Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Fully-channel regional attention network for disease-location recognition with tongue images☆

Yang Hu [a,b], Guihua Wen [a,*], Mingnan Luo [a], Pei Yang [a], Dan Dai [a], Zhiwen Yu [a], Changjun Wang [c], Wendy Hall [b]

[a] South China University of Technology, Guangzhou 510006, China
[b] University of Southampton, Highfield Campus, SO171BJ Southampton, United Kingdom
[c] Guangdong General Hospital, Guangzhou 510000, China

ARTICLE INFO

Keywords:
Tongue image modeling
Disease-position recognition
Convolutional networks
Regional detailed features
Attention mechanism

ABSTRACT

Objective: Using the deep learning model to realize tongue image-based disease location recognition and focus on solving two problems: 1. The ability of the general convolution network to model detailed regional tongue features is weak; 2. Ignoring the group relationship between convolution channels, which caused the high redundancy of the model.
Methods: To enhance the convolutional neural networks. In this paper, a stochastic region pooling method is proposed to gain detailed regional features. Also, an inner-imaging channel relationship modeling method is proposed to model multi-region relations on all channels. Moreover, we combine it with the spatial attention mechanism.
Results: The tongue image dataset with the clinical disease-location label is established. Abundant experiments are carried out on it. The experimental results show that the proposed method can effectively model the regional details of tongue image and improve the performance of disease location recognition.
Conclusion: In this paper, we construct the tongue image dataset with disease-location labels to mine the relationship between tongue images and disease locations. A novel fully-channel regional attention network is proposed to model the local detail tongue features and improve the modeling efficiency.
Significance: The applications of deep learning in tongue image disease-location recognition and the proposed innovative models have guiding significance for other assistant diagnostic tasks. The proposed model provides an example of efficient modeling of detailed tongue features, which is of great guiding significance for other auxiliary diagnosis applications.

1. Introduction

The amount of outpatients in Chinese hospitals is enormous [8,44]. The automated diagnostic assistant tools can markedly improve the efficiency of the doctor's diagnosis and treatment [22,31]. In the conventional medical system, tongue diagnosis is a crucial diagnostic method, which is painless and convenient [53]. In the past works, researchers have designed professional image acquisition equipment for tongue image taking [47,28,38] and proposed the pre-processing methods from color calibration, tongue segmentation, denoising, and so on [46,57], to provide doctors with high-quality tongue image and carry out the diagnosis by electronic terminals. These works did not analyze the tongue image and provide the auxiliary diagnostic results. Further, in some studies of tongue image analysis, [45] made the

☆ This study was supported by China National Science Foundation (Grant Nos. 60973083 and 61273363), Science and Technology Planning Project of Guangdong Province (Grant Nos. 2014A010103009 and 2015A020217002), Guangzhou Science and Technology Planning Project (Grant Nos. 201604020179, 201803010088), Guangdong Province Higher Vocational Colleges and Schools Pearl River Scholar Funded Scheme under Grant 2018, China National Natural Science Foundation (Grant Nos. 61722205, 61751205, 61751202, 61572199, 61502174, and U1611461), Science and Technology Planning Projects of Guangdong Province (Grant No. 2018B010107002).

* Corresponding author.
E-mail addresses: superhy199148@hotmail.com (Y. Hu), crghwen@scut.edu.cn (G. Wen), csluomingnan@mail.scut.edu.cn (M. Luo), cs.pyang@gmail.com (P. Yang), daidanjune@hotmail.com (D. Dai), zhwyu@scut.edu.cn (Z. Yu), gzwchj@126.com (C. Wang), wh@ecs.soton.ac.uk (W. Hall).

https://doi.org/10.1016/j.artmed.2021.102110
Received 6 May 2020; Received in revised form 6 April 2021; Accepted 11 May 2021
Available online 26 May 2021

statistical analysis on the color features of tongue image, [52,18,27] used the color, texture, geometry, and other features of tongue image to detect specific diseases such as diabetes and liver diseases. Disease-location is an essential diagnostic concept in Chinese medicine. It indicates the particular organ location of the illnesses. Studying automatic disease-location recognition plays a crucial assistant role in the diagnostic process. However, previous studies have relied on specialist design features, which are laborious and hard to cover all diseases. As far as we know, there is no research on disease-location recognition using the patient's body signs.

The convolutional neural networks (CNNs) are widely applied in computer vision [30,36], include the modeling of tongue images [29,19]. However, there is no application of tongue-based disease location recognition. Also, the modeling of the tongue image needs to capture detailed features on different parts of the tongue, like root, tip, and coating of tongue, which are often distributed in different visual areas and accompanied by complex noise. Traditional CNN models are inadequate to capture such scattered, extensive details of pathological features, their redundant structure also leads to monotony or even deviations of feature maps, seriously affecting the performance of disease location recognition. On the other hand, the vast networks often rely on massive data and immense computing power [9,41,15], while for medical data is not easy [23]. Moreover, in a limited training set, the internal construction of the basic CNN model lacks complementary collaboration. This defect makes the modeling of CNN low-efficient, leads to high structural risk and severe overfitting [54].

This paper studies the new task of automatic disease location recognition based on tongue image and uses deep learning methods to conduct experiments on the large-scale tongue image dataset. As shown in Fig. 1, the tongue image is input into the CNN model. After layer-by-layer modeling, the visual features of the tongue image are extracted automatically. Then multi-label pathological results can be obtained by fully-connected (FC) output layer, including heart, gallbladder, stomach, bladder, etc. This research has the following significance: 1. The disease-location recognition based on tongue image has a high application value. It can provide a reference for outpatient doctors and improve their efficiency. Besides, the diagnostic cases of renowned doctors are used as training data so that it can guide junior doctors; 2. Although tongue image is not the only basis for diagnosis, the studies of tongue image-disease location relations provide researchers valuable technical reference on the signs-based assisted diagnosis. It has great guiding significance; 3. The application of deep learning technology not only mine the links between tongue image and disease location but also promote the development of deep learning in the task of assistant diagnosis of Chinese medicine.

Technically, aiming at the low efficiency of tongue feature modeling by the general CNNs model, a novel structure based on fully-channel regional attention is proposed. It mainly contains two parts: 1. Stochastic regional pooling (SRP): the traditional global pooling squeezes the whole feature map as a signal, which ignores localized details. We intercept several local regions with different sizes and locations from each feature map, and we pool them separately to achieve simultaneous attention to the multi detailed features; 2. Inner-imaging channel-wise attention (InI): we rearrange the pooled signals of each convolution channel into inner-imaging matrices. Then, the convolution filters are applied to model the group relationships among the channels as well as regional visual features, and re-weight all feature maps. The main innovations of this study can be summarized as follows:

1. A new research task of disease-location recognition based on tongue image is researched, and we propose to apply deep learning methods to accomplish this task. The first large-scale clinical tongue image dataset with the diagnostic label is established, which contains more than $10,000$ anonymous tongue images and corresponding disease-location tags.
2. A novel inner-imaging channel-wise attention mechanism is proposed. The pooled signals of convolutional channels are projected onto pseudo-maps, and filters model the channel group relationships. So, the relations of CNN components can be captured more abundantly: not only the point-to-point links between feature maps but also their multiple group relationships. Finally, it reduces the redundancy of the CNNs model and improves its modeling efficiency.
3. A dynamic regional pooling mechanism is proposed. Several specific areas are intercepted from one feature map to form multiple concentrated signals and remove edge noise. Combine with the inner-imaging mechanism, the relationships between the detailed features from fully-channels are encoded. Such a design helps to represent the multi-detailed tongue features effectively.

The proposed two mechanisms have powerful universality on CNN structure and other CNN enhancement methods, like spatial attention mechanism.

The remainder of this paper is organized as follows. Section 2 discusses related works. Section 3 introduces the overall framework of the InI mechanism and its enhanced edition for residual networks. Section 4 presents our theoretical explanations for designing of the InI mechanism. Section 5 describes the experimental results and analysis. We conclude in Section 6.

2. Related works

Machine learning methods for tongue image modeling. Modeling tongue images based on machine learning is a sustained research topic [5,34]. [45] designs the professional features of tongue color and applies the support vector machine (SVM) method to classify tongue diagnostic results. [52] expands the previous work and considers more tongue features, like texture and geometric shape. [18] synthesizes the tongue features in diverse illumination environments, and it uses SVM to classify liver diseases. The researches above rely on expert features, which is not automatic enough and is not conducive to generalization. The deep learning method can automatically extract the features of tongue image, [29,19] use CNNs to realize body recognition and

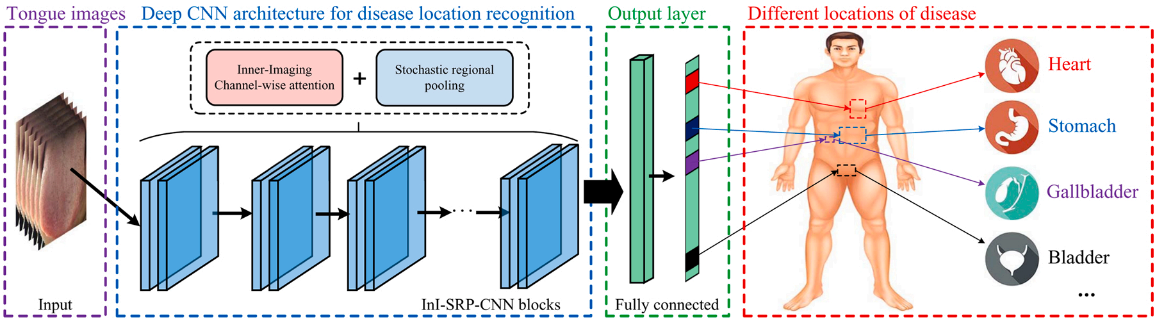
Tongue images
Deep CNN architecture for disease location recognition
Output layer
Different locations of disease
Inner-Imaging Channel-wise attention
+
Stochastic regional pooling
Input
InI-SRP-CNN blocks
Fully connected
Heart
Stomach
Gallbladder
Bladder
...

Fig. 1. Deep learning framework for disease-location recognition with tongue images.

herbal prescription generation, based on tongue image. However, these studies only directly run the routine CNN models, ignoring the detailed attentional features of tongue images. [27,42] focus on extracting the features of the tooth-marks in the tongue image, and modeled the captured details using CNN to identify the tooth-marks of the tongue image. However, they only capture the details of the tooth-marks on the edge of the tongue image, lack the modeling of other aspects of the tongue image features. Hence, they have a limited scope of application. Moreover, this step-by-step modeling method is not as automated as our end-to-end method.

