CNN performance [35]. It mimics the biological observation process of paying attention to more detailed information about the desired target and suppressing useless information. [9] was the first to propose an attention mechanism for processing natural language translation. [36] relied on self-attention to capture the dependencies of inputs for machine translation. Meanwhile, the attention mechanism has been used in the field of computer vision [37]–[39]. [37] and [38] used spatial attention for image classification and image captioning. [39] employed a dual attention mechanism to capture global features for semantic segmentation.

In many digital image segmentation tasks, the attention mechanism has also been adopted for better performance. Generally, attention modules can be plug-and-play in CNNs and help CNNs focus on more effective features of the target using spatial regions and channel interrelationships. Based on U-Net [6], AG Gate [13] focuses on the salient feature shape and size of the target through multiscale information. SE-Net [14] employed the squeeze and excitation (SE) block to recalibrate relevant channel feature maps and overgo irrelevant features. CBAM [40] emphasized the meaningful features in space and channels. It enhanced the feature representation of key regions related to the target. [41] designed an autofocus attention layer for semantic segmentation. It employed multiparallel attention branches, which had different scales of receptive fields to focus on the optimal scales. However, multiple branches increase the complexity of models and the difficulty of training. Inspired by previous methods, we hypothesize that the effective use of spatial information and channel-dependent features can improve the segmentation performance of our network.

III. MATH

Based on previous works, we use the effective architecture of the encoder and decoder as our backbone. As illustrated in Fig. 1, the architecture proposed by this paper has three major components: residual block, spatial attention block (SAB) and channel attention block (CAB). The encoder transforms the input image into multidimensional feature maps and extracts the segmentation information, and the decoder generates spatial feature maps across aggregating multiscale information and distributes the weight of feature map channels.

In the encoding stage, we use the residual block to retain more original information and extract the feature maps. In the decoding stage, the SAB redistributes the spatial pixel weights by aggregating pooled feature information from high-level and low-level stages. The CAB exploits the channel features, which uses global average pooling and global max pooling to excite more channel contextual information. It reassigns the relationship of every channel and its neighbors to highlight more important channel information. The details of these modules are described as below.

A. RESIDUAL BLOCK

With increasing network depth, the model generally has a better expression for tasks. However, it increases the risk of gradient degradation and explosion of the network at the same time. To solve these problems, [42] proposed the residual learning network, which employed the residual connection to ease the difficulty of network training and keep more learnable features.

Inspired by the residual learning framework, we use two 1 × 1 convolution blocks and one 3 × 3 convolution

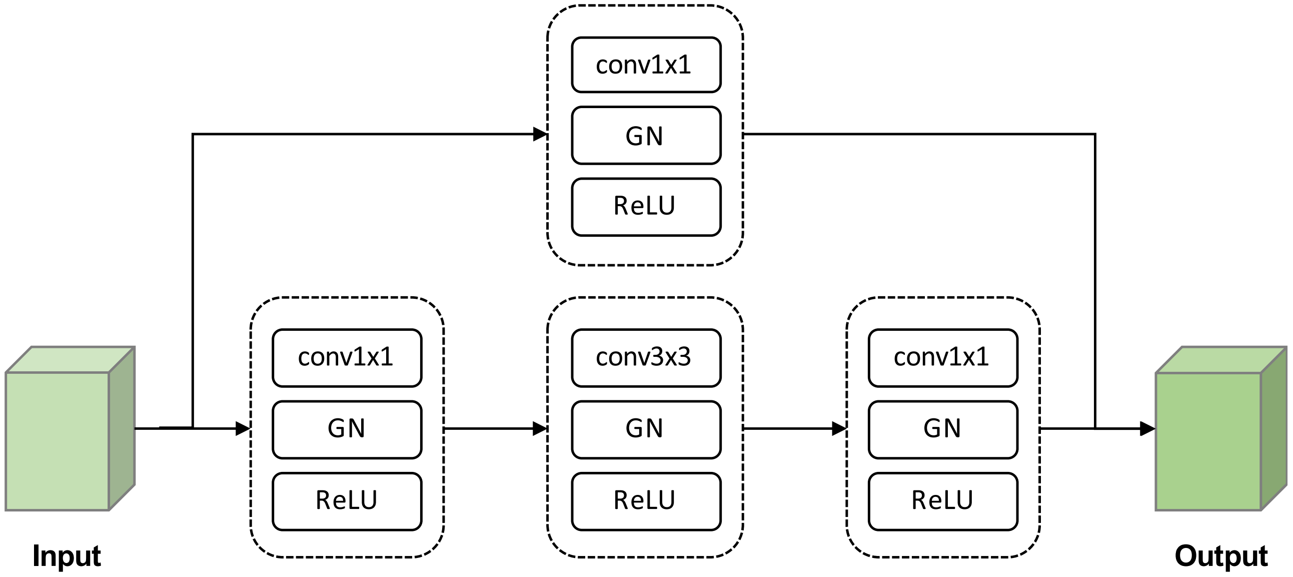


FIGURE 2. The residual block in our proposed method. We use GN layer to replace the original BN layer.

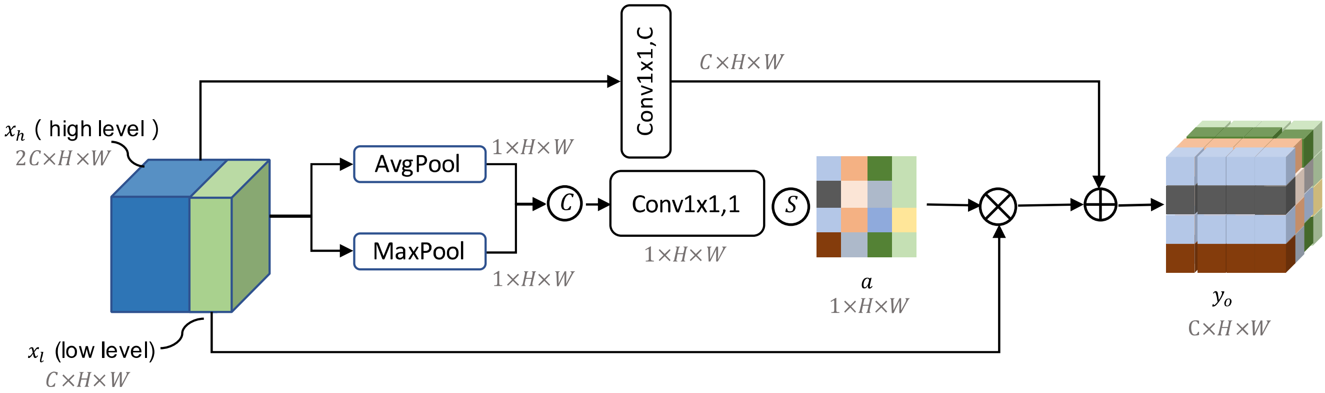


FIGURE 3. Structure of spatial attention block in our proposed method. The input image is $3C \times H \times W$ and the output image is $C \times H \times W$. The α presents the feature map.

