RSEA-Net: Residual Squeeze and Excitation Attention Network for Medical Image Segmentation

Shufen Liang
Wuyi University
Tian Wang
Wuyi University
Chen Chen
Wuyi University
Huilin Liu
Wuyi University
Chuanbo Qin
Wuyi University
Yue Feng (yfeng_wyu@wyu.edu.cn)
Wuyi University

Research Article

Keywords: Medical image segmentation, Squeeze and Excitation, Attention mechanism, Receptive field

Posted Date: March 30th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1419097/v1

RSEA-Net: Residual Squeeze and Excitation Attention Network for Medical Image Segmentation

Shufen Liang[1], Tian Wang[1] ,Chen Chen[1] ,Huilin Liu[1] ,Chuanbo Qin[1],Yue Feng[1*]

(1. Department of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China)

***Correspondence**: Yue Feng, e-mail:yfeng_wyu@wyu.edu.cn, Department of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China

Abstract

Background: In recent years, convolutional neural networks have been prominent in medicine image processing, but single convolution and frequent pooling operations very easily produce redundant information or miss key information. This paper designs a residual squeeze and excitation attention network (RSEA-Net) to solve the above problems.

Methods: The network has the following main advantages: (1) It can capture richer and more detailed features at different scales in our proposed step convolution module (SCM), which adopts a parallel structure design and has receptive fields of different sizes. (2) This paper also designs a residual squeeze and excitation attention module (RSEAM), which can improve the useful feature gain through space and channel. It can not only eliminate some redundant information but also improve the overall robustness of the model.

Results: This paper verifies the performance of the RSEA-Net in 2D lung CT and tongue image databases. This paper selects seven models for comparison. The experimental results showed that: The RSEA-Net has multi-size respective fields and can eliminate redundant information. In the 2D lung CT database, the Accuracy, Jaccard coefficient, and Dice coefficient reach 0.9939, 0.9705, and 0.9850, respectively. In the tongue image database, the Accuracy, Jaccard coefficient, and the Dice coefficient of the RSEA-Net reach 0.9954, 0.9794, and 0.9895, respectively, which are better than those of the other seven models.

Conclusion: We propose a new deep network model (RSEA-Net). The structure is composed of two U-shaped networks with left and right layers. These are precoding networks and precision segmentation networks. To avoid losing information or producing invalid information caused by frequent convolution and pooling operations, this paper designs an SCM to obtain more multi-scale features. This paper also designed an RSEAM, which can improve the useful feature gain through space and channel, remove some redundant information, and improve the network's overall robustness. Finally, we also reduced the number of down-sampling operations and simplified the longitudinal complexity. Experimental results show that our method is superior to the original U-Net and other state-of-the-art methods in two different datasets.

Keywords: Medical image segmentation; Squeeze and Excitation; Attention mechanism; Receptive field

Background

Since manual labeling of a large number of medical images is a very complicated task, it is prone to deviations. Therefore, automatic medical image segmentation is a very important step in medical diagnosis and analysis. For example, tongue image segmentation is a very important first step in the objectification of tongue diagnosis [1]. A large number of manual annotations may include additional redundant information, because the surface color of the tongue is similar to the lip, thereby affecting the correctness of the subsequent classification. Traditional medical image segmentation methods mainly include: threshold segmentation method [2,3], edge detection-based method [4,5], region-based segmentation method [6], based on specific theory segmentation method [7,8,9], but these methods generally have the following shortcomings: (1) Those methods are often sensitive to light changes, and can only adapt to controllable environments. (2) The edge segmentation effect is poor.

In recent years, convolutional neural networks (CNNs) have been widely used in medical image segmentation given their rapid development, especially the rise of fully convolutional networks (FCN) [10] and U-Net [11] in semantic segmentation. For example, in terms of tongue

segmentation, Xue et al. [12] used FCN to obtain the initial area of the tongue image and then combined it with the traditional method to optimize it. Lin et al. [13] trained a ResNet-based method to generate an end-to-end tongue segmentation model. To segment lung CT images, Skourt et al. [14] used the classic U-Net framework. Zhang et al. [15] further integrated an incepate-Res module and dense convolution module into the U-Net structure, which can train the segmentation model for medical images. For medical image segmentation, Gu et al. [16] proposed a context encoder network(CE-Net) to capture and save more high-level information and spatial information.

Most of the above methods are based on the classic U-Net framework. However, we cannot meet precision requirements when using only the U-Net structure to segment lesions because of the diverse lesion shapes and different organ structures.

Some scholars improved the U-Net codec structure [17,18]. For example, Wang et al. [17] proposed a deep neural network with a coordination guide for chest CT image segmentation. The lung area is first extracted using CT image segmentation technology. Then, the volume convolutional nerve was used to segment the pulmonary lobes. In another example, Wang et al. [18] adopted the residual structure for the input and output modules and the global aggregation module with an external nested residual structure for the upper and lower sampling process to reduce the loss of information, due to a single convolution operation.

In summary, compared with the classic U-Net, the improved model has a certain degree of improvement in segmentation accuracy and speed and can also adapt to more complex background conditions. However, these methods are not still suitable for solving U-Net’s problem on the production of redundant information and loss of key information after successive down-sampling and single-convolution operations. This paper proposes a residual squeeze and excitation attention network (RSEA-Net) to address these problems, and the main contributions of the model are as follows:

(1) We design a feature extraction module (SCM) with a parallel structure and receptive fields of different sizes, which can capture more detailed features of different scales.

(2) We also propose a residual squeeze and excitation attention module (RSEAM), which can enhance the useful feature gain and eliminate redundant information through space and channel.

(3) We experiment RSEA-Net for image segmentation in tongue and 2D lung CT datasets, whose results show that our method is superior to the original U-Net and other state-of-the-art methods.

Methods

The dataset

This paper verifies the performance of the RSEA-Net in 2D lung[1] CT and tongue[2] image databases.

The 2D lung CT images used in this article are from the Lung Nodule Analysis (LUNA) competition. The dataset contains 534 original images and labeled images (the size is 512×512).

The Tongue Image Database (TONDAT) contains 600 original images and label images (the size is 768×576). Tongue segmentation is an important first step in the objectification of tongue diagnosis. Accurate segmentation will directly affect the correctness of the subsequent tongue classification.

Related works

A. Multi-scale feature extraction

To further improve U-Net’s performance, some scholars improve the feature extraction bottleneck modules in codec units. Xiao et al. [19] introduced the residual structure based on the bottleneck module, which could reduce the partial information loss of the convolution process, and solve the problem of gradient explosion or gradient disappearance during training. To ensure better feature representation for segmentation tasks, Md et al. [20] proposed a Recurrent Convolutional Neural Network based on U-Net as well as a Recurrent Residual Convolutional Neural Network based on U-Net which are named RU-Net and R2U-Net respectively. Among them, RU and R2U modules are improved convolution variants, and they can gain richer feature accumulation. Nabil et al. [21] were inspired by the inception module [22]. They firstly connected 3×3, 5×5, and 7×7 convolutional layers in parallel to achieve multi-layer resolution analysis. Then, a smaller and lighter 3×3 convolution block is used to approximately replace the 5×5, 7×7 convolution operation, and the MultiRes module is proposed. Also learning from the inception module, Zhang et al. [23] combined it with the residual module and proposed an inception-ResNet-V2 module, which deepened the width of the network and could solve the problem of huge changes in the size of objects in medical images. Wang et al. [24] substituted it with four attention modules. Most of these improved modules use residual structure to obtain more detailed features, thereby enhancing the final segmentation effect.