Intelligent medical technology based on CNNs. In addition to tongue image modeling and analysis, CNNs also play an essential role in other medical assistant technologies [7,37,2]. Where [13,32,25] respectively apply CNNs to reconstruct CT images, extract organ regions and predict lung diseases. [35] propose a novel approach for 3D fully automatic and accurate breast tissue segmentation from MRI data. [50] combines CNN models and medical knowledge to recognize specific diseases. [3] uses the support vector machine after the CNN model and classifies hypercellularity. Although CNNs are well applied in the field of intelligent medicine, it still has a great space to improve in dealing with the detailed pathological features, and the CNN models are still affected by a lot of noise in real clinical data. [11] adds channel-wise attention on CNNs to classify skin diseases. However, it only applies one type of attention module, and it only considers using the random occluded visual area to improve the robustness of spatial feature modeling. Overall, the existing auxiliary medical models based on CNNs still have huge room for improvement, especially their ability to model detailed regional pathological features is not enough.

Attention mechanisms in CNNs. The attention mechanism is widely applied in the modeling of CNNs [33]. Usually, it aims to regulate the weights of visual-spatial areas [24]. The self-attention strategy is also used to modeling the global and long-range dependencies of CNN features [10]. Further, channel-wise attention is proposed to allocate weights for convolution channels effectively [17,26]. The differences between our study and the other attention mechanisms are 1. Stochastic regional area replaces the whole feature map as the source of channel concentration signal, which helps detailed local features to be added into channel relationship modeling; 2. The design of the inner-imaging structure is used so that filters can extract the diversified grouped channel relations.

3. Dataset

In this study, the first large-scale tongue image disease-location classification data set is established. All tongue images were collected from outpatients in various departments of the cooperative hospital. In order to collect tongue images in diverse environments as far as possible, the data collection lasted for 6 months, which is under different illumination, brightness, and shooting angles. The number of collected images reached 10,333. It makes our model more robust and does not rely on professional acquisition tools. Our shooting tools are ordinary digital cameras or smartphones. The disease-locations are labeled manually by outpatient doctors, the categories of which include: large intestine, gallbladder, lung, liver, bladder, spleen, kidney, stomach, small intestine, heart, health, and unknown [40]. Further, all labels of disease-location were double-checked by more than three chief physicians.

Naive pre-processing is conducted on the collected image, which is to automatically segment the tongue image area and remove the surrounding noise. Patients have consented all data acquisition work, and anonymous processing is carried out. Fig. 2 is the pretreatment schematic diagram of the tongue image.

Except acquisition and pre-processing of tongue images, other data information is listed in Table 1. The tongue images cover various types of diseases. The quality of the original image is guaranteed at more than 8 million pixels. The pre-processed image for tongue image detection is uniformly customized as fixed dimension RGB inputs. In each experimental run, we use the following data set cutting strategy: we randomly select 2000 samples from indices 1 to 2333 as the verification set and the test set (1000 samples are randomly taken out for each set), and the remaining 8333 is used as the training set. In a new round, 2000 samples are selected from the samples with indices 2001 $\sim$ 4333 as the verification set and test set, and so on: 4001 $\sim$ 6333, 6001 $\sim$ 8333, until 8001 $\sim$ 10,333.

Fig. 3(a and b) illustrates our experimental tongue data. Each tongue image may belong to multiple disease-location categories, and the amount distribution of each disease-location is shown as Fig. 3(c).

4. Proposed methods

The followed section elaborates on the research work of tongue image modeling and disease location recognition in three aspects: 1. The overall framework of disease location recognition based on the fully-channel regional attention network; 2. Technical details of the proposed architecture.

4.1. Overall framework

Task Description. Disease-location recognition can be defined as a multi-label classification problem. For set of disease-location categories

Table 1
Data informations of tongue image-disease location.

Information of tongue images	
Number of tongue images	10,333
Original image minimum shape	2600 × 3200
Intercepted tongue image shape (RGB)	224 × 224
Information of data argumentation	
Maximum cropping ratio (Original/Resize)	$224^2/256^2$
Picture rotation angle	180°
Information of disease position labels	
Number of labels	12
Average sample number of each position	2009
Number of health sample	124
Number of unknown sample	6

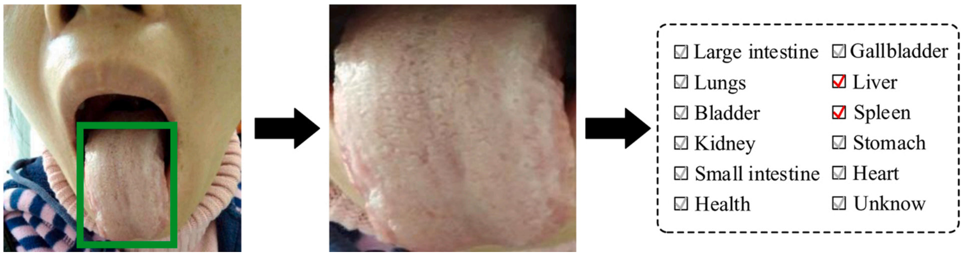

Fig. 2. Pre-processing of original tongue image and the corresponding disease-location label.

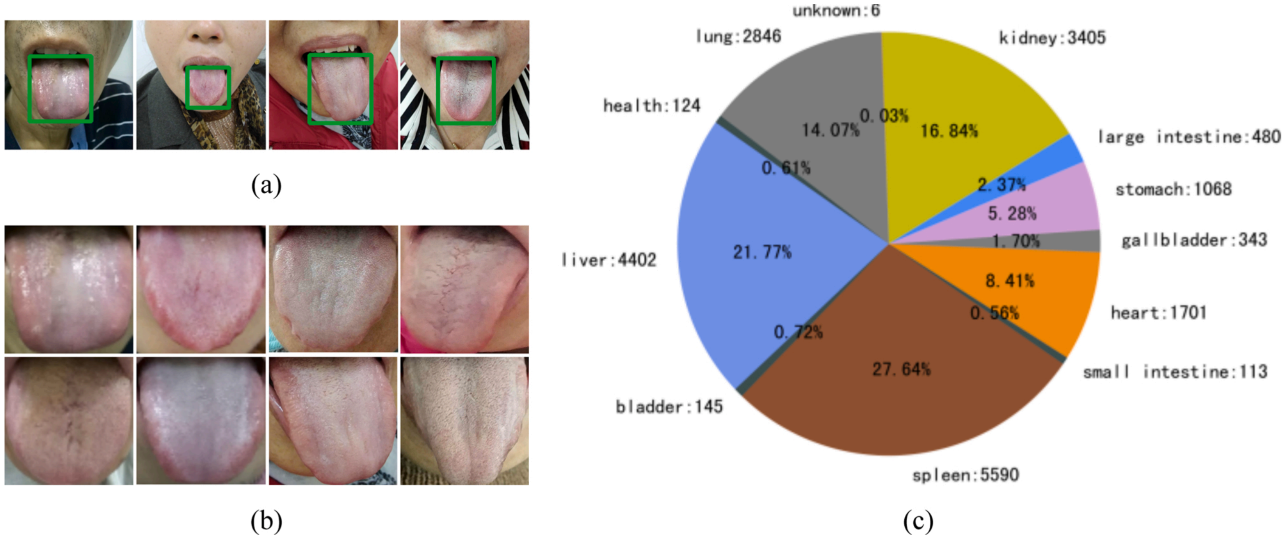

Fig. 3. Examples of tongue images and distribution of disease-locations. (a): original tongue images and their detection frame; (b) pre-processed tongue images; (c) disease location label distribution pie chart.

$T = \{t_1, t_2, \ldots, t_n\}$, after visual modeling of tongue images, feature encoding f_{tg} is obtained, it is input into the classifier and the output is the decision vector $P = [p_1, p_2, \ldots, p_n]$, the dim n of which is equal with the size of set T. The element of decision vector is $p_i \in \{0, 1\}$, when $p_i = 1$, the currently input case belongs to the disease-location t_i, else, the input case is unconcerned to t_i.

Tongue image modeling based on CNNs. Convolutional networks have been successfully applied in medical image modeling [19,11,36,13], CNN model $\mathcal{F}_{\text{cnn}}(\cdot, \mathrm{W}_k)$ extracts the visual features $C(X)$ from input X_{tg}, as:

$$C(X_{\text{tg}}) = \mathcal{F}_{\text{cnn}}(X_{\text{tg}}, \mathrm{W}_k), \tag{1}$$

where W_k refers to the parameters of the CNN model.

Then, features $C(X)$ are input into FC layer with binary classifiers, which is implemented by the sigmoid activation function $\sigma(\cdot)$. Thus, the disease-location results are obtained as follows:

$$\mathcal{P}(X_{\text{tg}}, \theta) = \sigma(C(X_{\text{tg}}), \mathrm{W}_o) = [p(t_1|X_{\text{tg}}, \theta), \ldots, p(t_n|X_{\text{tg}}, \theta)], \tag{2}$$

where W_o denotes the parameters of output FC layer, $\theta = \{\mathrm{W}_k, \mathrm{W}_o\}$ is the entire set of parameters, $\mathcal{P}(X_{\text{tg}}, \theta)$ is the output vector of disease-location recognition, which is consists of the probability $p(t_i|X_{\text{tg}}, \theta)$ of each disease-location.

We have classification labels $G = \{g_1, g_2, \ldots, g_n\}$, $g_i \in \{0, 1\}$, and averaged binary cross-entropy is applied as loss function, as follows:

$$\mathcal{J}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\Big(- g_i \log\big(p(t_i|X_{\text{tg}}, \theta)\big) - (1 - g_i)\log\big(1 - p(t_i|X_{\text{tg}}, \theta)\big)\Big), \tag{3}$$

Stochastic regional pooling of feature maps. In the general channel-wise attention module, the channel signal $\widehat{x}^c$ squeezed from each feature map is obtained by global average pooling from x^c, as follows:

$$\widehat{x}^c = \text{AvgPool}(x^c) = \frac{1}{w \times h}\sum_{i=1}^{w}\sum_{j=1}^{h} x^c[i, j], \tag{4}$$

where $\text{AvgPool}(\cdot)$ is the global average pooling, (w, h) is the shape of the feature map, and c indicates the order number of channel. Although channel-wise attention improves the modeling efficiency of CNNs, the model may still be trapped in the noise of non-tongue areas, and this prevents the model from accurately capturing the detailed pathological features.