block to generate the multidimensional feature maps, and another residual connection is employed to reserve the original feature information, which uses 1 × 1 convolution to adapt the number of channels. A small batch size may cause training gradient degradation and decrease the network performance. Thus, we use group normalization [43] instead of BN in our entire network. The residual block avoids the risk of vanishing gradients and accelerates network convergence. The residual block used in this paper is shown in Fig. 2.

B. SPATIAL ATTENTION BLOCK

Previous works [31], [41] show that a deep convolution network with atrous convolutional blocks and multikernel branches can effectively extract contextual features from images. However, using these blocks consumes considerable memory and increases the complexity of the model. To use the multiscale contextual information and the experimental platform's memory effectively, Attention-UNet [13] utilized AG Gate to capture the spatial features from multiscale information. Motivated by these methods, we design SAB to fuse adjacent features of high-level and low-level spatial feature maps from multiple stages. By extracting the relationship of spatial interpixels, SAB can focus on meaningful spatial features and highlight prospective information.

The SAB is shown in Fig. 3. x_l represents the low-level feature input from the encoder with the shape of $C \times H \times W$, where C denotes input channels and H, W indicate the height and width of input, respectively. x_h represents the input of high-level features with the shape of $2C \times H \times W$, which are upsampled from the previous decoder layer. Compared with x_l, x_h has a higher spatial resolution. First, we concatenate them into the shape of $3C \times H \times W$, then feed them into global average-pooled and global max-pooled functions along the spatial dimension with the shape of $1 \times H \times W$ and concatenate them by the channel dimension. One 1×1 convolution kernel with an output channel of 1 is employed to fuse the spatial feature. The *Sigmoid* activation function

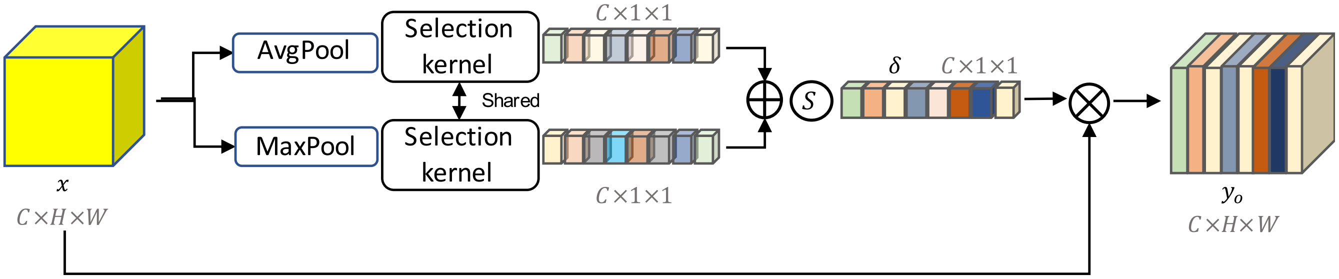


FIGURE 4. Structure of channel attention block in our proposed method. The δ means the channel attention coefficient.

is applied to gain a spatial wise statistic $\alpha \in [0, 1]_{H \times W}$. The size of feature map α is $1 \times H \times W$. To calibrate the spatial feature maps, x_l is subsequently multiplied by α. To reuse the feature of x_h, we employ the residual connection. x_h is compressed by 1×1 convolution with C output channels as x^h. Furthermore, the output is obtained as:

$$y_o = \Phi^C (x_h) + a * x_l \tag{1}$$

where Φ^C denotes the 1×1 convolution with C output channels and $*$ denotes the elementwise dot product. The number of output channels depends on the stage of x_l. Here, C is 128, 64, 32 and 16 for different dimensional stages.

C. CHANNEL ATTENTION BLOCK

The spatial feature maps from SAB contain considerable spatial interpixel information, as shown in Fig. 3. However, the output from SAB still contains unutilized channel feature information. To exploit critical features and suppress useless ones, we use CAB to redistribute the channel feature responses and strengthen important channel features provided by SAB. The details of CAB are shown in Fig. 4.

SE-Net [14] shows the effectiveness of the squeeze-and-excitation block, which specifies the interchannel relationship. However, it only uses global average-pooled information. Compared with SE-Block, we additionally use the global max-pooled information, which stores more channel contextual information. Taking x as an input with the shape of $C \times H \times W$, global average pooling and global maximal pooling are separately applied along the x channel dimension to obtain global channel information with the shape of $C \times 1 \times 1$. Inspired by ECA-Net [35], CAB employs one-dimensional kernel convolution with a kernel size of 5×5 to preliminarily capture the nonlinear cross-channel interaction. To decrease the parameters and complexity, the weights of convolution kernels are shared. We use the F_{add} function to fuse the obtained channel information, and the result is fed into the *sigmoid* active function to obtain the output δ with the shape of $C \times 1 \times 1$. Finally, the output y_o of our channel attention module is:

$$y_o = x * \delta \tag{2}$$

where $*$ denotes channelwise multiplication. The shape of y_o is $C \times H \times W$.

D. LOSS FUNCTION

Our proposed framework is an end-to-end training network. In our medical image segmentation tasks, we need to train our network to accurately predict the classification of each pixel. In recent years, the cross entropy loss function has been broadly used in the medical image segmentation field. However, some medical image segmentation objects often have a range of variations in scale and direction in the region of interest, particularly the pancreas and skin lesions. Accordingly, we used the soft dice loss function to alleviate the above problem. The soft dice loss function uses the predicted probability maps instead of thresholding and converts them into a binary mask. We used it in training and validation processing. It is described as:

$$L_{dice} = 1 - \frac{2 \sum_{pixels} y_{true} \times y_{pred}}{\sum_{pixels} y_{true}^2 + \sum_{pixels} y_{pred}^2} \tag{3}$$

where y_{true} denotes the ground truth point values and y_{pred} denotes the predicted probability point values.