B. Attention and gating mechanisms

In medical imaging, lesions or human organs are the main research objects. However, the structure of human organs is

1 https://www.kaggle.com/kmader/finding-lungs-in-ct-data/data/.
2 https://github.com/BioHit/TongeImageDataset.

quite different, the types of lesions are diverse, and the background is complex, which makes it difficult to segment medical images. Therefore, during the segmentation process, it is necessary to focus on target features and suppress irrelevant features. To enhance the effective features gain, "Attention" or "Squeeze and Excitation" modules are added generally to the codec. Those modules can be used to excite through space or channel.

The attention mechanism is derived from the study of human vision. The human eye is always accustomed to paying attention to the objects of interest, thereby ignoring other visible information. From the perspective of the computer, attention can be interpreted as biasing computing resources towards the most informative signal part of the method, the advantages of this mechanism have been proven in a series of applications [25-28]; The "Squeeze and Excitation" module was first proposed by Hu et al. [29] The initial design goal was by simulating the interdependence between channels to improve the performance of the network. Roy et al. [30] first introduced the "Squeeze and Excitation" module into various compiling networks, which can decompose the spatial correlation through the global average pool. It can learn channel-specific descriptors and recalibrate the feature gain to emphasize useful channels.

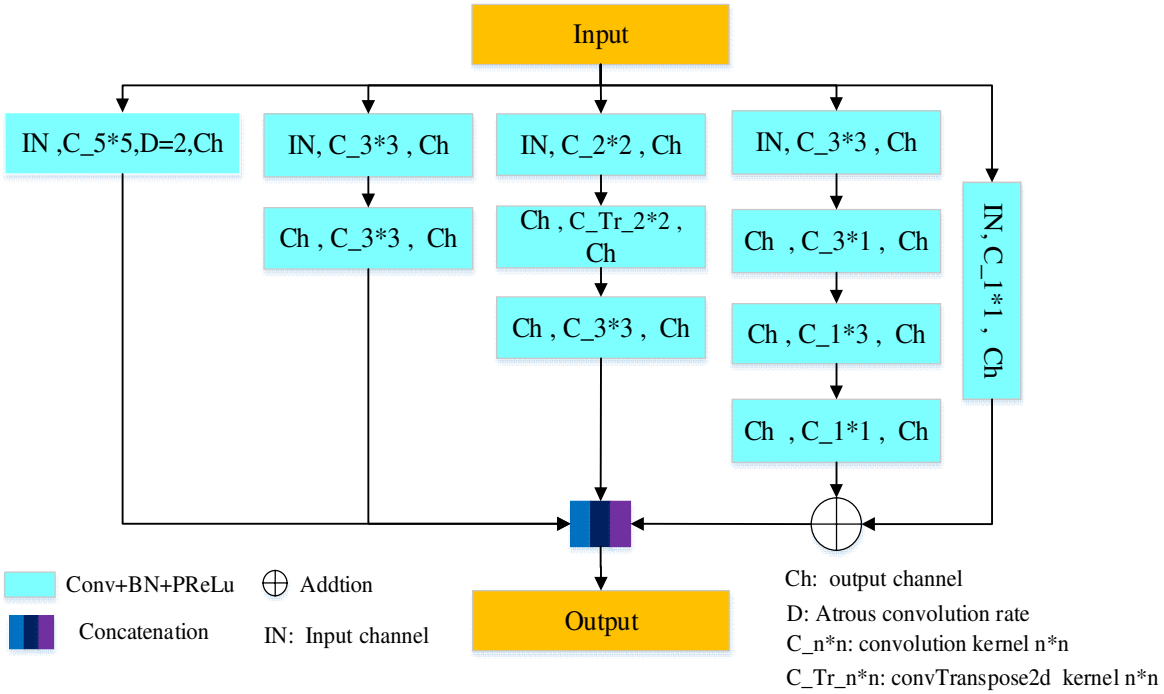

Fig. 1. The architecture of SCM.

Proposed Method

A. SCM

Semantic segmentation is done mainly through the continuous alternating use of convolution and pooling operations to gain local receptive fields of various depths and different global information. However, the segmentation objects of medical images are manifold, but also the divergence in the structure of the different organs is more prominent. To adapt to these complex segmentation situations and obtain more detailed information, this paper adopts the design method of parallel structure and uses the step convolution module (we call it the step convolution module, SCM for short because convolution in the module branch of convolution kernels increases gradually and is shaped like a ladder) to capture multi-scale context information. The module has four branches, and specific parameters are shown in Fig 1.

The above-mentioned convolutional module has strong adaptability and can be embedded in most deep neural networks; Therefore, we can derive different benefits from a variety of convolutional modules. To expand the receptive fields and control the amount of calculation in the feature extraction process, the first branch of SCM uses a 5×5 atrous convolution with a rate of 2. The atrous convolution was originally used to effectively calculate wavelet transform. Chen et al. [31] first introduced the concept of atrous convolution in deeplab v1 and thus extended the design of the deeplab v family [32].

The second branch of SCM still adopts a classic bottleneck module, which can identify the most simple contexts. Various networks use it for the feature extraction

module.

The successive pooling operations tend to lose part of key information, so the third branch of SCM substitutes the first convolution of the classic bottleneck with one convolution and one deconvolution (both with a size of 2×2). The 2×2 convolution is used to extract features, and the 2×2 deconvolution is applied to restore the part of lost information.

The last branch replaces the second convolution in the classic bottleneck with two successive convolutions with the sizes of 3 ×1 and 1×3. It is demonstrated that the n×n convolution is divided into n×1 and 1×n convolutions, obtaining a significantly improved performance [33]. For example, the 3×3 convolution is equivalent to 3×1 and 1×3 convolutions. We found that performance improved by 33%. In addition, to prevent gradient explosion or disappearance in the process of training, residual structures are added.

The experimental results illustrate that the proposed SCM feature extraction module performs better than the bottleneck module and even various improved convolution modules [19-21,34], as shown in Fig 2.