Therefore, we design the stochastic regional pooling mechanism. Global average pooling is performed only in some regions of the original feature map, ignoring the noise information. We record the scope of pooling as $([w_b : w_b + l], [h_b : h_b + l])$, the operation of regional pooling is:

$$\widehat{x}^c = \text{AvgPool}_{\{b:b+l\}}(x^c) = \frac{1}{l^2}\sum_{i=w_b+1}^{w_b+l}\sum_{j=h_b+1}^{h_b+l} x^c[i, j], \tag{5}$$

where l is the length of regional pooling area, and $0 \le w_b \le w - l, 0 \le h_b \le h - l$. The setting strategy of hyper-parameter l is elaborated in Sections 4.2 and 5.1. Concentrated signal $\widehat{x}_c$ indicate the whole feature map with local information, all signals formate the inner-imaged map $\widehat{V}$. The expressiveness of inner-imaged maps is further enriched.

Regional pooling is carried out in multiple local regions stochastically to consider the various detailed part. Channel regional signals $\{\widehat{x}_1^c, \ldots, \widehat{x}_r^c\}$ are obtained as follows:

$$\{\widehat{x}_1^c, \ldots, \widehat{x}_r^c\} = \{\text{AvgPool}_{\{b_1:b_1+l_1\}}(x^c), \ldots, \text{AvgPool}_{\{b_r:b_r+l_r\}}(x^c)\}, \tag{6}$$

Inner-imaging channel-wise attention network. The typical channel-wise attention [17] reduces the redundancy of CNNs. However, it ignores the group relationships between channels, which plays a vital role in the overall performance of the convolution channels. To solve this dilemma, as shown by Fig. 5, we project the channel compressed signal $\widehat{x}^c$ onto a matrix $\widehat{V}$ and call it "inner-imaged map", usually its shape is set as $(C/16, 16)$, C is the channel number at each layer. After that, filters $\{\mathrm{W}^{(a_1 \times b_1)}, \ldots, \mathrm{W}^{(a_f \times b_f)}\}$ can be leveraged to scan the inner-imaged maps, thus represent the collaboration of channels as groups. Next, the FC layers W^1 and W^2 are applied to encoding channel relations and output attentional weight for each channel, as follows:

$$s_L = \sigma(\text{ReLU}(\overline{V}, \mathrm{W}^1), \mathrm{W}^2), \tag{7}$$

where,

$$\overline{V} = \mathcal{F}_{\text{flatten}}\Big(\text{Conv}(\widehat{V}, \mathrm{W}^{(a_1 \times b_1)}) \bowtie \cdots \bowtie \text{Conv}(\widehat{V}, \mathrm{W}^{(a_f \times b_f)})\Big), \tag{8}$$

and $\mathcal{F}_{\text{flatten}}(\cdot)$ is the matrix reshape function, filters $\{\mathrm{W}^{(a_1 \times b_1)}, \ldots, \mathrm{W}^{(a_f \times b_f)}\}$ own different sizes, and f is their number of types, operator $\bowtie$ denotes matrices concatenation. Compare with the typical channel-wise attention mechanism [17], our proposed framework first rearranges the squeezed channel signals into the inter-imaging map $\widehat{V}$, then applies $\text{Conv}(\widehat{V}, \mathrm{W}^{(a_1 \times b_1)})$ to model their group relationship on $\widehat{V}$, and combines the group relational modeling results at multiple scales. Finally, we use the two fully connected layers to output the attentional value. The inner-imaged map $\widehat{V}$ in our approach is not the real image feature map, which is applied to organize the channel signals, and then supports the filters to model the group relationship between channels, we call them "group filters". The purpose of integrating multi-size group filters is to improve the completeness of channel relationship encoding.

The channel-wise attentional vector $\mathrm{s}_L = [s^1_L, s^2_L, \dots, s^C_L]$ of layer L, they are multiplied back to each channel to achieve weight adjusting for feature maps, as follows:

$$X_{L+1} = \mathrm{s}_L \circ C_L(X_L), \tag{9}$$

where $\circ$ refers to the element-wise product.

Fig. 4shows the overall architecture of the proposed network. The proposed novel channel-wise attention framework is applied at the end of each block in the CNN backbones. We project regional signals $[\widehat{x}^1_1, \dots, \widehat{x}^1_r, \dots, \widehat{x}^C_1, \dots, \widehat{x}^C_r]$ of all channels onto the inner-imaged map $\widehat{V}$. So, the channel-wise attention module of inner-imaging cannot only model the coordination between channels, but also model the relationship between feature regions of channels, and enhance their linkage.

4.2. Technical details and extensions

Some technical details of the proposed method are described in this section, including: 1. Stochastic tailoring strategy of the regional features; 2. Integration with the current spatial attention mechanism.

Strategy of channel attentional regions selection. The regional pooling area $([w_b : w_b + l], [h_b : h_b + l])$ is stochastically selected within a reasonable range. Firstly, the pooling area cannot exceed the scope of the feature map; secondly, too small pooling area may lead to the excessive loss of visual information. Thus, the regional pooling areas are selected according to the following strategies, as Fig. 6 shows, we put the original $h \times w$ feature map onto the two-dimensional coordinate axis, then set the minimum and maximum pooling sizes to be (l^-, l^-) and (l^+, l^+). When selecting the pooling area, we first determine a center, then select a square area with the side length of $l \in [l^-, l^+]$ based on the center, this pooling center is randomly selected from a rectangular region of shape (r_h, h_w) at the center of the feature map, and:

$$r_h = h - \frac{l^-}{2}, \quad r_w = w - \frac{l^-}{2}. \tag{10}$$

Therefore, the regional pooling areas have at least size for $(l^-)^2$, to ensure the feature map information is not overly lost.

In addition, as shown in Fig. 6(b), for the case that multiple pooled regions are intercepted from the same feature map, we can obtain the attention values $\{s^c_1, \dots, s^c_r\}$ for each sub-region of the feature map, they will be integrated as $s^c = \frac{1}{r}\sum^r_{i=1} s^c_i$, where c denotes the index of the feature map, r is the number of selected regions from each feature map. As Eq. (9), feature map x^c will be regulated by channel attention weight s^c.

Combining with spatial attention. The proposed fully-channel regional attention mechanism has strong generality and can well collaborate with various enhancement mechanisms of CNNs, including visual-spatial attention [33].

The spatial attention mechanism is designed to dynamically adjust the weight at the pixel level on the feature map, as shown in Fig. 7, two modes of spatial attention are implemented. The first is the standard mode that contains only one spatial attention map, and the second is the advanced mode, which combines multiple attention maps; it also costs more parameters than the former. The spatial attention learns the importance of each spatial location on the feature maps, and the attention map is a_L, which indicates using on layer L. We add our fully-channel regional attention module after executing spatial attention, as follows:

$$X_{L+1} = \mathrm{s}_L \circ \mathcal{F}_{\mathrm{sa}}(C_L(X_L), \mathrm{a}_L), \tag{11}$$

where $\mathcal{F}_{\mathrm{sa}}(\cdot)$ is the function of spatial attention, each feature map of the convolution layer is multiplied with the spatial attention map, as feature maps of layer L are $C_L(X_L) = [\mathrm{x}^1_L, \dots, \mathrm{x}^C_L]$, we obtain $\mathcal{F}_{\mathrm{sa}}(C_L(X_L), \mathrm{a}_L) = [\mathrm{a}_L \circ \mathrm{x}^1_L, \dots, \mathrm{a}_L \circ \mathrm{x}^C_L]$.

In the mode with multiple spatial attention maps, each attention map multiplies to a group of feature maps, then we conduct the fully-channel regional attention module, as follows:

$$\begin{aligned} X_{L+1} &= \mathrm{s}_L \circ \mathcal{F}_{\mathrm{sa}}(C_L(X_L), \{\mathrm{a}^1_L, \dots, \mathrm{a}^m_L\}) \\ &= \mathrm{s}_L \circ [\mathrm{a}^1_L \circ \mathrm{x}^1_L, \dots, \mathrm{a}^1_L \circ \mathrm{x}^{\frac{C}{m}}_L, \dots, \mathrm{a}^m_L \circ \mathrm{x}^{C-\frac{C}{m}+1}_L, \dots, \mathrm{a}^m_L \circ \mathrm{x}^C_L] \end{aligned} \tag{12}$$

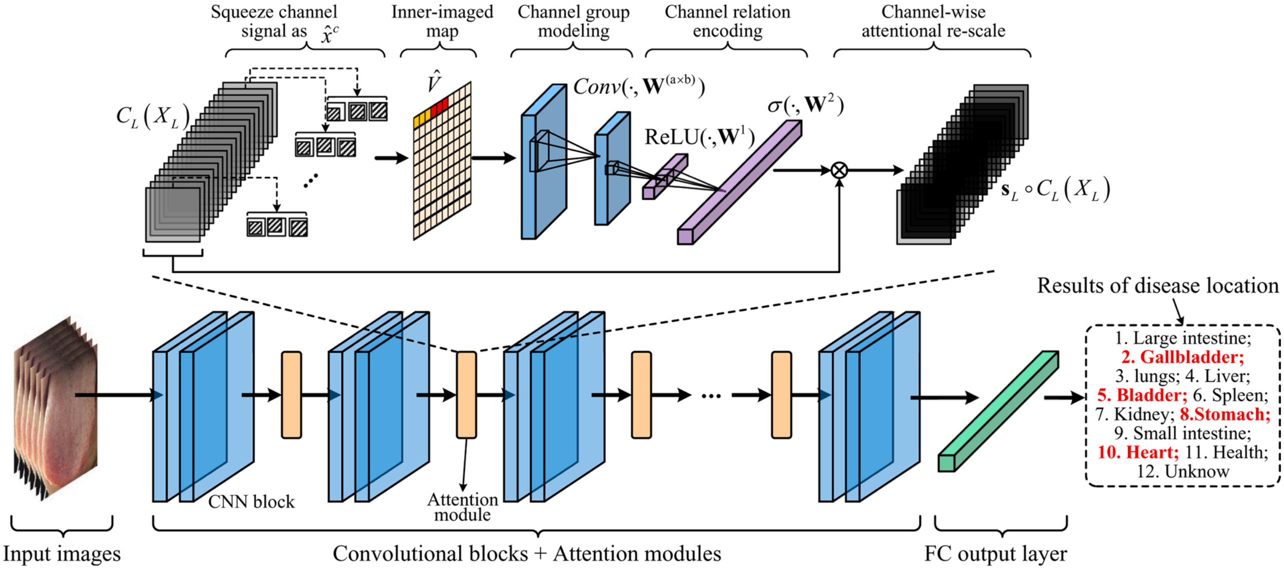
Squeeze channel signal as $\hat{x}^c$
Inner-imaged map
Channel group modeling
Channel relation encoding
Channel-wise attentional re-scale
$C_L(X_L)$
$\hat{V}$
$Conv(\cdot, \mathbf{W}^{(a\times b)})$
$\sigma(\cdot, \mathbf{W}^2)$
$\mathrm{ReLU}(\cdot, \mathbf{W}^1)$
$\mathbf{s}_L \circ C_L(X_L)$
Results of disease location
1. Large intestine;
2. Gallbladder;
3. lungs; 4. Liver;
5. Bladder; 6. Spleen;
7. Kidney; 8.Stomach;
9. Small intestine;
10. Heart; 11. Health;
12. Unknow
CNN block
Attention module
Input images
Convolutional blocks + Attention modules
FC output layer

Fig. 4. The architecture of fully-channel regional attention convolution network. This architecture is significantly different from the ordinary SE-Net, 1. Each feature map is pooled from multiple stochastic regions, get the squeezed signal representing different visual areas; 2. All squeezed signals are rearranged into the Inner-imaging maps, and filters can be used on it to modeling the group relationships of different regions of channels.