IV. EXPERIMENTS AND RESULTS

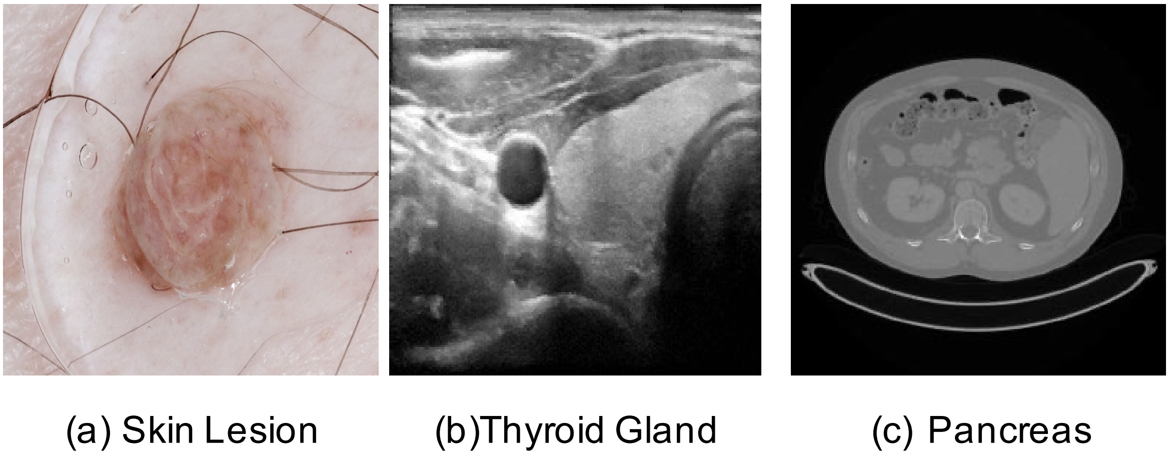
A. DATASET

To assess the effectiveness of the proposed method, we applied our network to three medical image segmentation tasks: skin lesion segmentation from ISIC 2018, thyroid gland segmentation, and pancreas segmentation. Each task has its own challenge, and the sample of three datasets is shown in Fig. 5.

B. IMPLEMENTION DETAILS

During our experiment, the input images were resized with a uniform size of 256×256 and normalized by the mean value and standard deviation. Fivefold cross-validation was employed to assess the performance of the proposed model. The dataset was randomly split at ratios of 70%, 10% and 20% for training, validation and testing, respectively. To reduce the risk of overfitting, we randomly rotated the training dataset at an angle of $(-\pi/9, \pi/9)$, which increased the number of training images.

Our framework was implemented on the PyTorch platform. The training batch size was 16, and the Adaptive Moment Estimation (Adam) optimizer was employed to train the network. The initial learning rate is 10^{-4}, and the weight



(a) Skin Lesion (b)Thyroid Gland (c) Pancreas

FIGURE 5. Three samples for each segmentation task. (a) Skin lesion segmentation from ISIC 2018. (b) Thyroid gland segmentation. (c) Pancreas segmentation.

decay is 10^{-8} for our experiments. For each task, we iterate the network for 300 epochs. The experimental hardware used is one NVIDIA Tesla P100 with 16 GB for all experiments. The soft dice loss function is used to train our network. During the process of validation, we saved the best performing model with the smallest loss. It was used in the test dataset to evaluate model performance.

C. EVALUATION METHODS

To quantitatively evaluate the segmentation performance of networks, we used the following evaluation methods, which are shown below:

$$Dice = \frac{2\,|A \cap B|}{|A| + |B|} \tag{4}$$

$$IOU = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

$$RAVD = \frac{|A| - |B|}{|B|} \tag{6}$$

$$ASSD = \frac{1}{|S_a| + |S_b|} \times \left(\sum\nolimits_{a \in S_a} d\,(a, S_b) + \sum\nolimits_{b \in S_b} d\,(b, S_a)\right) \tag{7}$$

where A denotes the region of the predicted probability segmentation map and B denotes the ground truth binary image. S_a denotes the set of segmentation boundary points, and S_b is defined as the set of ground truth boundary points. $d(v, S_a) = min_{x \in S_a}(|v - x|)$ denotes the shortest Euclidean distance between point x and all points of S_a. Additionally, the Bland–Altman plot, which is a commonly used method for analyzing the consistency of two technologies in medical statistics, is applied to visualize the potential bias between the areas segmented by the automatic method and in a manual manner.

D. ABLATION ANALYSIS

To prove the validity of SAB and CAB in our proposed SCA-Net, we evaluate the proposed module by ablation analysis. Each module performance was tested by segmenting skin lesions from the ISIC 2018 dataset. The residual block replaces all convolutional layers in U-Net [6] as our backbone.

In the next ablation experiment, the residual block is used in the encoder path, and the decoder path integrates the SAB and CAB to extract the feature information from feature maps separately. The skip connection is used for concatenating features between the encoder and the decoder as implemented in the U-Net architecture [6].

The results of the quantitative comparison of these methods are shown in Table 1. U-Net with residual block is assumed to be the backbone. For the skin lesion segmentation task, the performance of the backbone with SAB and CAB is improved separately. Additionally, our proposed SCA-Net can significantly enhance the performance of medical image segmentation. Compared with the backbone, our proposed network improved the average Dice from 0.8944 to 0.9292. The visual segmentation result is shown in Fig. 6. We could learn that SAB shows up the target space region and that CAB pays attention to the edge information. Our SCA-Net achieves better segmentation result.

E. SKIN LEISION SEGMENTATION

To assess the performance of our proposed SCA-Net, we first put its paces on the skin lesion segmentation dataset from ISIC 2018. The dataset contains 2594 images with their ground truth [44], [45]. The skin lesion boundaries vary in scale, shape and color, necessitating automated segmentation methods to be extremely sensitive to these variations [46].

We present the comparison of our method with other state-of-the-art networks. The comparison was made with seven existing networks, including U-Net [8], ResUNet [48], U-Net++ [25], CE-Net [31], Attention-UNet [13], FCA-Net [17] and Singh *et al.* [18]. All of them are adopted with the original implementation, and the soft dice loss function is used uniformly.