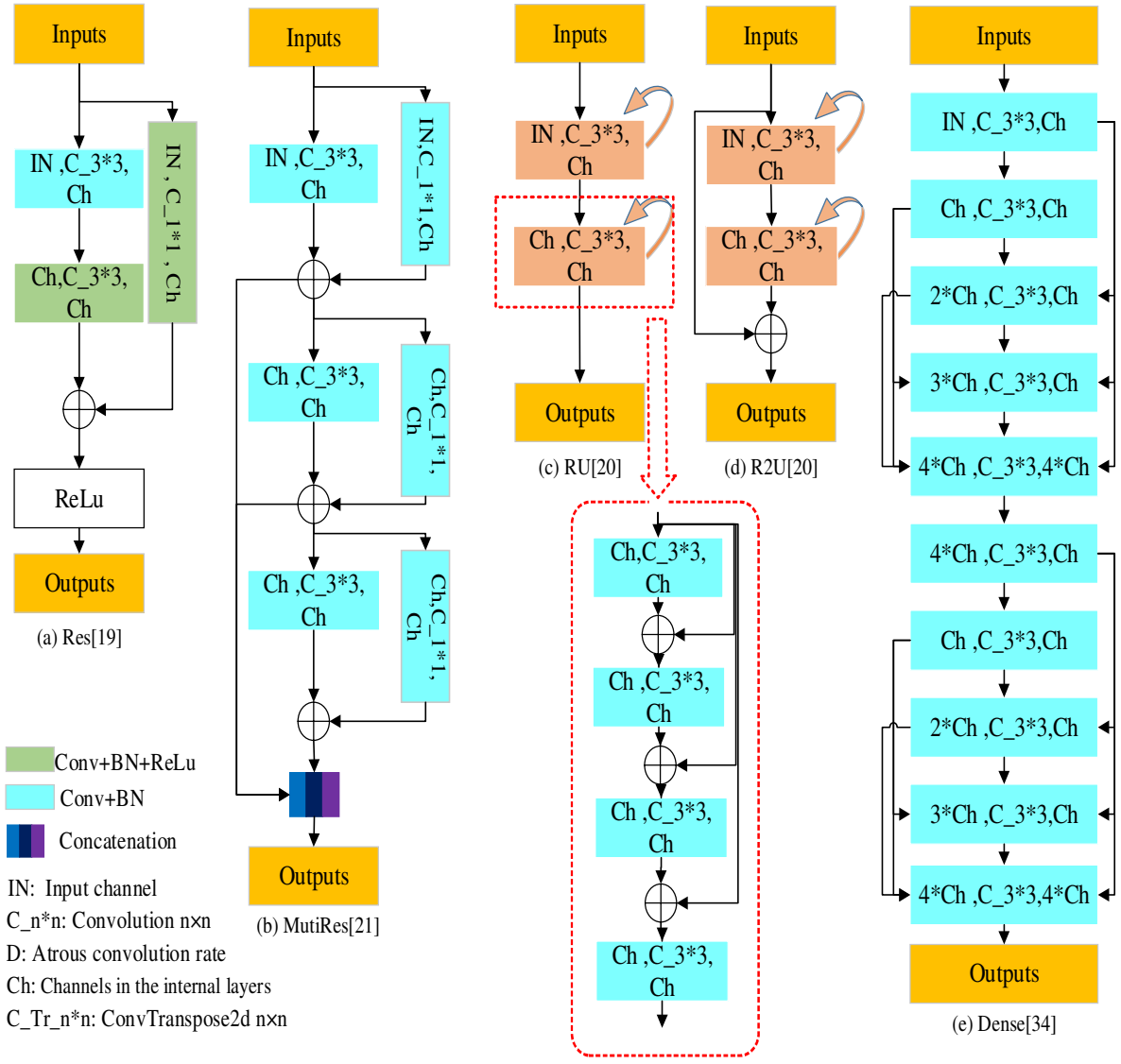

Fig. 2. The diagrams of the various improved convolution module structure

B. SEAM

The classic codec structure's basic components are the convolutional layers, which can extract each channel's space and channel information layer by layer. This information is gradually mapped to the high-dimensional image feature space. Finally, end-to-end classification is realized based on the semantics of each pixel. However, in the process of layer-by-layer convolution operation, redundant information is easily produced. To solve this problem, we combine the

"Squeeze and Excitation" component with an attention mechanism and introduce a residual structure. Therefore, an RSEAM (shown in Fig 3) is proposed. It can increase the useful characteristic gain through the space and channel of the gate signal(GS) and eliminate part of the redundant information of the input feature(X). It also uses residual structures to recover some of the missing key information to improve the overall robustness of the network.

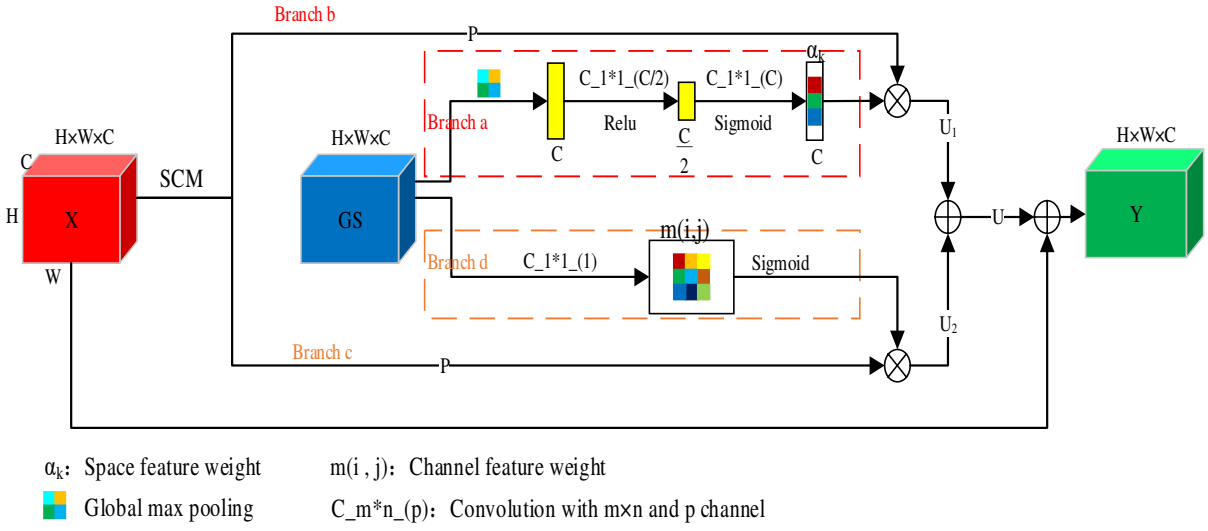

Fig. 3. The architecture of RSEAM.

1) Space Squeeze and Channel Excitation: Branches a and b represent the process of space squeezing and channel excitation, respectively. Suppose the input feature $X \in R^{H\times W\times C}$ passes through an encoding or decoding unit (the SCM feature extraction module is used in this article) to generate the output feature map $P \in R^{H\times W\times C}$ (H and W represent the length and width of the feature images, respectively, and C represent the number of channels). GS is regarded as a channel combination($G_k = \{g_1, g_2, \mathrm{K}, g_k\}$), and each feature map is gathered to a point using the global max pooling function to perform spatial compression to generate a vector $q_k \in R^{1\times 1\times C}$.

$$q_k = MAX(\sum_{i}^{H}\sum_{j}^{W} G_k(i,j)) \tag{1}$$

To obtain the recalibrated feature gain, the vector q_k is further subjected to compression and ReLu activation function mapping operations and restored to its original length. Finally, we use the sigmoid function to obtain the adjusted feature weights restricted to the interval [0, 1]:

$$\alpha_k = [\delta(F_c(q_1)), \delta(F_c(q_2)), ..., \delta(F_c(q_k))] \tag{2}$$

$F_c(\cdot)$ represents the operation process of convolution compression and ReLu function activation.$\delta(\cdot)$ represents the activation process of the sigmoid function. α_k represents a vector of feature weight. The vector α_k is used as an incentive to act on the output feature $P \in R^{H\times W\times C}$, and a new feature U_1 is calculated:

$$U_1 = \alpha_k \times P(k = C, p \in R^{H\times W\times C}) \tag{3}$$

2) Channel Squeeze and Spatial Excitation: Branches c and d represent the channel squeezing and spatial excitation processes, respectively. Assuming that the GS is still $G_k = \{g_1, g_2, \mathrm{K}, g_k\}$, channel squeezing is used to compress on the channel. We extrude the number of the original channel C to the number of Channel 1:

$$m(i,j) = \sum_{k=1}^{C} s_k \times G_k(i,j) \tag{4}$$

This formula represents the linear combination of GS at the spatial position (i, j); Sk denotes the channel coefficient. m ($m \in R^{H\times W}$) is a plane, similar to the Gs projection, the projection by a sigmoid function, used to generate the weight of the characteristics of the space restrictions on [0, 1] again, as incentives, used to recalibrate output characteristic map P:

$$U_2 = [\delta(m(1,1)\times p(1,1)), \delta(m(1,2)\times p(1,2)), ..., \delta(m(H,W)\times p(H,W))] \tag{5}$$

We then combine the "Space Squeeze and Channel Excitation" and "Channel Squeeze and Spatial Excitation" processes. The formula is as follows:

$$U = U_1 + U_2 \tag{6}$$

3) Residual structure: This paper uses SCM as the feature extraction module and sets up an attention mechanism. We use the excitation and squeeze module to adjust the threshold of attention. We also emphasize the useful features of input X and filter some redundant information. Although the above operations can improve local performance, they may

also cause the degradation of high-resolution features. Therefore, this paper introduces the residual structure in RSEAM and directly cascades the inputs X and U to recover some of the information lost due to information screening or the expansion of the perception field. The final output feature (Y) is as follows:

$$Y = U + X \tag{7}$$

II. RSEA-Net

The U-Net mainly operates to obtain image features through a series of convolutional and pooling operations. But successive convolution tends to produce some redundant information. For obtaining multi-scale features of different depths, the model needs to be downsampled many times during the encoding process. Although part of the loss information is recovered by deconvolution in the subsequent decoding process, some key information is still irreversibly lost during the downsampling process. To solve the above problems, this paper designs a two-layer superimposed RSEA-Net based on the U-Net framework (shown in Fig 4). The model is divided into two modules: a predicted network used to provide gate signals and a precise segmentation network using the RSEAM.