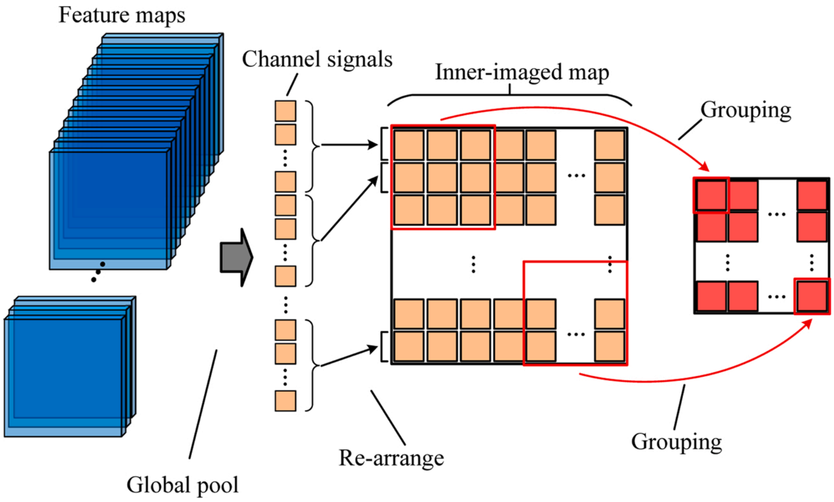

Fig. 5. Detailed framework of inner-imaging mechanism.

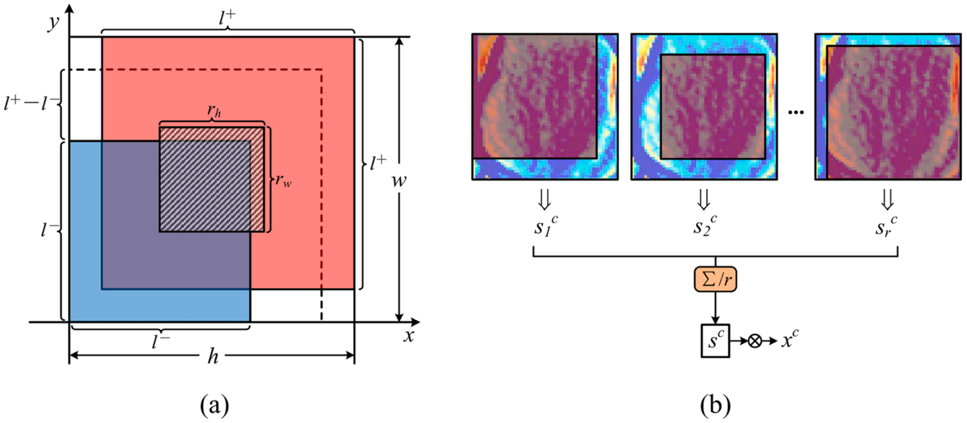

Fig. 6. Technical details of stochastic region pooling. (a) Selection of regional pooling area; (b) Fusion of multi-area channel signals.

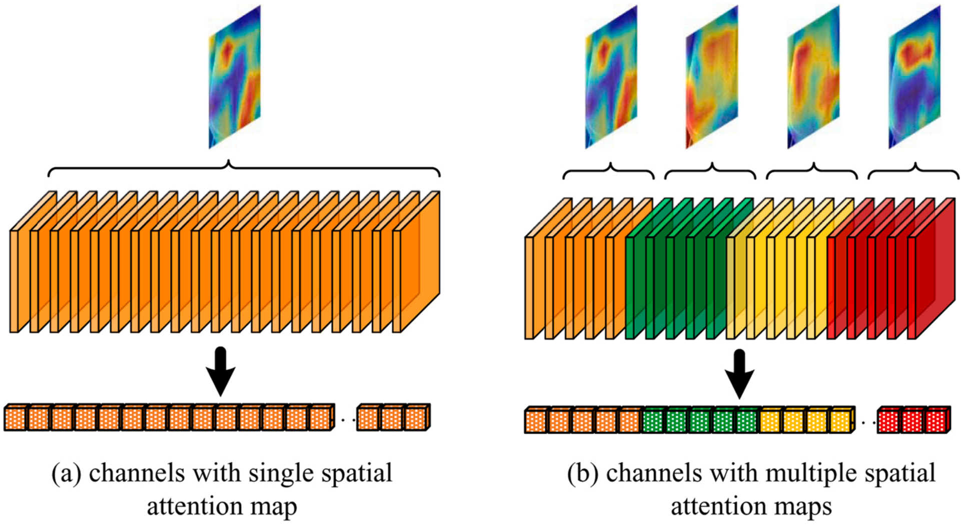
(a) channels with single spatial attention map
(b) channels with multiple spatial attention maps

Fig. 7. Visual spatial attention mechanism. (a) mundane attention mode; (b) multi-attention mode.

where m is the number of spatial attention maps.

5. Experiments

This section describes the experiments of tongue image-based disease-location recognition and verifies the effectiveness of the proposed fully-channel regional attention model in this task. The main contents include 1. Influence of some hyper-parameters; 2. Ablation Studies of all proposed techniques; 3. Comparative experiments with baseline methods; 4. Analysis and discussions.

5.1. Implementation details

Networks. The proposed fully-channel regional attention network can be based on any CNN model, such as VGG-19 [41], ResNet-34, ResNet-50 [15], ResNeXt-50 [49], which are chosen as backbones. All hyper-parameter configurations follow the default settings of the backbones. Besides, the basic channel-wise attention mechanism [17] is also deployed to the backbones as part of the comparison model. In the proposed fully-channel regional attention, the dimension of FC layer is $C/16$, batch normalization [21] is used after the inner-imaging group filters by default. MXNet and GluonCV[1] are used for implementation.

Training. We first pre-train all the experimental models on the ImageNet dataset [9] until convergence, retain their training weights, then we continue to fine-tune the models with the tongue image dataset. The SGD with 0.9 Nesterov momentum is used to train models for 100 epochs with batch-size 32. The learning rate starts at 0.1 and is divided by 10 at epochs 30, 60, and 90. The standard data augmentation (clipping/flipping) is performed during training [14].

Setting of hyper-parameters and model variants. The proposed full-channel regional attention network mainly contains two parts: 1. Inner-imaging channel-wise attention; 2. Stochastic regional pooling, we set prefixes “InI” and “SRP” for them respectively and indicate corresponding parts by prefixing the model name. For inner-imaging channel-wise attention (InI), there are two different modes: basic and composite. The former only uses the (3×3) filter to model the channel group relationships. In contrast, the latter uses three different filters $(1 \times 1), (3 \times 3), (5 \times 5)$ to encode more diverse grouped channel relationships, and we indicate the composite mode with the prefix “InI3”. At any layer, if the size of the group filter exceeds the size of the inner-imaged map, it will be discarded automatically.

For stochastic regional pooling (SRP), the upper limit of the local pooling area size l^+ is set as: $l^+ = \min(h,w)$, where $\min(\cdot)$ denotes picking the smaller value. The lower limit l^- for the pooling size and the number of pooling areas r will be discussed in later sections. Similarly, the prefix “SRP*r*” is used to denote SRP with r pooling areas.

In the case of introducing spatial attention in the experimental models, prefix “SA*m*” is used to indicate them, and m denotes the number of spatial attention maps. In each CNN block, we first conduct spatial attention, then perform other operations. For the case of using multiple mechanisms, we concatenate their prefixes. The naming strategies are included in Table 2.

Evaluation metrics. Expect accuracy [12] of disease-location classification. Precision, recall, and F1-value are also considered, which are calculated by the micro-average strategy and widely applied in multi-label classification tasks [1]. We abbreviate them as Accuracy, Micro-P, Micro-R, and Micro-F1.

5.2. Ablation studies

In this section, abundant ablation studies are carried out, and the effects of the proposed technologies and a few hyper-parameters on the performance of disease-location recognition are analyzed. Fig. 8 shows the curves in the number of regional pooling areas, the lower limit of the pooling area, and the number of group filters in different settings. It can be seen that, with the number of regional pooling areas increasing, the accuracy of disease location recognition is generally promoting. For the lower limit of the size of the pooling area, an excessively low limit will cause some decline in recognition accuracy, because the model may discard too many visual features.

Table 2
Naming strategy of fully-channel regional attention and the other methods.

	Prefix	Example
Full name of the proposed modules		
Inner-Imaging	InI	InI-model
(with multiple grouping filters)	InI3	InI3-model
Stochastic region concentrate	SRP	SRP-model
(with multiple focused region)	SRP*r*	SRP*r*-model
Spital attention	SA	SA-model
(with multiple attention maps)	SA*m*	SA*m*-model
Fully configuration	InI-SRP-SA-model	
Other method name		
Squeeze-and-excitation	SE	SE-model
Convolutional block attention module	CBAM	CBAM-model
Double attention	A^2	A^2-model
Gather-excite	GE	GE-model
Selective kernel	SK	SK-model

When the proportion of the lower limit to the feature map is set to 0.7 or 0.8, the performance reaches the optimal value, so 0.7 is a relatively reasonable choice. When the proportion reaches 0.9, the accuracy somewhat falls back, because it affects the diversity of regional pooling areas. In the subsequent experiments, the ratio of the lower limit of the local pooling area is set to 0.7 by default.