TABLE 1. The quantitative comparison of ablation analysis. The value of dice, IoU, ASSD and RAVD are based on mean standard deviation.

Method	Dice	IoU	ASSD	RAVD
U-Net(backbone)	0.8970±0.0545	0.8207±0.0637	1.1360±1.4196	-0.0275±0.1320
U-Net + SAB	0.9165±0.0332	0.8520±0.0511	0.7076±0.6722	-0.0804±0.1035
U-Net + CAB	0.9194±0.0312	0.8566±0.0477	0.6182±0.6252	-0.0364±0.1061
SCA-Net	**0.9292±0.0361**	**0.8730±0.0570**	**0.5838±0.0570**	**-0.0061±0.1491**

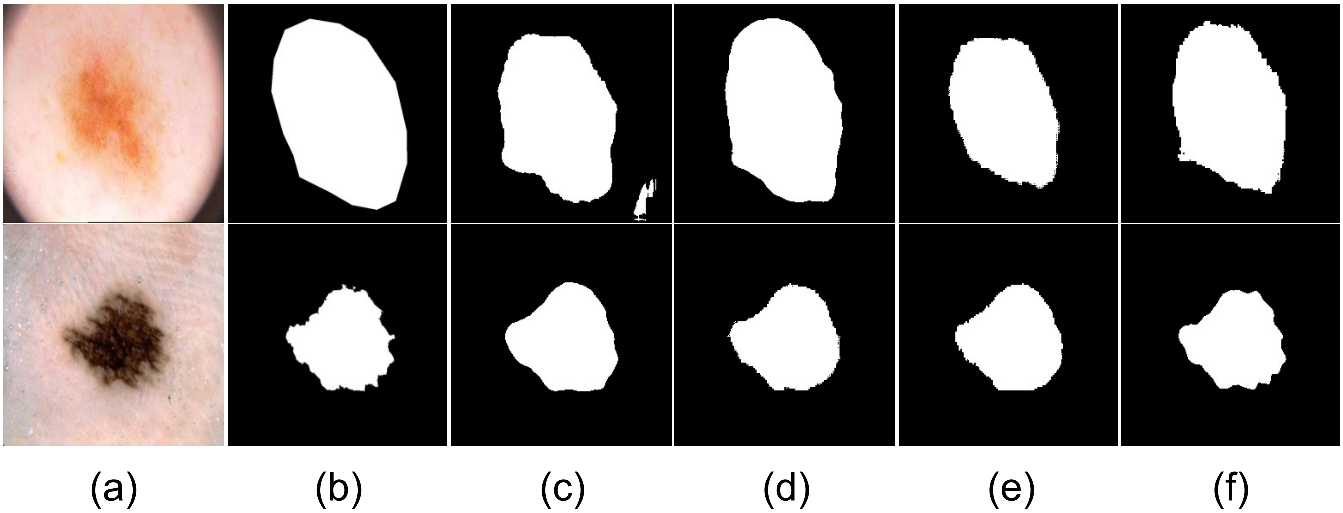


FIGURE 6. Visual comparison between different attention modules for skin lesion segmentation. (a) Original image. (b) Ground truth. (c) Backbone: U-Net with residual block. (d) Backbone with SAB (spatial attention block). (e) Backbone with CAB (channel attention block). (f) Our proposed SCA-Net.

TABLE 2. The result of skin lesion segmentation compared with other state-of-the-art methods.

Method	Dice	IoU	ASSD	RAVD
U-Net [6]	0.8893±0.0375	0.8094±0.0560	0.9848±0.7377	-0.0427±0.1442
ResUNet [48]	0.8932±0.0485	0.8168±0.0707	1.1501±1.6196	-0.0289±0.2222
UNet++ [25]	0.9096±0.0440	0.8420±0.0659	1.1613±1.6653	-0.0422±0.1632
CE-Net [31]	0.8983±0.0391	0.8237±0.0588	0.9859±0.9537	-0.1433±0.1290
Attention-UNet [13]	0.9035±0.0461	0.8326±0.0677	1.3076±1.969	-0.0219±0.2256
FCA-Net[17]	0.9222±0.0357	0.8615±0.0550	0.6481±0.8056	0.0568±0.1572
Singh et al. [18]	0.9214±0.0443	0.8612±0.0656	0.8291±1.4779	-0.0432±0.1609
SCA-Net	**0.9292±0.0361**	**0.8730±0.0570**	**0.5838±0.0570**	**-0.0061±0.1491**

The properties describing the detailed results are displayed in Table 2. We calculated the means and standard deviation of the four assessed metrics in all experiments. Fig. 7 shows the visual segmentation results, and it is obvious that our framework outperformed other state-of-the-art methods in skin lesion segmentation. Our SCA-Net achieved a Dice score of 0.9292, an IoU score of 0.8730, an ASSD of 0.5079 and an RAVD of -0.0061 for skin lesion segmentation. The training parameters of U-Net [6] have 20.96 M but our network only has 13.36 M, showing that the complexity of the model is higher than our model, but our network performs better. From the sample performance images, the segmentation results show that other state-of-the-art networks produce missegmentation due to color and hair interference. Fig. 8 depicts the Bland–Altman plots for the comparison difference between the segmentation areas of the ground truth and automatic segmentation methods. Compared with Singh *et al.* [18], our proposed method has a lower average deviation, which illustrates that our model is much more robust.

F. THYROID GLAND SEMGENTATION

We conducted the following evaluation task: thyroid gland segmentation [47], which consists of sixteen records of 3D volumes and their matching ground truth. To match our network input format, the 3D volumes and their corresponding ground truth were split into 4762 individual slices with the shape of 256 × 256. According to the ground truth marked by the sonographer, the unmarked slices were removed from the dataset. Finally, 3999 images and the corresponding ground