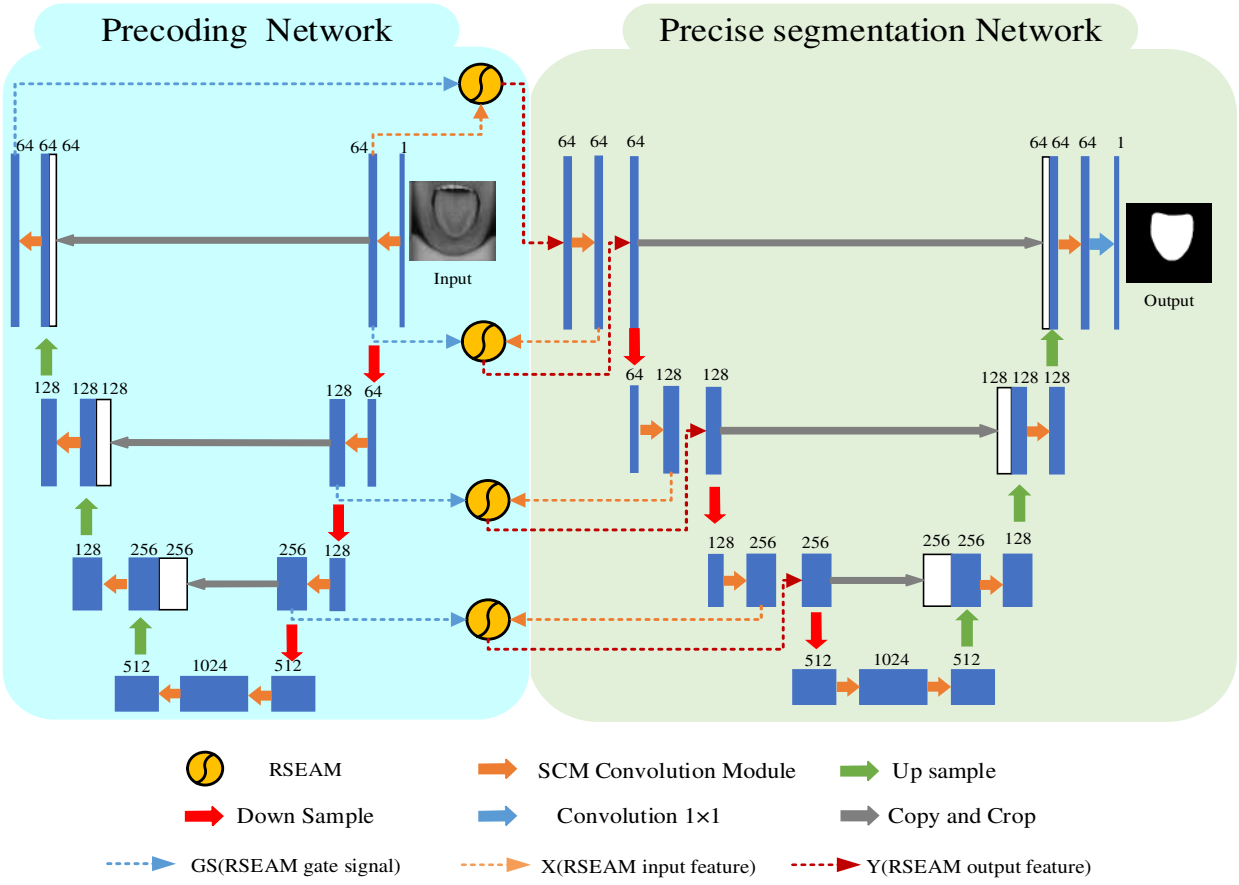

Fig. 4. RSEA-Net framework

1) Predict network: To solve the loss of key information caused by multiple down-sampling, the precoding network first reduces the number of down-sampling of the U-Net from four to three times. However, to ensure that sufficient multi-scale features are obtained, the bottom layer still uses two SCMs for feature extraction. It can make the convolution module of the precoding network use the same number as U-Net. Then, to obtain feature maps of receptive fields of different sizes, this paper replaces the bottleneck module of the U-Net with the SCM. The precoding network is mainly used to prepare for the subsequent precise segmentation.

2) Precision segmentation network: To readjust the feature weight of X, and get a more accurate segmentation effect, the RSEAM is used repeatedly in the precise segmentation network. First, we take the final output of the precoding network as the GS of the RSEAM. Then, the original image after one SCM convolution is used as the input X of the RSEAM. Finally, the output Y of RSEAM is taken as the input of the precise segmentation network. The RSEAM applies the space and channel of the GS to increase the useful feature gain and suppresses some invalid information. Therefore, it makes the final cut edge more precise and softer. RSEAM has been used a total of 4 times in RSEA-Net. By analogy, all GS of the RSEAM corresponds to the decode unit in the precoding network.

To prevent the "Squeeze and Excitation" module of RSEAM from suppressing too much useful information, a residual structure is introduced in the RSEAMs. In other words, the input feature X and output feature Y is directly cascaded to restore the key information. The precise segmentation network is also consistent with the precoding network in the number of down-sampling and still uses two SCM at the bottom layer for feature extraction.

In summary, the RSEA-Net can acquire rich multi-scale features and effectively suppress some invalid information. In addition, the SCM used in this model has strong adaptability. It is no need to use any pretraining methods. The segmentation effect of the applied SCM on the U-Net framework is better compared with those of other different improved convolution modules. It more easily adapts to different working environments. Table 1 shows the detailed configuration of various improved convolution modules (corresponding to the improved convolution modules introduced in Section 2.1 above) individually applied to the U-Net framework.

"Res", "MutiRes", "RU", "R2U", "Dense", and "SCM" in Table 1 represent different kinds of improved convolution modules. "E-" is the encoding process. "D-" denotes the decoding process. "IN" shows the input channel number. "CH" denotes the number of channels in the internal layers of convolutional modules. "O" indicates the output channel number. Taking the SCM u-net model in Table 1 as an example, we use the SCM module designed in this chapter to replace the classic bottleneck module of u-net. Compared with the classic u-net model, the depth, frame shape, up and down sampling methods of the SCM u-net model remain unchanged. The only difference is that in the feature extraction process, the classic module of bottleneck composed of convolution is replaced by the SCM designed in this chapter. The specific SCM U-net model framework is shown in Fig 5.

Table 1. Detailed configuration of various improved convolution modules individually applied to the U-Net framework.