Tables 3 and 4 list the results of ablation studies for the proposed module, including the inner-imaging channel-wise attention (InI), stochastic regional pooling (SRP), the InI module with multiple group filters (InI $\times$ 3) and combination of numerous regional pooling areas (SRP $\times$ 4). It can be seen that with the more and more complete module of the fully-channel regional attention network, the accuracy of disease-location recognition is constantly improving. And we can enhance the performance of recognition by increasing the number of group filters and stochastic pooling regions, with a tiny amount of additional parameters. Compared with the baseline network, the fully-channel regional attention network can improve the disease-location recognition accuracy by 5%, and the proposed method can also improve the Accuracy, Micro-P, Micro-R, and Micro-F1 by more than 3% compared with the classic channel-wise attention network [17], especially on Micro-Recall, the improvements can reach 4.5% at least.

From the results of ablation studies, it can be seen that the performance of model disease location classification can be effectively improved on multiple metrics, whether it uses Inner-imaging channel-wise attention or stochastic regional pooling. Especially when the two are combined, the improvement of model performance is more prominent.

5.3. Compare with the other tongue modeling methods

To our knowledge, there is no similar study to identify the disease-position based on the tongue image. Therefore, in this part, we compare some tongue image modeling methods and apply them to the proposed dataset for the classification of disease-position.

The comparison methods used in this section are traditional machine learning methods and models based on deep learning or deep features. They are: 1. Tongue image modeling methods using color features [45], color and texture features [34], comprehensive features on color, texture, and geometric [52], we use support vector machine (SVM) as the classifier, and bayesian network classifier as in [34]; 2. The CNN to model after capturing the crack [51] and tooth-marked [27] features, the Dual-pipeline CNN model [19], and TongueNet [56]. For TongueNet, we removed the image segmentation interface and replaced it with the sigmoid classifier same as the proposed approach. For TongueNet and our InI-SPA-model, we apply ResNet with 18 layers as the backbone.

[1] https://gluon-cv.mxnet.io.

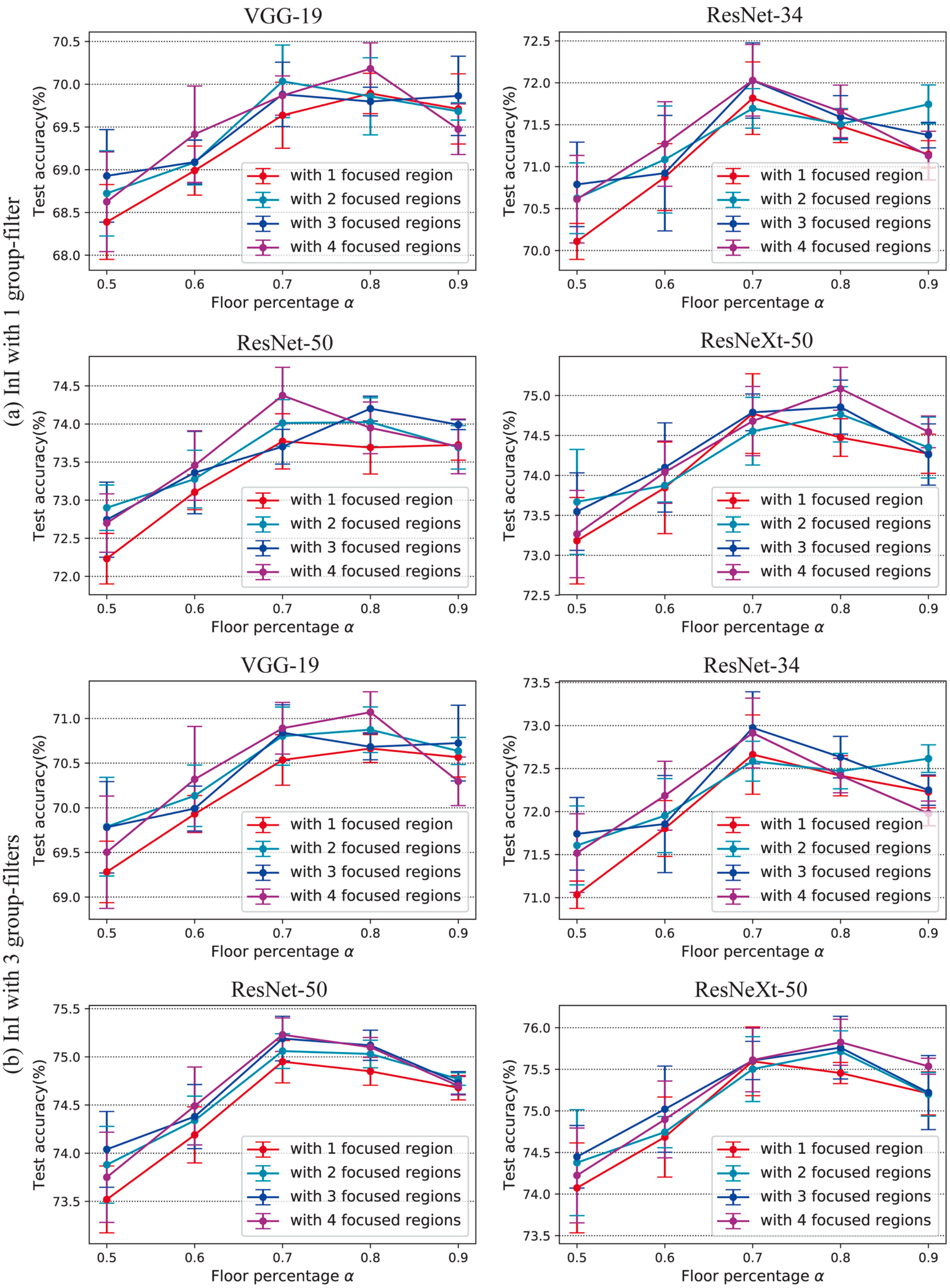

Fig. 8. The accuracy of disease location recognition concerning the number of random pooling regions and the lower limit of the size of the pooling area. (a) Single group filter is used in the inner-imaging mechanism; (b) 3 group filters are used in the inner-imaging mechanism.

Table 3
Test accuracy ((mean ± std) %) of multiple modes of InI-SRP-models on various backbones over 5 runs for disease position recognition. Results that surpass all competing methods are **bold**, and the overall best results are *italics*.

Model	InI	InI × 3	SRP	SRP × 4	Params.	Accuracy
VGG-19 [41]	–	–	–	–	143.7M	65.91 ± 0.41
InI-SRP-VGG-19 (ours)	√	–	√	–	144.8M	69.64 ± 0.39
InI3-SRP-VGG-19 (ours)	√	√	√	–	144.8M	70.54 ± 0.28
InI3-SRP4-VGG-19 (ours)	√	√	√	√	144.8M	**70.89 ± 0.28**
ResNet-34 [14]	–	–	–	–	21.8M	68.25 ± 0.34
SE-ResNet-34 [17]	–	–	–	–	23.6M	69.67 ± 0.42
SRP-SE-ResNet-34 (ours)	–	–	√	–	23.6M	70.94 ± 0.38
InI-SRP-ResNet-34 (ours)	√	–	√	–	23.8M	71.82 ± 0.43
InI3-SRP-ResNet-34 (ours)	√	√	√	–	24.2M	72.66 ± 0.46
InI3-SRP4-ResNet-34 (ours)	√	√	√	√	25M	**72.91 ± 0.41**
ResNet-50 [14]	–	–	–	–	25.6M	70.86 ± 0.42
SE-ResNet-50 [17]	–	–	–	–	28.1M	72.12 ± 0.34
SRP-SE-ResNet-50 (ours)	–	–	√	–	28.1M	72.91 ± 0.29
InI-ResNet-50 (ours)	√	–	–	–	28.2M	72.65 ± 0.33
InI-SRP-ResNet-50 (ours)	√	–	√	–	28.2M	73.77 ± 0.36
InI3-SRP-ResNet-50 (ours)	√	√	√	–	28.6M	74.95 ± 0.22
InI3-SRP4-ResNet-50 (ours)	√	√	√	√	29.5M	*75.23 ± 0.17*

As listed in Table 5, the proposed method has distinct advantages in comparison to the baseline method on various metrics, with an advantage of more than 15%. By analyzing these baseline methods, it can be found that the methods based on the SVM classifier can achieve better results when synthesizing more types of features, however, the performance of these baselines is far inferior to the proposed neural network method, and SVM will consume more time than the Bayesian network classifier. Based on the existing performance of these baseline methods, we did not further try the *k*-Nearest Neighbors (*k*-NN) classifier. Moreover, the baseline methods based on deep features are also not effective, and the reason is they only consider one detailed tongue features and lack comprehensiveness. Besides, none of these comparison methods are designed for the disease-location recognition task studied in this paper, so the results in this part cannot completely confirm the excellent ability of the proposed method. Therefore, we compare more deep CNN models of the same level in the following sections.

5.4. *Compare with the other deep CNN models*

We conduct comparative experiments by using other popular CNN models as baselines: 1. The popular CNN structures VGG [41], ResNet [14], Xception [6], DenseNet [20] and ResNeXt [49], and their channel-wise attention version [17]; 2. The multi-attention CNN structures CBAM-Net [48], A^2-Net [4], GE-Net [16], and SK-Net [26].

From Table 6 and 7, we can see the comparative results of the proposed network with other baseline networks. Compared with the neural networks with the same depth or scale, the proposed fully-channel regional attention network can achieve the optimal performance. It is noteworthy that: although the smaller model InI3-SRP4-ResNet-34 uses fewer layers and parameters, it can achieve better results than ResNet-50 and SE-ResNet-50.