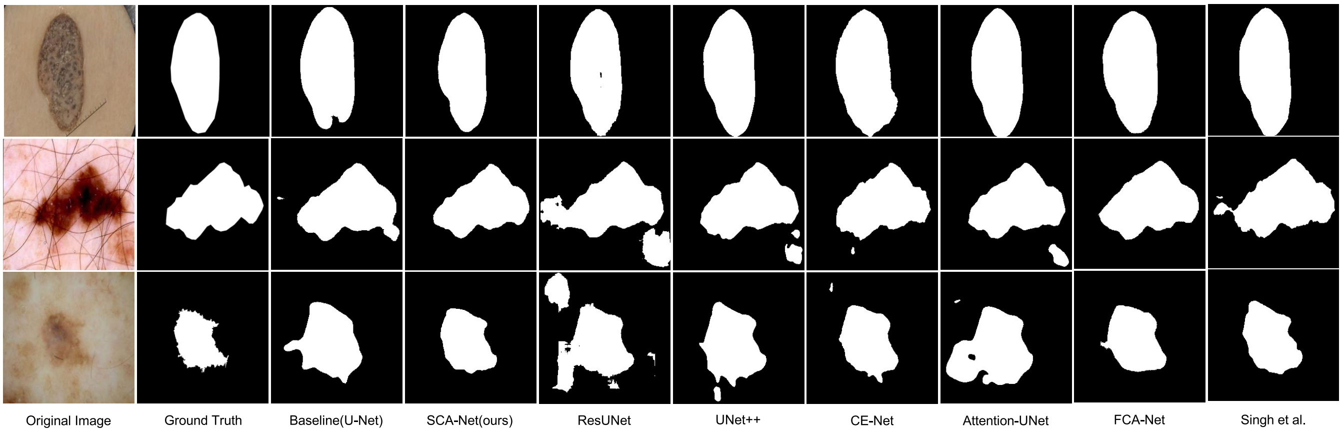


FIGURE 7. Visual comparison between SCA-Net and state-of-the-art networks for skin lesion segmentation. Our proposed SCA-Net has obvious performance compared with other segmentation performance.

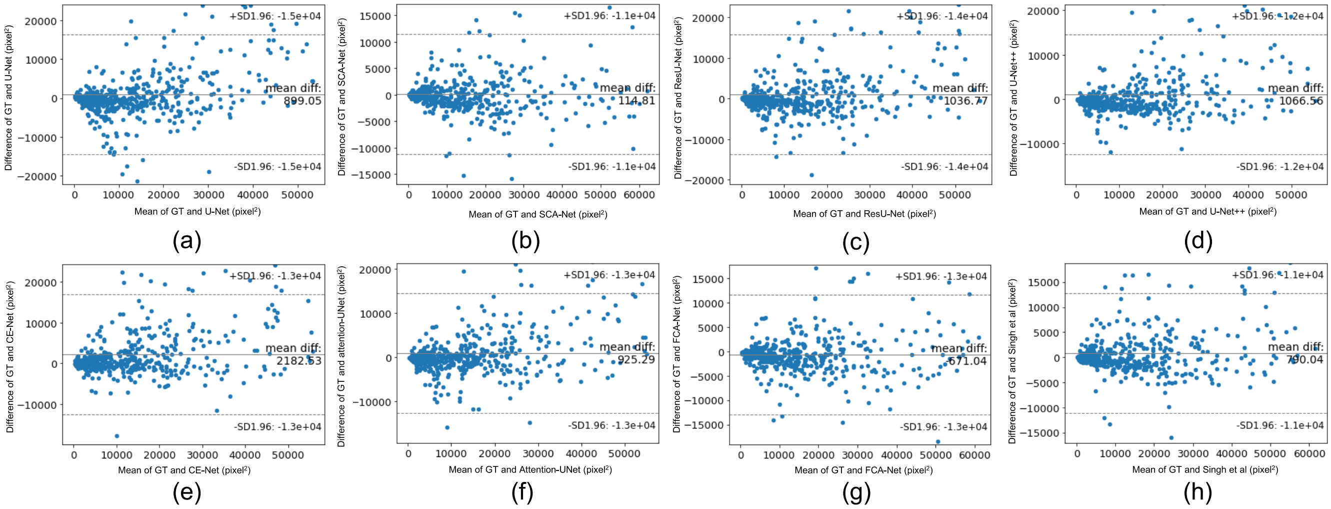


FIGURE 8. Bland-Altman plots for area comparisons between ground truth (GT) and automated segmentation results of skin lesion.

truth were screened. The main challenge of this task is the diversity of thyroid tissue size and morphology in the thyroid ultrasound images. The complexity of the peripheral tissue also affects the segmentation performance.

The result shown in Table 3 shows that our network is successful and achieves higher efficiency compared to other state-of-the-art networks. Our proposed network outperformed U-Net with a Dice score of 0.9825, an IoU score of 0.9661, an ASSD score of 0.0508, and a RAVD score of 0.0021. U-Net can segment the general outline of thyroid glands. However, it lacks the ability to segment both blurred and prominent edges. ResUNet has a better performance than U-Net, in which the residual connection enhances the segmentation ability. CE-Net, U-Net++, FCA-Net and Singh *et al.* have slight oversegmentation and undersegmentation, respectively. Compared with other networks, our model can segment the details of the thyroid edge. We show some samples of segmentation results for visual comparison in Fig. 9. In Fig. 10, all automatic methods dealing with the thyroid gland segmentation task present consistency with manual segmentation, and our proposed method performs better with a lower average deviation and smaller dispersion.

G. PANCREAS SEGMENTATION

Pancreas segmentation is the last experimental task. This dataset comes from The Nation Institutes of Health Clinical Center, which consists of 82 abdominal contrast-enhanced 3D CT scans from 53 male and 27 female subjects. Pancreases corresponding to the ground truth were manually slice-by-slice segmented by a medical student and inspected by an experienced radiologist. The anatomical structure of the pancreas is complex, and it is mainly located in the posterior peritoneum, with very high shape and volume variability in morphology among different slices. It is surrounded by adjacent tissues, and these tissues are close to the pancreas in CT images, which causes blurring of segmentation boundaries. Together with the noise of the CT

TABLE 3. The result of thyroid gland segmentation compared with other state-of-the-art methods.