Architecture with different blocks	Detailed configuration								
	E-1	E-2	E-3	E-4	E-5	D-1	D-2	D-3	D-4
Res U-Net[19]	IN:1	IN:64	IN:128	IN:256	IN:512	IN:1024	IN:512	IN:256	IN:128
	CH:64	CH:128	CH:256	CH:512	CH:1024	CH:512	CH:256	CH:128	CH:64
	O:64	O:128	O:256	O:512	O:1024	O:512	O:256	O:128	O:64
MutiRes U-Net[21]	IN:1	IN:64	IN:128	IN:256	IN:512	IN:1024	IN:512	IN:256	IN:128
	CH:64	CH:128	CH:256	CH:512	CH:1024	CH:512	CH:256	CH:128	CH:64
	O:64	O:128	O:256	O:512	O:1024	O:512	O:256	O:128	O:64
RU-Net[20]	IN:1	IN:64	IN:128	IN:256	IN:512	IN:1024	IN:512	IN:256	IN:128
	CH:64	CH:128	CH:256	CH:512	CH:1024	CH:512	CH:256	CH:128	CH:64
	O:64	O:128	O:256	O:512	O:1024	O:512	O:256	O:128	O:64
R2U-Net[20]	IN:1	IN:64	IN:128	IN:256	IN:512	IN:1024	IN:512	IN:256	IN:128
	CH:64	CH:128	CH:256	CH:512	CH:1024	CH:512	CH:256	CH:128	CH:64
	O:64	O:128	O:256	O:512	O:1024	O:512	O:256	O:128	O:64
Dense U-Net[34]	IN:1	IN:64	IN:128	IN:256	IN:512	IN:1024	IN:512	IN:256	IN:128
	CH:16	CH:32	CH:64	CH:128	CH:256	CH:128	CH:64	CH:32	CH:16
	O:64	O:128	O:256	O:512	O:1024	O:512	O:256	O:128	O:64
SCM U-Net(ours)	IN:1	IN:64	IN:128	IN:256	IN:512	IN:1024	IN:512	IN:256	IN:128
	CH:16	CH:32	CH:64	CH:128	CH:256	CH:128	CH:64	CH:32	CH:16
	O:64	O:128	O:256	O:512	O:1024	O:512	O:256	O:128	O:64

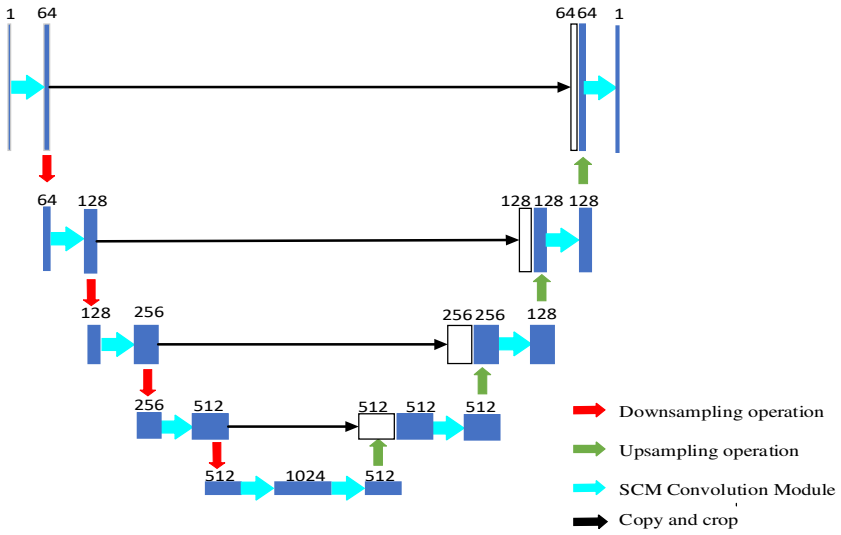

Fig. 5. SCM U-Net framework

Results and Discussion

In this section, this paper first introduces the experimental setup in the process of training and testing and then lists the image segmentation evaluation metrics used in this paper. Finally, it verifies the performance of the RSEA-Net in 2D lung[3] CT and tongue[4] image databases.

I. Implementation details

The optimization method in this article adopts the Adam optimizer and its parameter settings: batch size is 1 and 300 training rounds. The initial learning rate is set to 2e-3. In the training process, the data are amplified; that is, the input image adopts random horizontal flip, average subtraction, and random ratio operations. We set the random angle to $\{0°, 90°, 180°, 270°\}$.In the research experiment, the network framework uses Window10 operating system with 16GB memory, CPU environment of Intel(R) Core(TM) I9-10900K, and GPU 1080 Ti GPU acceleration environment of NVIDIA GeForce RTX. In the Python language environment, Pytorch has been used in CUDA version 10.0 general computing framework, and the cudnn acceleration module is used to accelerate the experimental training process.

II. Evaluation Metrics

To facilitate the analysis of the segmentation performance and test the segmentation effect, common indicators in the field of medical image segmentation include Accuracy, Sensitivity, Specificity, Precision, Jaccard Similarity, and the Dice coefficient. In formulae (8)-(13), True positive (TP) refers to the set of pixels correctly marked as positive; false positive (FP) means a set of pixels incorrectly labeled as positive; false negative (FN) represents a set of pixels incorrectly labeled as negative, and true negative (TN) represents a set of pixels that has been correctly labeled negative.

Accuracy is the proportion of all positive pixels in all pixels, which can measure the recognition ability of positive pixels by the segmentation method. The accuracy can be calculated as

$$Accuracy = \frac{sum(TP+TN)}{sum(TP+TN+FP+FN)} \tag{8}$$

The Sensitivity coefficient, also known as the recall rate, represents the proportion of all positive pixels with the correct pixel set, which can measure the segmentation method's ability to recognize positive pixels. The sensitivity coefficient can be calculated by the defined type (9).

$$Sensitivity = \frac{sum(TP)}{sum(TP+FN)} \tag{9}$$

The Specificity index, also known as the true negative rate, represents the proportion of the set of correctly labeled pixels in all negative pixels, which can measure the ability of the segmentation method to recognize negative pixels, and the coefficient is calculated by equation (10).

$$Specificity = \frac{sum(TN)}{sum(FP+TN)} \tag{10}$$

Precision refers to the proportion of pixel sets marked as positive that are actually positive. The Precision value is calculated by Formula (11), mainly to measure the model's ability to predict positive pixel points.

$$Precision = \frac{sum(TP)}{sum(TP+FP)} \tag{11}$$

Given two sets, positive and negative pixels A and B, the Jaccard coefficient represents the ratio of the intersection between sets A and B and the union between them. The

[3] https://www.kaggle.com/kmader/finding-lungs-in-ct-data/data/.
[4] https://github.com/BioHit/TongeImageDataset.

coefficient can be calculated by Formula (12).

$$Jaccard = \frac{sum(TP)}{sum(TP + FN + FP)} \tag{12}$$

The Dice coefficient of the range of values defined as [0, 1] presents the image overlap between the two target shape area of the total area ratio. The larger the coefficient, the better the segmentation result. Medical image segmentation often uses this index as the main evaluation metric, and the analysis of experimental results in this article also uses this as the main criterion. The coefficient is calculated by Formula(13).

$$Dice = \frac{sum(2 * TP)}{sum(2 * TP + FP + FN)} \tag{13}$$

III. Lung Segmentation

It is not necessary for the SCM designed in this article to use any pre-training methods. The SCM achieves a better segmentation effect when applied to the U-Net framework compared with other similar convolution modules. In this paper, we experimented on SCM and RSEA-Net which are compared with other convolution modules and state-of-the-art networks, respectively.

Table 2 shows the evaluation metrics of six improved convolution modules. The "Res" variant refers to the residual structure, solving the loss of information caused by multiple convolutions to a certain extent. Therefore, compared with the U-Net, the Accuracy, Jaccard similarity index, and Dice coefficient are slightly improved.

Because of the added 1×1 convolution, Res U-Net is 0.0081s slower than U-Net for the single-frame segmentation. The "MutiRes" variant uses repeatedly the residual structure to further improve the ability to extract detailed features, so it consumes a longer segmentation time. The "R2U" and "RU" variants combine residual connection and cyclic convolution, but because of excessively frequent convolution, the single-frame segmentation takes the highest time, and the Accuracy, Jaccard, and Dice coefficient are not high.