Further, to verify the adaptability of the fully-channel regional attention module to spatial attention mechanism, Table 8 lists the comparative results of our network with the other hybrid attentional networks. Two attention mechanisms are usually considered: channel-wise attention and visual-spatial attention. The proposed fully-channel regional attention model can achieve the best results by combining the spatial attention mechanism. And, with the increase of the number of attention maps, the recognition accuracy of disease-location is rising. The model InI3-SRP4-SA4-ResNet-50 with four attention maps achieves the first-best result, and its test accuracy is higher than the baseline method by at least 2%, on the other metrics, we can get improvement by about 3%. The performance of InI3-SRP4-SA-ResNet-50 with only one

Table 4
Test micro-P, micro-R, and micro-F1 ((mean ± std) %) of multiple modes of InI-SRP-models on various backbones over 5 runs for disease position recognition. Results that surpass all competing methods are **bold**, and the overall best results are **italics**.

Model	InI	InI × 3	SRP	SRP × 4	Micro-P	Micro-R	Micro-F1
VGG-19 [41]	–	–	–	–	56.18 ± 0.47	45.95 ± 0.21	50.55 ± 0.30
InI-SRP-VGG-19 (ours)	√	–	√	–	60.29 ± 0.24	50.28 ± 0.28	54.84 ± 0.24
InI3-SRP-VGG-19 (ours)	√	√	√	–	61.15 ± 0.26	51.19 ± 0.25	55.73 ± 0.25
InI3-SRP4-VGG-19 (ours)	√	√	√	√	**61.74 ± 0.53**	**51.73 ± 0.35**	**56.29 ± 0.41**
ResNet-34 [14]	–	–	–	–	58.85 ± 0.22	49.96 ± 0.31	54.04 ± 0.25
SE-ResNet-34 [17]	–	–	–	–	60.21 ± 0.39	51.30 ± 0.15	55.40 ± 0.25
SRP-SE-ResNet-34 (ours)	–	–	√	–	61.99 ± 0.09	54.09 ± 0.11	57.77 ± 0.10
InI-SRP-ResNet-34 (ours)	√	–	√	–	63.28 ± 0.19	55.27 ± 0.36	59.00 ± 0.29
InI3-SRP-ResNet-34 (ours)	√	√	√	–	64.43 ± 0.26	56.01 ± 0.20	59.93 ± 0.22
InI3-SRP4-ResNet-34 (ours)	√	√	√	√	**64.71 ± 0.26**	**56.21 ± 0.23**	**60.16 ± 0.25**
ResNet-50 [14]	–	–	–	–	61.69 ± 0.29	51.70 ± 0.39	56.26 ± 0.35
SE-ResNet-50 [17]	–	–	–	–	63.51 ± 0.43	55.48 ± 0.37	59.22 ± 0.30
SRP-SE-ResNet-50 (ours)	–	–	√	–	64.49 ± 0.21	56.51 ± 0.16	60.24 ± 0.18
InI-ResNet-50 (ours)	√	–	–	–	64.20 ± 0.41	56.28 ± 0.28	59.98 ± 0.34
InI-SRP-ResNet-50 (ours)	√	–	√	–	65.24 ± 0.34	57.10 ± 0.54	60.90 ± 0.45
InI3-SRP-ResNet-50 (ours)	√	√	√	–	66.51 ± 0.15	58.17 ± 0.43	62.06 ± 0.31
InI3-SRP4-ResNet-50 (ours)	√	√	√	√	*66.83 ± 0.15*	*59.23 ± 0.33*	*62.80 ± 0.23*

Table 5
Test accuracy, micro-P, micro-R, and micro-F1 ((mean ± std) %) of different tongue classification methods over 5-folds validation for disease position recognition. Best competing results are **bold**.

Methods	Accuracy	Micro-P	Micro-R	Micro-F1
Color + SVM [45]	39.05 ± 0.51	30.28 ± 0.66	22.94 ± 0.48	25.33 ± 0.75
Color + Texture + BayesNet [34]	38.62 ± 0.67	31.15 ± 0.81	23.45 ± 0.54	26.81 ± 0.88
Color + Texture + SVM	41.16 ± 0.35	31.69 ± 0.47	25.13 ± 0.63	27.99 ± 0.59
Color + Texture + Geometry + SVM [52]	44.36 ± 0.49	37.43 ± 0.65	30.67 ± 0.79	34.83 ± 0.61
Cracked + CNN + SVM [51]	33.45 ± 0.54	27.98 ± 0.48	23.46 ± 0.78	26.11 ± 0.77
Tooth-marked + CNN + SVM [27]	29.88 ± 0.62	22.17 ± 0.53	17.53 ± 0.78	19.87 ± 0.91
Dual-pipelines CNN [19]	50.25 ± 0.41	44.02 ± 0.25	35.28 ± 0.42	38.26 ± 0.40
TongueNet-18 [56]	48.28 ± 0.58	40.85 ± 0.42	36.61 ± 0.39	39.02 ± 0.45
InI-SRP-ResNet-18 (ours)	66.34 ± 0.19	56.72 ± 0.23	46.03 ± 0.44	50.19 ± 0.21
InI3-SRP4-ResNet-18 (ours)	**68.30 ± 0.24**	**59.01 ± 0.28**	**47.76 ± 0.30**	**52.86 ± 0.36**

Table 6
Test accuracy ((mean ± std) %) of compared methods over 5 runs for disease position recognition. Results that surpass all competing methods are **bold**, and the overall best results are **italics**.

Model	Depth	Params.	Accuracy
VGG-16 [41]	16	138M	63.72 ± 0.45
VGG-19 [41]	19	143.7M	65.91 ± 0.41
InceptionV3 [43]	–	23.8M	66.29 ± 0.25
InI3-SRP4-VGG-19 (ours)	19	144.8M	**70.89 ± 0.28**
ResNet-50 [14]	50	25.6M	70.86 ± 0.42
SE-ResNet-50 [17]	50	28.1M	72.12 ± 0.34
InI3-SRP4-ResNet-34 (ours)	34	25M	72.91 ± 0.41
InI3-SRP4-ResNet-50 (ours)	50	29.5M	**75.23 ± 0.17**
Xception [6]	65	22.9M	68.54 ± 0.44
DenseNet [20]	121	7.9M	71.15 ± 0.45
ResNeXt-50 [49]	50	25M	71.56 ± 0.45
SE-ResNeXt-50 [17]	50	27.5M	72.89 ± 0.25
InI3-SRP4-ResNeXt-50 (ours)	50	28.8M	*75.61 ± 0.38*

attention map is nearly 1% better than that of the other hybrid attention network on each metric. Moreover, even without spatial attention mechanisms, our proposed network can outperform most other hybrid attention networks in disease-location recognition tasks.

Table 7
Test micro-P, micro-R, and micro-F1 ((mean ± std) %) of compared methods over 5 runs for disease position recognition. Results that surpass all competing methods are **bold**, and the overall best results are **italics**.

Model	Micro-P	Micro-R	Micro-F1
VGG-16 [41]	54.31 ± 0.24	43.59 ± 0.32	48.36 ± 0.29
VGG-19 [41]	56.18 ± 0.47	45.95 ± 0.21	50.55 ± 0.30
InceptionV3 [43]	56.78 ± 0.06	46.75 ± 0.27	51.28 ± 0.19
InI3-SRP4-VGG-19 (ours)	**61.74 ± 0.53**	**51.73 ± 0.35**	**56.29 ± 0.41**
ResNet-50 [14]	61.69 ± 0.29	51.70 ± 0.39	56.26 ± 0.35
SE-ResNet-50 [17]	63.51 ± 0.43	55.48 ± 0.37	59.22 ± 0.30
InI3-SRP4-ResNet-34 (ours)	64.71 ± 0.26	56.21 ± 0.23	60.16 ± 0.25
InI3-SRP4-ResNet-50 (ours)	**66.83 ± 0.15**	**59.23 ± 0.33**	**62.80 ± 0.23**
Xception [6]	59.03 ± 0.33	48.58 ± 0.19	53.30 ± 0.25
DenseNet [20]	63.66 ± 0.23	53.89 ± 0.48	58.37 ± 0.38
ResNeXt-50 [49]	64.05 ± 0.19	54.50 ± 0.45	58.89 ± 0.34
SE-ResNeXt-50 [17]	64.63 ± 0.29	56.16 ± 0.25	60.10 ± 0.26
InI3-SRP4-ResNeXt-50 (ours)	*67.10 ± 0.44*	*59.89 ± 0.17*	*63.29 ± 0.28*

5.5. Visualization

In Fig. 9, the visualization method CAM [55] and Grad-CAM [39] give the heat maps corresponding to some tongue image inputs in different models, thus exhibiting the attentional areas of CNNs when performing the disease-location recognition task. In the two kinds of visualization results, the darker red areas and the brighter pixels indicate the areas where the CNN models focus. To evaluate the ability of the model to perform visual modeling, we summarize the following basic principles: 1. The model that pays attention to more non-tongue areas has made wrong modeling; 2. Focusing on the model scattered in many details rather than broad areas is better, more targeted pathological features can be captured.

Based on the above criteria, it can be seen that for tongue images, the proposed fully-channel regional attention network can focus on more comprehensive and scattered pathological feature regions, while other models often focus incorrectly on edge noise areas. After combining the spatial attention mechanism, the proposed fully-channel regional attention network can pay attention to more scattered detailed features. From the output probabilities of disease-location recognition, the proposed models are more accurate, the outputs of the correct category are more significant, and the interference of the wrong category outputs is less.

We illustrate the confusion matrix for each individual disease location classification in Figs. 10 and 11. It can be seen that the proposed InI3-SRP4-ResNet-50 model (as shown in Fig. 11) is better than the baseline SE-ResNet-50 (as shown in Fig. 10) in most categories. However, the unbalanced label allocation seriously affects some categories' recognition. For the “unknown” category, due to its meager number of occurrences, all methods failed: it is mistakenly classified into other classes, so we count its accuracy as 0.

5.6. Discussions

Two attentional mechanisms proposed in this paper, i.e., internal imaging channel attention and random area pool, have their respective roles in improving the model. The former enhances the efficiency of convolution network modeling. At the same time, the latter helps the model capture details better, such as the cleft tongue, red spots on the tip of the tongue, tongue coating, etc. When the two are combined, the model can more effectively model details of tongue image and their relationships.

The study proposes to apply the deep neural network to implement

Table 8
Test accuracy, micro-P, micro-R, and micro-F1 ((mean ± std) %) of methods with multiple attention mechanisms over 5 runs for disease position recognition. Best competing results are **bold**.