Method	Dice	IoU	ASSD	RAVD
U-Net	0.9402±0.0085	0.8923±0.0136	0.2921±0.0726	0.1278±0.0408
ResUNet	0.9520±0.0103	0.9120±0.0171	0.2368±0.1061	0.0393±0.0460
UNet++	0.9453±0.0113	0.9098±0.0181	0.3141±0.1649	-0.0265±0.0504
CE-Net	0.9555±0.0057	0.9176±0.0097	0.2048±0.0577	-0.0147±0.0283
Attention-UNet	0.9582±0.0090	0.9225±0.0153	0.1971±0.1101	0.0099±0.0379
FCA-Net	0.9640±0.0045	0.9325±0.0077	0.1273±0.0305	0.1185±0.0189
Singh et al.	0.9665±0.0096	0.9371±0.0164	0.1378±0.1003	0.0465±0.0356
SCA-Net	**0.9825±0.0033**	**0.9661±0.0062**	**0.0508±0.0217**	**0.0021±0.0108**

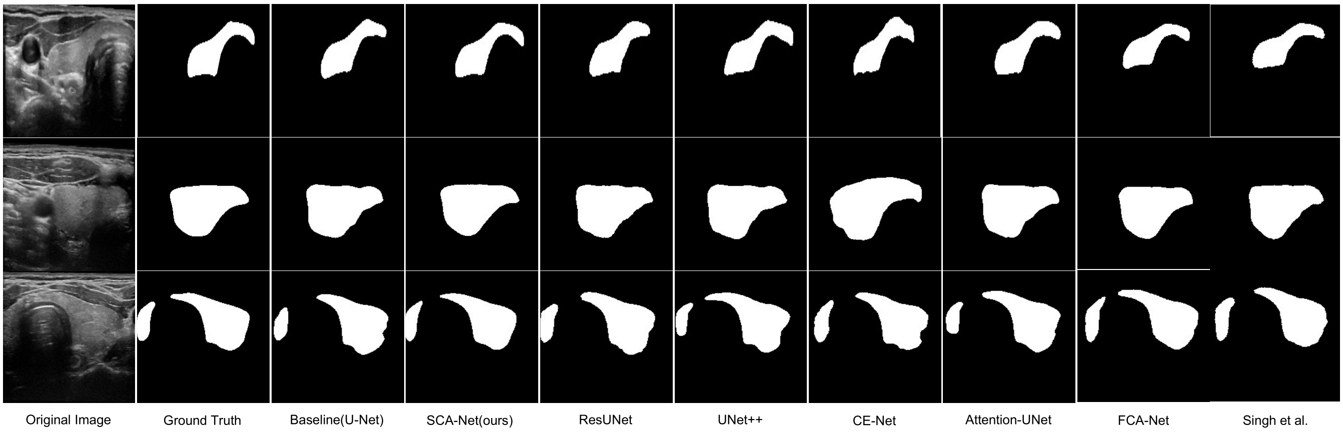


FIGURE 9. Visual comparison between SCA-Net and state-of-the-art networks for thyroid gland segmentation. Our proposed network achieves the best performance. The U-Net++ and CE-Net have incomplete segmentation or excessive segmentation, separately.

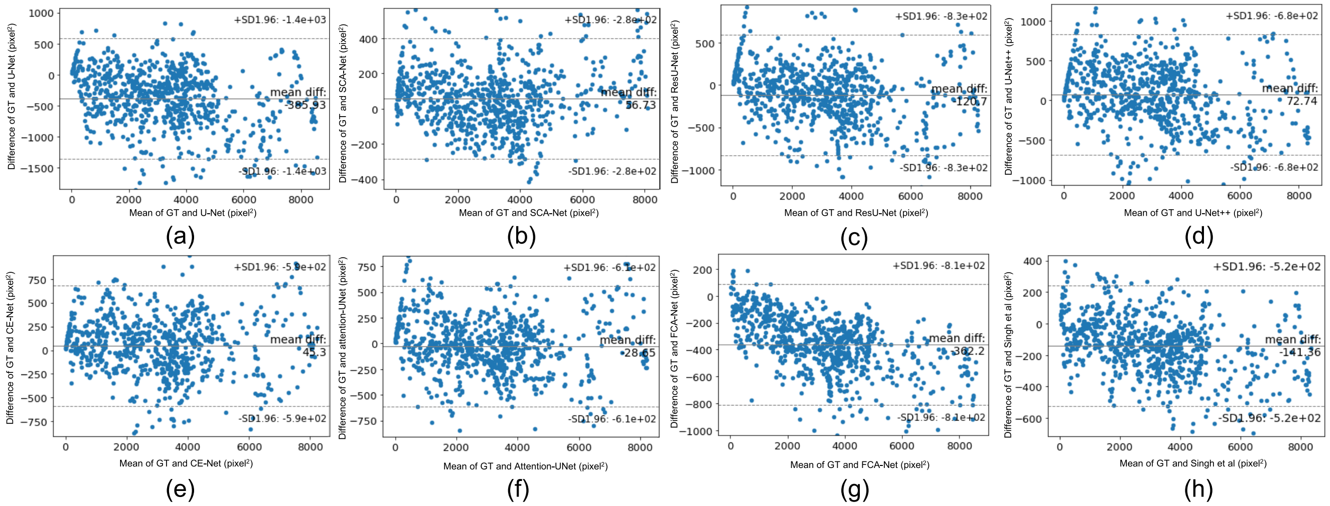


FIGURE 10. Bland-altman plots for area comparisons between ground truth (GT) and automated segmentation results of thyroid gland.

images themselves, local body effects and the influence of tissue motion, pancreas segmentation is a very challenging problem.

The result is displayed in Table 4. Our network has a better performance than other state-of-the-art networks. The model obtained the best Dice score of 0.9137, IoU score of 0.8530, ASSD of 0.3079 and RAVD of −0.0069. The samples of segmentation results for visual comparison are illustrated in Fig. 11, and the Bland–Altman plots of these methods are presented in Fig. 12. In comparison with the segmentation samples of other state-of-the-art networks, we find that SCA-Net is slightly worse in complex boundary segmentation than CE-Net and Singh *et al.*, which has more parameters and better fitting segmentation boundaries. However, they are more complex than our network, and our model excels at focusing on specific target areas. Although our SCA-Net has

TABLE 4. The result of pancreas segmentation compared with other state-of-the-art methods.