The "Dense" variant splices multiple levels of output features together, retaining more detailed features, so the Accuracy, Jaccard, and Dice coefficient are closer to the segmentation index of the "SCM". The "SCM" adopts a parallel structure and introduces the residual structure and atrous ideas to obtain multi-scale features. Therefore, the Accuracy, Jaccard, and Dice coefficients are highest, reaching 0.9939, 0.9698, and 0.9846, respectively, which are 0.16%, 0.83%, and 0.43% higher than in the U-Net. Moreover, it only takes 0.1104s to process the single-frame, which is slightly higher than the 0.0839s of the U-Net. Experiments show that the SCM has a stronger ability to extract features.

This paper selects seven models for comparison. FCN is the pioneer of semantic segmentation. U-Net introduces skip connection in its basis. SegNet improves the upsampling method in its decoding network. Therefore, U-Net and SegNet have improved Accuracy, Jaccard, and the Dice coefficient compared with FCN. Because FCN has two fully connected layers, it takes the most time among all similar models. Large kernel matters take the FCN structure as the basic framework and introduce a boundary refinement module with relatively strong detail-capture capabilities. It is also observed from Fig 6 that the edge segmentation of the large kernel matters model is relatively complete.

The AttU-Net introduces an attention mechanism that can suppress part of invalid information, and the segmentation effect is better. The DeepLabv3+ introduces the theory of atrous convolution, which guarantees the acquisition of multi-scale feature information while avoiding the problem of losing key information caused by multiple down-sampling operations. Therefore, the segmentation result is better than that of the U-Net.

Table 2 Evaluation metrics of different improved convolution module on LUNA.

Model Type	Sensitivity	Specificity	Precision	Jaccard	Accuracy	Dice Coefficient	Time(s)
U-Net[11]	0.9929	0.9924	0.9684	0.9615	0.9923	0.9803	0.0838
Res U-Net[19]	0.9898	0.9936	0.9741	0.9643	0.9927	0.9817	0.0919
MutiRes U-Net[21]	0.9911	0.9943	0.9773	0.9687	0.9936	0.9840	0.0909
RU-Net[20]	0.9947	0.9923	0.9691	0.9616	0.9926	0.9806	0.1980
R2U-Net[20]	0.9855	0.9946	0.9780	0.9640	0.9925	0.9816	0.2005
Dense U-Net[34]	0.9919	0.9938	0.9751	0.9673	0.9934	0.9833	0.1319
SCM U-Net(ours)	0.9907	0.9948	0.9788	**0.9698**	**0.9939**	**0.9846**	0.1104

Table 3 Comparison between the model designed in this paper and similar models on LUNA

Model Type	Sensitivity	Specificity	Precision	Jaccard	Accuracy	Dice Coefficient	Time(s)
FCN-8S[10]	0.9873	0.9929	0.9710	0.9590	0.9916	0.9790	0.2190
U-Net[11]	0.9929	0.9924	0.9684	0.9615	0.9923	0.9803	0.0838
SegNet[35]	0.9826	0.9942	0.9767	0.9600	0.9918	0.9795	0.1754
Large Kernel Matters[36]	0.9878	0.9949	0.9788	0.9670	0.9932	0.9831	0.1238
Deeplabv3+[32]	0.9870	0.9942	0.9764	0.9639	0.9926	0.9816	0.1104
AttU-Net[37]	0.9903	0.9946	0.9781	0.9687	0.9936	0.9840	0.0895
BASNet[38]	0.9906	0.9888	0.9631	0.9611	0.9929	0.9828	0.1856
RSEA-Net(Ours)	0.9910	0.9947	**0.9792**	**0.9705**	**0.9939**	**0.9850**	0.1933

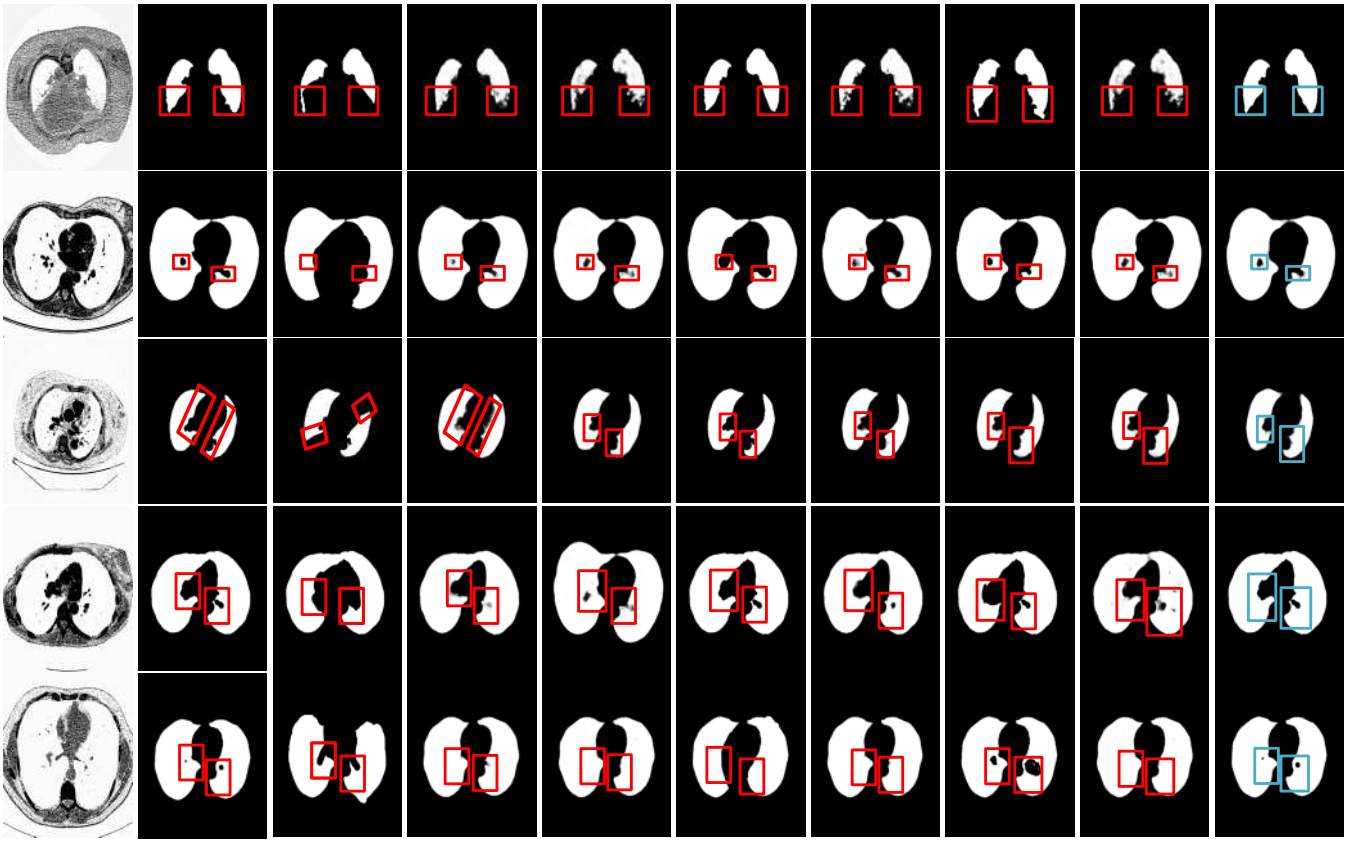

Fig. 6. Segmentation results of seven different models in LUNA. (a) Original image, (b) Labels, (c) FCN, (d) U-Net,(e) SegNet, (f) Large Kernel Matters, (g) Deeplab V3+, (h) AttU-Net , (i) BASNet, and (j) RSEA-Net (Ours).