Model	Params.	Accuracy	Micro-P	Micro-R	Micro-F1
CBAM-ResNet-50 [48]	28.1M	73.08 ± 0.47	64.24 ± 0.51	56.01 ± 0.26	59.85 ± 0.35
A^2-ResNet-50 [4]	28.4M	74.64 ± 0.51	66.39 ± 0.32	57.96 ± 0.29	61.89 ± 0.29
GE-ResNet-50 [16]	33.7M	75.54 ± 0.41	66.94 ± 0.38	59.66 ± 0.04	63.09 ± 0.19
SK-ResNet-50 [26]	28.5M	72.24 ± 0.36	63.73 ± 0.56	55.85 ± 0.41	59.53 ± 0.48
InI3-SRP4-ResNet-50 (ours)	29.5M	75.23 ± 0.17	66.83 ± 0.15	59.23 ± 0.33	62.80 ± 0.23
InI3-SRP4-SA-ResNet-50 (ours)	29.5M	76.56 ± 0.36	68.02 ± 0.39	60.49 ± 0.44	64.04 ± 0.38
InI3-SRP4-SA2-ResNet-50 (ours)	29.6M	77.18 ± 0.34	68.60 ± 0.42	61.20 ± 0.28	64.69 ± 0.34
InI3-SRP4-SA4-ResNet-50 (ours)	29.8M	**77.79** ± **0.51**	**69.45** ± **0.26**	**61.46** ± **0.23**	**65.21** ± **0.24**

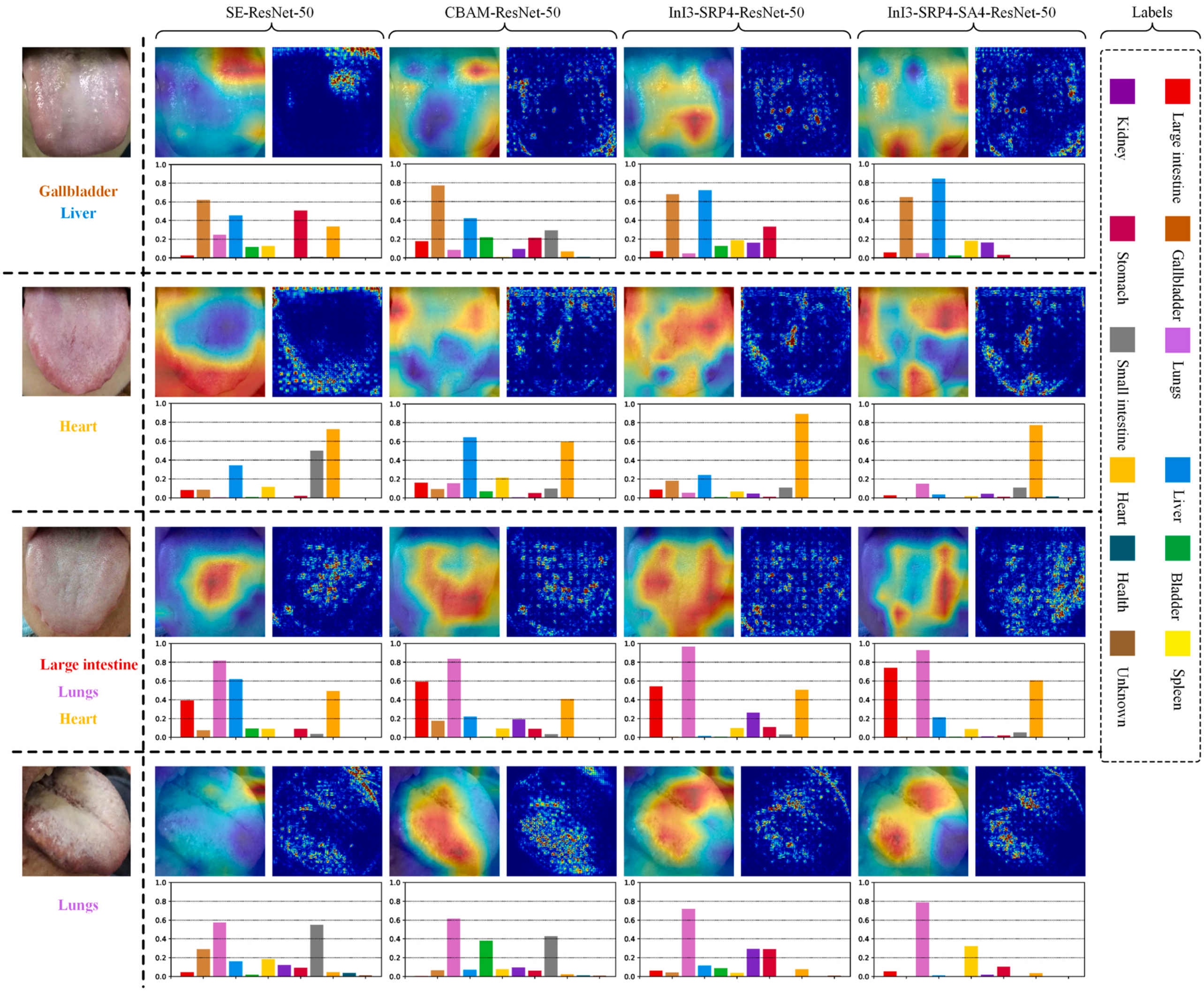

Fig. 9. The CAM [55] and Grad-CAM [39] visualization for different models with the backbone of ResNet-50, on our tongue image dataset. And the output probability of disease location recognition of these models. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

the auxiliary tongue diagnosis. It plays a pivotal assistant role in the real-world outpatient diagnosis. The proposed system has a broad application prospect. Although this study only explores the issue of disease-location recognition based on tongue image, it still has momentous reference significant for the follow-up studies, such as other body signs modeling and auxiliary diagnostic functions. Overall, our proposed method has high practicality and can be transferred to other ancillary diagnostic tasks easily.

From the experimental results, we can see that the fully-channel regional attention network can effectively extract the detailed pathological features on multiple concerned areas, and improve the accuracy of disease location recognition. The proposed design can enhance all kinds of convolutional network structure. It effectively improves the efficiency of CNNs, and the cost is just a tiny amount of additional

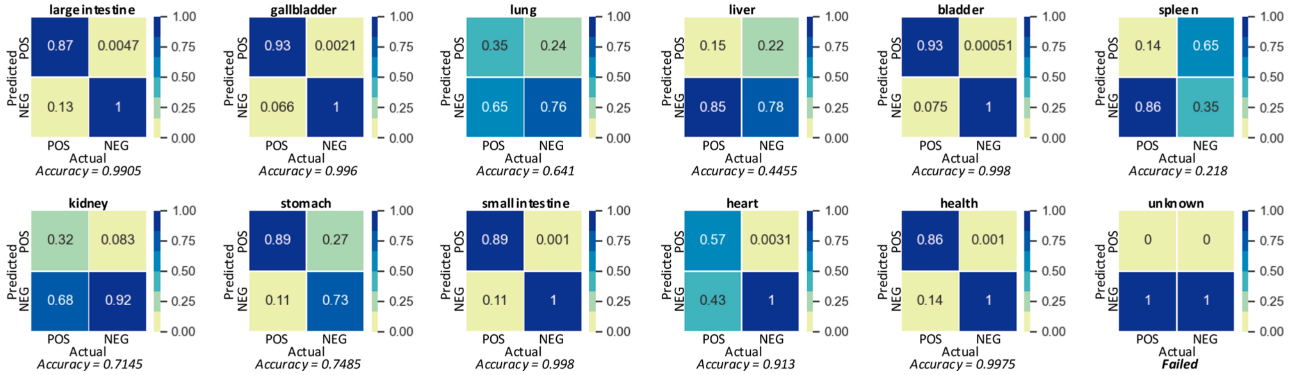

Fig. 10. Confusion matrices for each individual disease location category, with SE-ResNet-50 [17].

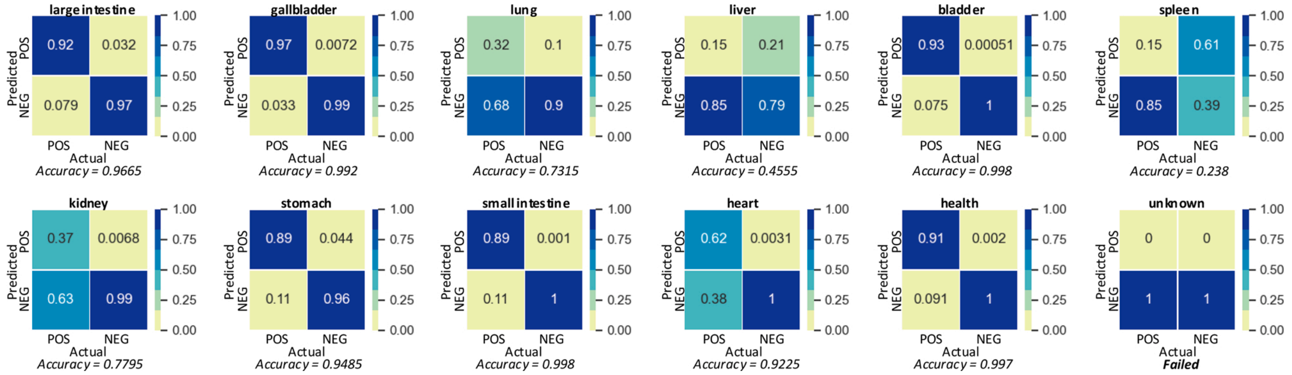

Fig. 11. Confusion matrices for each individual disease location category, with the proposed InI3-SRP4-ResNet-50.

parameters. Moreover, the proposed method can also integrate other enhancement modules of CNNs, like spatial attention mechanism, and exert more powerful effects. After spatial attention regulation, the next two modules: regional feature pooling and inner-imaging channel relationship encoding, can enhance the feature connection of specific visual regions, then increase the expressive ability of CNNs.

Despite the aforementioned progress, it will still be affected when facing unbalanced data. Moreover, the current models cannot successfully recognize categories with few-shot samples. For scenarios with few or even zero samples, further research still needs to be carried out to upgrade the proposed fully-channel regional attention network.

This study has another contribution. To research disease-location recognition, we establish the first large-scale tongue image dataset for disease-location recognition. We effectively alleviated the over-fitting under the limited training data through data argumentation and ImageNet pre-training.