Method	Dice	IoU	ASSD	RAVD
U-Net	0.8798±0.0257	0.8072±0.0322	0.7958±0.5592	-0.0727±0.1037
ResUNet	0.8744±0.0239	0.8000±0.0301	0.7824±0.5200	-0.0527±0.1086
UNet++	0.8866±0.0236	0.8160±0.0311	0.5605±0.3278	-0.0800±0.0850
CE-Net	0.9054±0.0151	0.8411±0.0212	0.3739±0.2093	-0.0194±0.0532
Attention-UNet	0.9061±0.0159	0.8423±0.0223	0.3696±0.1883	-0.0282±0.0616
FCA-Net	0.9081±0.0151	0.8450±0.0214	0.3762±0.1999	-0.0382±0.0554
Singh et al.	0.9049±0.0235	0.8411±0.0311	0.4538±0.8066	-0.0180±0.0922
SCA-Net	**0.9137±0.0124**	**0.8530±0.0179**	**0.3079±0.1279**	**-0.0069±0.0475**

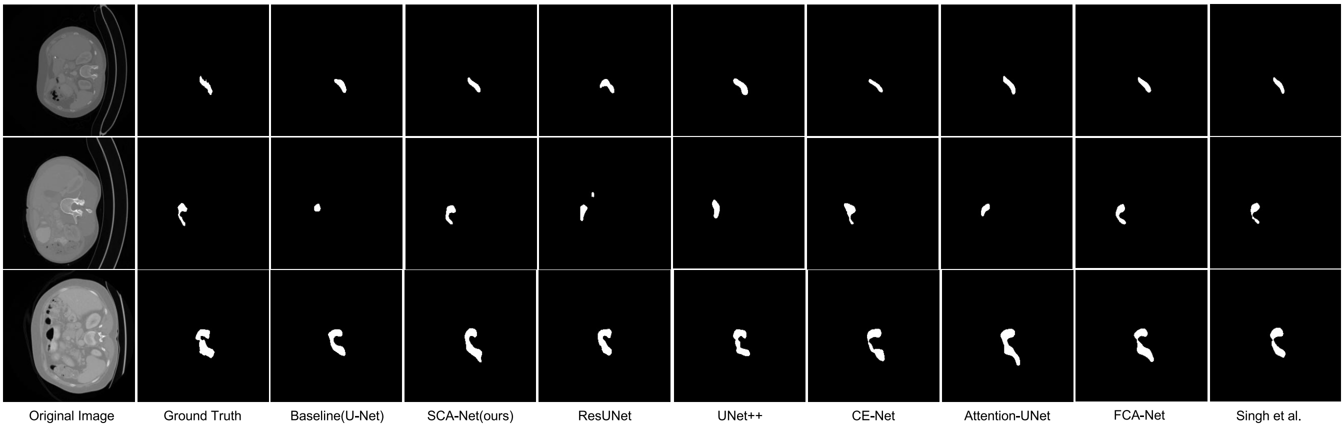


FIGURE 11. Visual comparison between SCA-Net and state-of-the-art networks for pancreas segmentation. With the complex segmentation task, our network has a better performance compared with other networks.

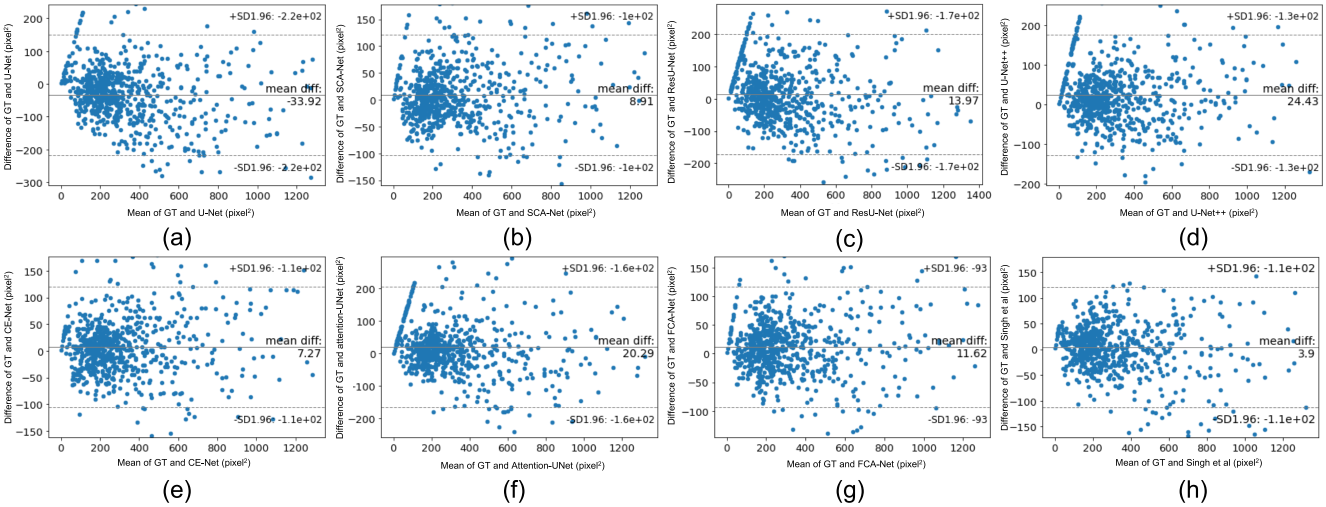


FIGURE 12. Bland-altman plots for area comparisons between ground truth (GT) and automated segmentation results of pancreas.

a slightly higher confidence interval than FCA-Net in Fig. 12, the difference is not obvious, and our proposed network has a lower bias.

V. DISCUSSION

For medical image segmentation tasks, better segmentation results help clinicians make a considerable preclinical diagnosis and assist them in clinical treatment. The variety of shapes, sizes and target locations, such as skin lesions, requires the network to have strong robustness. Original methods based on CNNs produced many channel feature maps and saved important features relying on simple concatenation operations. However, the relevant information is not utilized efficiently between multiscale spatial and channel features. The component structure of attention mechanisms handles the relevant feature maps, which improves the performance of segmentation tasks. Thus, we conceive a novel framework for medical image segmentation. The SAB connects the high- and low-level information from multiple stages to produce more representative contextual features. Additionally, the CAB redistributes the channel feature responses and strengthens important channel features.

To further verify the validity and robustness of the model, we conducted tests in three different medical image domains, including RGB images, MRI slices and ultrasound images. Compared with state-of-the-art networks, our SCA-Net has a significant improvement over three representative datasets, which shows that SCA-Net has better performance for different medical image segmentation tasks. We are more interested in applying our network to 3D data in the future.