The RSEA-Net has multi-size respective fields and can eliminate redundant information. It can be observed from Fig 6 that the current model has the best result and the highest accuracy for edge segmentation. The Accuracy, Jaccard coefficient,and Dice coefficient reach 0.9939, 0.9705, and 0.9850, respectively. It is the best among all

segmentation models.However, the single-frame segmentation time of the model reaches 0.1807s. It is only faster than FCN, because of the precoding network and SCM.

IV. Tongue Segmentation

In this section, we also experimented on SCM and RSEA-Net which are compared with other convolution modules and state-of-the-art networks, respectively.

Table 4 still gives six improved variants for comparison. The U-Net has a single-frame segmentation time of 0.0756 s, which is the shortest time consumed among other models. The SCM U-Net achieves the Accuracy, Jaccard coefficient, and Dice coefficient of 0.9948, 0.9763, and 0.9879, respectively. Among the above six improved variants, these values are the highest, segmentation result is the best. However, it reaches 0.1608s in single-frame segmentation time, more than twice U-net, due to the four parallel branches.

It is shown in table 5 that the Accuracy, Jaccard coefficient, and the Dice coefficient of the RSEA-Net reach 0.9954, 0.9794, and 0.9895, respectively, which are better than those of the other seven models. As seen from Fig 7 that because the color of the tongue is similar to that of the lip. There are two kinds of segmentation results: containing irrelevant regions (such as lips and teeth) or segmenting incompletely.

Nevertheless, the SCM has a strong detail processing ability, and the RSEAM can effectively suppress the non-target area. Therefore, the RSEA-Net model has a better edge segmentation effect and a higher segmentation accuracy. The more complex the model, the longer the segmentation time.

Discussion

The traditional codec network mainly uses a large number of convolution, up-sampling, and down-sampling operations to transform the dimensions, but these operations can easily bring redundant information to the network or lose some key information. In response to this, this article first optimizes the classic convolution module to obtain more detailed feature information. Then, an attention module is designed to fully filter some redundant information in the training process. Finally, a network with two U-shaped superimposed is proposed.

Table 4 Evaluation metrics of different improved convolution modules on TONDAT

Model Type	Sensitivity	Specificity	Precision	Jaccard	Accuracy	Dice Coefficient	Time(s)
U-Net[11]	0.9869	0.9924	0.9753	0.9646	0.9916	0.9812	0.0756
Res U-Net[19]	0.9873	0.9967	0.9874	0.9753	0.9945	0.9864	0.0905
MutiRes U-Net[21]	0.9861	0.9971	0.9890	0.9753	0.9946	0.9874	0.1571
RU-Net[20]	0.9899	0.9982	0.9932	0.9634	0.9918	0.9812	0.1640
R2U-Net[20]	0.9863	0.9914	0.9698	0.9736	0.9900	0.9842	0.1690
Dense U-Net[34]	0.9851	0.9937	0.9756	0.9542	0.9897	0.9742	0.1611
SCM U-Net(ours)	0.9874	0.9969	0.9885	**0.9763**	**0.9948**	**0.9879**	0.1608

Table 5 Comparison between the proposed models and similar models on TONDAT

Model Type	Sensitivity	Specificity	Precision	Jaccard	Accuracy	Dice Coefficient	Time(s)
FCN-8S[10]	0.9866	0.9892	0.9803	0.9682	0.9860	0.9774	0.1955
U-Net[11]	0.9869	0.9924	0.9753	0.9646	0.9916	0.9812	0.0756
SegNet[35]	0.9870	0.9959	0.9843	0.9716	0.9938	0.9855	0.1500
Large Kernel Matters[36]	0.9868	0.9954	0.9822	0.9694	0.9934	0.9843	0.1828
Deeplabv3+[32]	0.9847	0.9960	0.9849	0.9699	0.9934	0.9846	0.0655
AttU-Net[37]	0.9875	0.9969	0.9884	0.9761	0.9947	0.9879	0.0797
BASNet[38]	0.9856	0.9849	0.9542	0.9536	0.9922	0.9862	0.1451

RSEA-Net(Ours)	0.9871	0.9979	0.9921	**0.9794**	**0.9954**	**0.9895**	0.2098

Fig. 7. Segmentation results of seven different models in TONDAT. (a) Original image, (b) Labels, (c) FCN, (d) U-Net,(e) SegNet, (f) Large Kernel Matters, (g) Deeplab V3+, (h) AttU-Net , (i) BASNet, and (j) RSEA-Net (Ours).

The SCM designed in this paper has 4 branches. In addition to the branches of including the bottleneck module, the ideas of atrous convolution and residual structure are also introduced in other branches to obtain more context information. It is not necessary for the SCM designed in this article to use any pre-training methods. The SCM achieves a better segmentation effect when applied to the U-Net framework compared with other similar convolution modules. The results of tables 2 and 4 show that the dice coefficient of SCM U-Net is 0.26% and 0.53% higher than other improved convolution modules on average. However，it takes a longer time because the SCM structure is more complex.

This article also designs an RSEAM to eliminate redundant information. Among them, the "Squeeze and Excitation" module and attention mechanism are introduced to obtain new weights of feature information. It is also combined with the residual structure to recover some key information. The module has two inputs. The one is the gate signal(GS) used to control the feature gain. The other is the input signal (X), and finally, the re-adjusted weight feature signal(Y) is output. RSEA-Net uses 4 RSEAMs in total. The GS of each RSEAM is provided by a precoding network. Y is used as the input for the precise segmentation network. RSEA-Net also reduces the number of pooling operations to cut down the loss of information. The experimental results in Tables 2 and 4 show that the RSEA-Net, applying to tongue image segmentation and 2D lung computed tomography image segmentation, is superior to the U-Net and other state-of-the-art methods. It can be observed from the red rectangles in Fig 6 and 7 that the segmentation edges of RSEA-Net are softer and more precise, and the overall effect is better.

Conclusions

In this paper, we propose a new deep network model (RSEA-Net). The structure is composed of two U-shaped networks with left and right layers. These are precoding networks and precision segmentation networks. To avoid losing information or producing invalid information caused by frequent convolution and pooling operations, this paper designs an SCM to obtain more multi-scale features. This paper also designed an RSEAM, which can improve the useful feature gain through space and channel, remove some redundant information, and improve the network's overall robustness. Finally, we also reduced the number of down-sampling operations and simplified the longitudinal complexity. Experimental results show that our method is superior to the original U-Net and other state-of-the-art

methods in two different datasets. However, the RSEA-Net requires a precoding network to provide the gate signal and the SCM structure is complicated. It takes a long time to segment single-frame. In the future, we will try to combine different technologies with this model to further increase its speed and reduce its number of parameters.

Abbreviations
CT: Computed tomography;ReLU: Rectified Linear Unit.

Acknowledgements
The 2D lung CT images used in this article are from the

Lung Nodule Analysis (LUNA) competition.

Authors information

Shufen Liang master of control theory and control engineering, Professor. Graduated from Hebei University of technology in 2002.Worked in WuYi university.Her research interests include Pattern recognition,information processing and communication research.

Tian Wang graduate student of WuYi University. His main research interest is image processing.

Chen Chen graduate student of WuYi University. His main research interest is image processing.

Huilin Liu graduate student of WuYi University. Her main research interest is image processing.(e-mail:mookkkk@163.com)

Chuanbo Qin received the Ph.D. in pattern recognition and intelligence system from South China University of Technology, Guangdong, China, in 2015. He is currently a Associate Professor with the College of Intelligent Manufacturing, Wuyi University, Guangdong. His current research interests include medical image processing and pattern recognition.