6. Conclusion

In this paper, we focus on the recognition of the patient's disease-location based on tongue image, and our system provides the auxiliary references for outpatient doctors. Using the theory of tongue diagnosis in Chinese medicine, we build a deep learning framework for the patients' tongue image, automatically predict the disease-location. The proposed system can improve the diagnostic efficiency of doctors in the case of a large number of patients. This study establishes the tongue image data set and designs the fully-channel regional attention structure for various convolution network structures. The proposed system does not rely on high-quality tongue images. Mobile devices collect experimental tongue images. Experiments on large-scale tongue image data sets show that the proposed method can improve the efficiency of neural network modeling and obtain excellent performance of multi-focus specific pathological features.

In this pioneering research, for the first time, the deep learning algorithm is applied to solve the problem of automatic disease-location prediction. This study cannot only provide a high-quality assistant function for the outpatient doctors but also provide a reference for follow-up researches. This study is also preliminary research. In the future, we can further improve our system from the following directions: 1. Validate the effectiveness of the proposed method on larger-scale of tongue image datasets; 2. The proposed model is planned to be upgraded to the few-shot and zero-shot paradigms; 3. Introduce more heuristic medical knowledge as a guide to help the model capture pathological features more accurately; 4. Use more diversified features visualization tools to enhance the interpretability of the neural networks in assistant diagnosis tasks.

We will integrate the 3rd and 4th plans. We are going to invite experts to conduct empirical research on symptom feedback of tongue image. We also plan to introduce medical expertise to the proposed system. Our goal is improving the interpretability of the disease location recognition model.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Alali A, Kubat M. Prudent: a pruned and confident stacking approach for multi-label classification. IEEE Trans Knowl Data Eng 2015;27:2480–93.

[2] Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, et al. Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification. Artif Intell Med 2019;97:79–88.

[3] Chagas P, Souza L, Araújo I, Aldeman N, Duarte A, Angelo M, et al. Classification of glomerular hypercellularity using convolutional features and support vector machine. Artif Intell Med 2020;103:101808.

[4] Chen Y, Kalantidis Y, Li J, Yan S, Feng J. Aˆ2-nets: double attention networks. Neural Inf Process Syst 2018:352–61.

[5] Chiu CC. A novel approach based on computerized image analysis for traditional chinese medical diagnosis of the tongue. Comput Methods Programs Biomed 2000;61:77–89.

[6] Chollet F. Xception: deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition 2017:1800–7.

[7] Cuomo S, De Michele P, Piccialli F. 3d data denoising via nonlocal means filter by using parallel gpu strategies. Computational and mathematical methods in medicine 2014. 2014.

[8] Deng F, Lv J, Wang H, Gao J, Zhou Z. Expanding public health in China: an empirical analysis of healthcare inputs and outputs. Public Health 2017;142:73–84.

[9] Deng J, Dong W, Socher R, Li L, Li K, Feifei L. Imagenet: a large-scale hierarchical image database. Proceedings of the IEEE conference on computer vision and pattern recognition 2009:248–55.

[10] Gao S, Qiu JX, Alawad M, Hinkle JD, Schaefferkoetter N, Yoon HJ, et al. Classifying cancer pathology reports with hierarchical self-attention networks. Artif Intell Med 2019;101:101726.

[11] Gessert N, Sentker T, Madesta F, Schmitz R, Kniep H, Baltruschat I, et al. Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. IEEE Trans Biomed Eng 2019.

[12] Godbole S, Sarawagi S. Discriminative methods for multi-labeled classification. Pacific-Asia conference on knowledge discovery and data mining 2004:22–30.

[13] Gupta H, Jin KH, Nguyen HQ, Mccann MT, Unser M. Cnn-based projected gradient descent for consistent ct image reconstruction. IEEE Trans Med Imaging 2018;37:1440–53.

[14] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition 2016:770–8.

[15] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. European conference on computer vision 2016:630–45.

[16] Hu J, Shen L, Albanie S, Sun G, Vedaldi A. Gather-excite: exploiting feature context in convolutional neural networks. Neural Inf Process Syst 2018:9401–11.

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2018.

[18] Hu MC, Lan KC, Fang WC, Huang YC, Ho TJ, Lin CP, et al. Automated tongue diagnosis on the smartphone and its applications. Comput Methods Programs Biomed 2017;174:51–64.

[19] Hu Y, Wen G, Liao H, Wang C, Dai D, Yu Z. Automatic construction of chinese herbal prescriptions from tongue images using cnns and auxiliary latent therapy topics. IEEE Trans Cybern 2019.

[20] Huang G, Liu Z, Der Maaten LV, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2017:2261–9.

[21] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. International conference on machine learning 2015:448–56.

[22] Jiang W, Yang X, Wu W, Liu K, Ahmad A, Sangaiah AK, et al. Medical images fusion by using weighted least squares filter and sparse representation. Comput Electr Eng 2018;67:252–66.

[23] Lee S, Rajan S, Jeon G, Chang JH, Dajani HR, Groza VZ. Oscillometric blood pressure estimation by combining nonparametric bootstrap with gaussian mixture model. Comput Biol Med 2017;85:112–24.

[24] Li W, Abtahi F, Zhu Z, Yin L. Eac-net: deep nets with enhancing and cropping for facial action unit detection. IEEE Trans Pattern Anal Mach Intell 2018:1.

[25] Li X, Shen L, Xie X, Huang S, Xie Z, Hong X, et al. Multi-resolution convolutional networks for chest x-ray radiograph based lung nodule detection. Artif Intell Med 2019:101744.

[26] Li X, Wang W, Hu X, Yang J. Selective kernel networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2019:510–9.

[27] Li X, Zhang Y, Cui Q, Yi X, Zhang Y. Tooth-marked tongue recognition using multiple instance learning and cnn features. IEEE Trans Cybern 2018;49:380–7.

[28] Lu J, Yang Z, Okkelberg KZ, Ghovanloo M. Joint magnetic calibration and localization based on expectation maximization for tongue tracking. IEEE Trans Biomed Eng 2018;65:52–63.

[29] Ma J, Wen G, Wang C, Jiang L. Complexity perception classification method for tongue constitution recognition. Artif Intell Med 2019;96:123–33.

[30] Ma Y, Luo Y, Yang Z. Pcfnet: deep neural network with predefined convolutional filters. Neurocomputing 2020;382:32–9.

[31] Nalepa J, Lorenzo PR, Marcinkiewicz M, Bobek-Billewicz B, Wawrzyniak P, Walczak M, et al. Fully-automated deep learning-powered system for dce-mri analysis of brain tumors. Artif Intell Med 2020;102:101769.

[32] Nardelli P, Jimenez-Carretero D, Bermejo-Pelaez D, Washko GR, Rahaghi FN, Ledesma-Carbayo MJ, et al. Pulmonary artery-vein classification in ct images using deep learning. IEEE Trans Med Imaging 2018;37:2428–40.

[33] Nguyen TV, Zhao Q, Yan S. Attentive systems: a survey. Int J Comput Vis 2018;126:86–110.

[34] Pang B, Zhang D, Li N, Wang K. Computerized tongue diagnosis based on bayesian networks. IEEE Trans Biomed Eng 2004;51:1803–10.

[35] Piantadosi G, Sansone M, Fusco R, Sansone C. Multi-planar 3d breast segmentation in mri via deep convolutional neural networks. Artif Intell Med 2020;103:101781.

[36] Qureshi MNI, Oh J, Lee B. 3d-cnn based discrimination of schizophrenia using resting-state fmri. Artif Intell Med 2019;98:10–7.

[37] Savelli B, Bria A, Molinara M, Marrocco C, Tortorella F. A multi-context cnn ensemble for small lesion detection. Artif Intell Med 2020;103:101749.

[38] Sebkhi N, Sahadat N, Hersek S, Bhavsar A, Siahpoushan S, Ghovanloo M, et al. A deep neural network-based permanent magnet localization for tongue tracking. IEEE Sens J 2019.

[39] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. International conference on computer vision 2017:618–26.

[40] Shen ZY. Basic theory of traditional chinese medicine. Chin J Integr Tradit West Med 1997;17:643.

[41] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International conference on learning representations 2015.

[42] Sun Y, Dai S, Li J, Zhang Y, Li X. Tooth-marked tongue recognition using gradient-weighted class activation maps. Fut Internet 2019;11:45.

[43] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition 2016:2818–26.

[44] Tang C, Dong X, Lian Y, Tang D. Do chinese hospital services constitute an oligopoly? Evidence of the rich-club phenomenon in a patient referral network. Fut Gener Comput Syst 2020;105:492–501.

[45] Wang X, Zhang B, Yang Z, Wang H, Zhang D. Statistical analysis of tongue images for feature extraction and diagnostics. IEEE Trans Image Process 2013;22:5336–47.

[46] Wang X, Zhang D. A new tongue colorchecker design by space representation for precise correction. IEEE J Biomed Health Inform 2013;17:381–91.

[47] Woo J, Murano EZ, Stone M, Prince JL. Reconstruction of high-resolution tongue volumes from mri. IEEE Trans Biomed Eng 2012;59:3511–24.

[48] Woo S, Park J, Lee J, Kweon IS. Cbam: convolutional block attention module. European conference on computer vision 2018:3–19.

[49] Xie S, Girshick RB, Dollar P, Tu Z, He K. Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2017:5987–95.

[50] Xie Y, Xia Y, Zhang J, Song Y, Feng D, Fulham M, et al. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. IEEE Trans Med Imaging 2018;38:991–1004.

[51] Xue Y, Li X, Cui Q, Wang L, Wu P. Cracked tongue recognition based on deep features and multiple-instance svm. Pacific rim conference on multimedia 2018:642–52.

[52] Zhang B, Kumar BV, Zhang D. Detecting diabetes mellitus and nonproliferative diabetic retinopathy using tongue color, texture, and geometry features. IEEE Trans Biomed Eng 2014;61:491–501.

[53] Zhang D, Zhang H, Zhang B. Introduction to tongue image analysis. Singapore: Springer Singapore; 2017. p. 3–18. chapter 1.

[54] Zhang X, Zhou X, Lin M, Sun J. Shufflenet: an extremely efficient convolutional neural network for mobile devices. Proceedings of the IEEE conference on computer vision and pattern recognition 2018:6848–56.

[55] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition 2016:2921–9.

[56] Zhou C, Fan H, Li Z. Tonguenet: accurate localization and segmentation for tongue images using deep neural networks. IEEE Access 2019;7:148779–89.

[57] Zhuo L, Zhang P, Qu P, Peng Y, Zhang J, Li X. A k-plsr-based color correction method for tcm tongue images under different illumination conditions. Neurocomputing 2016;174:815–21.