We also find that our SCA-Net outperforms other networks in the thyroid gland segmentation task, but there are no significant segmentation differences. The reason we believe is that the boundary and the shape of the thyroid gland have small differences, and the distribution of locations is similar. Our proposed network can discern the boundary effectively. Compared with the skin lesion segmentation task, the color of the ultrasound image is gray, which may make it easier to learn the characteristics of the thyroid gland. In the pancreas segmentation tasks, SCA-Net scored significantly higher than the other networks. Our network shows more effectiveness of segmentation. Compared with other state-of-the-art methods, SCA-Net has fewer parameters and higher efficiency.

VI. CONCLUSION

Medical image segmentation tasks are crucial for clinical analysis and diagnosis. Due to the large variation in shape and texture of segmented targets, higher demands are placed on the robustness and performance of medical image segmentation networks. We introduced a spatial and channel attention network (SCA-Net) in this study, aiming to enhance the segmentation performance of medical image segmentation methods. Specifically, we design the SAB to consider the multiscale spatial information and the CAB to recalibrate the channel information. We train our SCA-Net, and the result demonstrates the superiority of our method in different tasks, including skin lesion segmentation, thyroid gland segmentation and pancreas segmentation. Our model can be used in a new application by fine-tuning using a new dataset and the manual ground truth.

In this paper, we conducted three experiments to verify the effectiveness of our network on 2D medical images. In possible future work, we will develop an extension to process 3D data.

REFERENCES

[1] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci. Rep.*, vol. 10, no. 1, p. 13724, Dec. 2020.

[2] T. Lei, R. Wang, Y. Wan, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," 2020, *arXiv:2009.13120*.

[3] S. Pandey, H. Tekchandani, and S. Verma, "A literature review on application of machine learning techniques in pancreas segmentation," in *Proc. 1st Int. Conf. Power, Control Comput. Technol. (ICPC T)*, Raipur, India, Jan. 2020, pp. 401–405.

[4] X. Yao, Y. Song, and Z. Liu, "Advances on pancreas segmentation: A review," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 6799–6821, Mar. 2020.

[5] H. Kumar, S. V. DeSouza, and M. S. Petrov, "Automated pancreas segmentation from computed tomography and magnetic resonance images: A systematic review," *Comput. Methods Programs Biomed.*, vol. 178, pp. 319–328, Sep. 2019.

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

[7] 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.

[8] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2018, pp. 1529–1537.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, 2015, pp. 2048–2057.

[11] Y. Pu, M. Min, Z. Gan, and L. Carin, "Adaptive feature abstraction for translating video to text," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, New Orleans, LO, USA, Feb. 2018, pp. 7284–7291.

[12] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," 2017, *arXiv:1703.03130*.

[13] O. Oktay, J. Schlemper, L. Le Folgoc, 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," Apr. 2018, *arXiv:1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[15] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: Object context network for scene parsing," 2018, *arXiv:1809.00916*.

[16] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9215–9223.

[17] V. K. Singh, M. Abdel-Nasser, H. A. Rashwan, F. Akram, N. Pandey, A. Lalande, B. Presles, S. Romani, and D. Puig, "FCA-Net: Adversarial learning for skin lesion segmentation based on multi-scale features and factorized channel attention," *IEEE Access*, vol. 7, pp. 130552–130565, 2019.

[18] V. K. Singh, M. Abdel-Nasser, F. Akram, H. A. Rashwan, M. M. K. Sarker, N. Pandey, S. Romani, and D. Puig, "Breast tumor segmentation in ultrasound images using contextual-information-aware deep adversarial learning framework," *Expert Syst. Appl.*, vol. 162, Dec. 2020, Art. no. 113870.

[19] G. O. Young, "Synthetic structure of industrial plastics," in *Plastics*, vol. 3, 2nd ed., J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.

[20] K. Zhou, Z. Gu, W. Liu, W. Luo, J. Cheng, S. Gao, and J. Liu, "Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 2724–2727.

[21] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 267–275.

[22] A. Krizhevsky, I. Sutskever, and G. H. Imagenet, "ImageNet clasification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Procesing Syst.*, Lake Tahoe, NA, USA, 2012, pp. 1097–1105.

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

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

[25] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[26] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput. Med. Syst. (CBMS)*, Rochester, MN, USA, Jul. 2020, pp. 558–564.

[27] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, Aug. 2019.

[28] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.

[29] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (DV)*, Oct. 2016, pp. 565–571.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[31] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[32] J. Zhang, Y. Jin, J. Xu, X. Xu, and Y. Zhang, "MDU-Net: Multi-scale densely connected U-Net for biomedical image segmentation," 2018, *arXiv:1812.00352*.

[33] P. Zhang, W. Liu, Y. Lei, H. Lu, and X. Yang, "Cascaded context pyramid for full-resolution 3D semantic scene completion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7801–7810.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 346–361.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[37] 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.

[38] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 375–383.

[39] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[41] Y. Qin, K. Kamnitsas, S. Ancha, J. Nanavati, G. Cottrell, A. Criminisi, and A. Nori, "Autofocus layer for semantic segmentation," in *Proc. MICCAI*, Sep. 2018, pp. 603–611.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image cognition," in *Proc. CVPR*, 2016, pp. 770–778.

[43] Y. Wu and K. He, "Group normalization," 2018, *arXiv:1803.08494*.

[44] N. Codella, V. Rotemberg, P. Tschandl, M. Emre Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[45] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," 2018, *arXiv:1803.10417*.

[46] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.

[47] T. Wunderling, B. Golla, P. Poudel, C. Arens, M. Friebe, and C. Hansen, "Comparison of thyroid segmentation techniques for 3D ultrasound," in *Proc. Med. Imag., Image Process.*, Orlando, FL, USA, Feb. 2017, Art. no. 1013317.

[48] X. Han, "Automatic liver lesion segmentation using a deep convolutional neural network method," 2017, *arXiv:1704.07239*.

TONG SHAN was born in Baicheng, China, in 1997. He is currently pursuing the M.A.Eng. degree in biomedical engineering with the University of Shanghai for Science and Technology. His research interests include medical image processing and machine learning.

JIAYONG YAN was born in China, in 1975. He received the bachelor's degree from the Department of Biomedical Engineering, Xi'an Jiaotong University, in 1994, and the Ph.D. degree from the Department of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2004. He is currently an Associate Professor with the Department of Biomedical Engineering, Shanghai University of Medicine and Health Sciences, Shanghai. His research interests include medical signal and image processing and medical instrument design.