Yue Feng received his Ph.D. from the Royal Melbourne Institute of Technology, Melbourne, Australia in 2006. His main research interests include computer vision and image processing.(e-mail: yfeng_wyu@wyu.edu.cn)

Authors' contributions
SL and YF designed the framework for Residual Squeeze and Excitation Attention Network for Medical Image Segmentation. TW and CC designed the experiments and analyzed the results. HL and CQ analyzed the experimental dataset. TW is major contributor in writing and editing the manuscript. SL and YF edited the manuscript. All authors read and approved the final manuscript.

Funding
The study was supported by the National Natural Science Foundation of China under Grant No. 33518001, Grant No.61372193, and Grant No.61901304, and Guangdong Natural Science Foundation under Grant No.07010869, and Basic Research and Applied Basic Research Key Project in General Colleges and Universities of Guangdong Province under Grant No.2021ZDZX1032. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. There was no additional external funding received for this study.

Data availability
The database used in this paper can be applied from this web page:
https://www.kaggle.com/kmader/finding-lungs-in-ct-data/data/.
https://github.com/BioHit/TongeImageDataset.

Ethics and consent to participate
The dataset used in this work is openly accessible and free to the public. No direct interaction with a human or animal entity was conducted in this work. And All procedures were performed in accordance with relevant guidelines

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Intelligent Manufacturing, Wuyi University, Jiangmen, 529020, China.

References

[1] Huang SQ, Zhang YL, Zhou J, et al. Research progress on the objectification, quantitation, and standardization of tongue manifestation in traditional Chinese medicine[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2017,32(4):1625-1627.

[2] Senthilkumaran N, Vaithegi S. Image Segmentation By Using Thresholding Techniques For Medical Images[J]. Computer Science & Engineering An International Journal, 2016,6(1):1-13.

[3] Shan J, Cheng H D, Wang Y. Completely automated segmentation approach for breast ultrasound images using multiple domain features[J]. Ultrasound in Medicine & Biology, 2012, 38(2):262-275.

[4] Gong M, Liang Y, Shi J, et al. Fuzzy C means clustering with local information and kernel metric for image segmentation[J]. IEEE Transactions on Image Process, 2013, 22(2):573-584.

[5] Cui Z, Zuo W, Zhang H, Zhang D.Automated Tongue Segmentation Based on 2D Gabor Filters and Fast Marching[J]. New York, NY: Springer, 2013:328-335.

[6] Kim K H, Do J H, Ryu H, et al. Tongue diagnosis method for extraction of effective region and classification of tongue coating[J]. Image Processing Theory, Tools and Applications, 2008. IPTA 2008. First Workshops on, 2008:1-7.

[7] Tao X, Chun M.X, Fei F, et al. A method of tongue image segmentation based on kernel FCM. 9th International Congress on Image and Signal Processing[J], BioMedical Engineering and Informatics. 2016: 319-324.

[8] Li X, Li J, Wang D. Automatic tongue image segmentation based on histogram projection and matting[J]. IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2014:76-81.

[9] Wentao X, Ratchadaporn K, Dong X, et al. An Automatic Tongue Detection and Segmentation Framework for Computer-Aided Tongue Image Analysis[C]//13th International Conference on e-Health Networking, Applications and Services. 2011:189-192.

[10]Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Trans Pattern Anal Mach Intell. 2015;39(4):640-651.

[11]Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[J]. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors MICCAI 2015. LNCS. Cham: Springer; 2015(9351): 234-241.

[12]Xue Y, Li X, Wu P, Li J, Wang L, Tong W, W. Automated tongue segmentation in Chinese medicine based on deep learning[C]//Paper presented at: Proceedings of the International Conference on Neural Information Processing. Lecture Notes in Computer Science. 2018:542-553.

[13]n B, Xle J, Li C, Qu Y. Deeptongue: tongue segmentation via resnet[C]//Paper presented at: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018(2018):1035-9.

[14]B. AA Skourt BA, El Hassani A, Majda A. Lung CT image segmentation using deep neural networks[C]. Procedia Computer Science. 2018(127):109-113.

[15]Zhang Z, Wu C, Coleman S, Kerr D. DENSE-INception U-net for medical image segmentation[J]. Computer Methods Programs Biomedicine. 2020(192):105395.

[16]GBates R, Irving B, Markelc B, Kaeppler J, Brown G, Muschel RJ, et al. Segmentation of Vasculature From Fluorescently Labeled Endothelial Cells in Multi-Photon Microscopy Images[J]. IEEE Trans Med Imaging. 2019;38(1):1-9.

[17]Wang W, Chen J, Zhao J, et al. Automated segmentation of pulmonary lobes using coordination-guided deep neural networks[J]. International Symposium on Biomedical Imaging; 2019: 1353-1357.

[18]Wang Z, Zou N, Shen D, Ji S. Non-local U-nets for biomedical image segmentation[J]. AAAI. National Conference on Artificial Intelligence. 2020,34(4):6315-6322.

[19]Xiao X, Lian S, Luo Z, Li SZ. Weighted res-UNet for high-quality retina vessel segmentation[C]//In: 9th International Conference on Information Technology in Medicine and Education (ITME). 2018:327-331.

[20]Alom Md. Zahangir, Mahmudul Hasan, Chris Yakopcic. Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. 2018; arXiv preprint arXiv:1802.06955.

[21]Ibtehaz N, Rahman MS. MultiResU-Net: rethinking the U-Net architecture for multimodal biomedical image segmentation[J]. Neural Networks. 2020(121):74-87.

[22]Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[J]. Computer Vision and Pattern Recognition, 2016:2818-2826.

[23]Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[C]//National Conference on Artificial Intelligence, 2016:4278-4284.

[24]Wang C, He Y, Liu Y, He Z, He R, Sun Z. Scleraseg net: an improved u-net model with attention for accurate sclera segmentation[C]//In: 2019 International Conference on Biometrics (ICB), 2019:1-8.

[25]Bluche T. Joint line segmentation and transcription for end-to-end handwritten paragraph recognition[J]. In NIPS, 2016.(2):25-30.

[26]Cao C., Liu X, Yang Y, Yu Y, Wang J, Wang Z, Huang Y, Wang L, Huang C, Xu W, Ramanan D, and Huang T. S.. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks[C]//In ICCV, 2015:2-8.

[27]Jaderberg M, Simonyan K, Zisserman A, and Kavukcuoglu K. Spatial transformer networks[J]. In NIPS, 2015.(1): 2-8.

[28]Miech A, Laptev I, and Sivic J. Learnable pooling with context gating for video classification. arXiv:1706.06905, 2017. 2

[29]Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//CVPR. 2018:2-10.

[30]Roy AG, Navab N, Wachinger C. Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks[C]//International conference on medical image computing and computer-assisted intervention, 2018:421-9.

[31]Chen L C, Papandreou G, Kokkinos I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. Computer Science, 2014(4):357-361.

[32]Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Trans Pattern Anal Mach Intell. 2018,40(4):834-48.

[33]Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision[C]//In: Proceedings of the IEEE conference on computer vision and pattern recognition 2016:2818-2826.

[34]Guan S, Khan AA, Sikdar S, Chitnis PV. Fully Dense UNet for 2-D Sparse Photoacoustic Tomography Artifact Removal[J]. IEEE J Biomed Health Informat. 2019;24(2):568-76.

[35]Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image

Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017:1-9.

[36]Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters-improve semantic segmentation by global convolutional network[C]//In: Proceedings of the IEEE conference on computer vision and pattern recognition,2017:4353-61.

[37]Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention U-net: learning where to look for the pancreas[J].Computer Vision and Pattern Recognition,2018(3):124-130.

[38]Qin X, Zhang Z, Huang C, Gao C, Dehghan M, Jagersand M, BAS-Net: boundary-aware salient object detection[C]//In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019:7479-7489.