Attention-Based Multi-Model Ensemble for Automatic Cataract Detection in B-Scan Eye Ultrasound Images

1st Xiaofei Zhang
College of Computer science
Sichuan University
Cheng Du, China
zhangxiaofei@stu.scu.edu.cn

2nd Jiancheng Lv
College of Computer science
Sichuan University
Cheng Du, China
lvjiancheng@scu.edu.cn

3rd Heng Zheng
Sichuan Yanting County People's Hospital
Yan Ting, China
1463556256@qq.com

4th Yongsheng Sang*(corresponding author)
College of Computer science
Sichuan University
Cheng Du, China
sangys@scu.edu.cn

***Abstract*—Accurate detection of early-stage cataract is essential for preventing blindness, but clinical cataract diagnosis requires the professional knowledge of experienced ophthalmologists, which may present difficulties for cataract patients in poverty-stricken areas. Deep learning method has been successful in many image classification tasks, but there are still huge challenges in the field of automatic cataract detection due to two characteristics of cataract and its B-scan eye ultrasound images. First, cataract is a disease that occurs in the lens of the eyeball, but the eyeball occupies only a small part of the eye B-ultrasound image. Second, lens lesions in eye B-ultrasound images are diverse, resulting in small difference and high similarity between positive and negative samples. In this paper, we propose a multi-model ensemble method based on residual attention for cataract classification. The proposed model consists of an object detection network, three pre-trained classification networks: DenseNet-161, ResNet-152 and ResNet-101, and a model ensemble module. Each classification network incorporates a residual attention module. Experimental results on the benchmark B-scan eye ultrasound dataset show that our method can adaptively focus on the discriminative areas of cataract in the eyeball and achieves an accuracy of 97.5%, which is markedly superior to the five baseline methods.**

***Index Terms*—cataract classification, deep learning, attention mechanism, ensemble learning, B-scan eye ultrasound images**

I. INTRODUCTION

Cataract is a common eye disease as well as the leading cause of blindness in the world. Early diagnosis and treatment of cataract is the best way to prevent blindness. The disorder of lens metabolism leads to the degeneration of lens protein, which makes the original transparent lens become opaque or milky white. The light is blocked by the turbid lens and cannot be projected onto the retina, resulting in blurred vision of the patient, which is exactly cataract. The diagnosis of cataract requires extensive expertise and clinical experience of ophthalmologists, which is time consuming and expensive. Accordingly, computer-aided automatic cataract detection is of great significance.

Eye B-ultrasound has the advantages of no damage, no pain, simple, convenient operation, high repeatability and good accuracy. It has been used as a routine examination in fundus examination before cataract surgery [1], [2]. For some cataracts with turbid lens, it is very difficult to clearly observe the fundus, while ultrasound is a kind of examination method that can clearly observe the posterior segment of the eyeball without being affected by lens turbidity [3]. The normal lens in the eye B-ultrasound image is a double-curved light band with thin and smooth periphery, good internal sound transmission and lack of echo [4]. However, cataract has different manifestations in the eye B-ultrasound image. The anterior and posterior capsules and the cortical area of the lens are banded with relatively strong echo. The center of the lens has relatively strong echo light spots or facula with different sizes and shapes [5].

Deep learning method has achieved great success in the field of computer vision. Deep learning method can automatically learn the critical features and integrate the feature learning into the process of building the model, which can reduce the incompleteness caused by the manual design features. CNN (Convolutional Neural Network) [6]–[9] is a representative method. In the field of medical imaging, CNN has also been successfully used for the diagnosis of diseases such as pulmonary nodules [10] and skin cancer [11]. In the classification task of cataract, Gao, Lin and Wong [12] proposed a CRNN (Convolutional-Recursive Neural Network), which feds extracted ROIs (nucleus, anterior cortex and posterior cortex) within detected lens structures and learned local filters into a CNN, and then into RNNs to automatically learn the features for nuclear cataract classification from slit lamp images. Kim, Jun, Kim and Eom [13] introduced tournament-based rank

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.

CNN. It consists of tournament structures and binary CNN models, so as to balance the biased number of images among classes. Zhang, Li, Han, Liu, Yang, Wang et al. [14] used eight layers of CNN to automatically detect and classify retinal fundus images.

The design of the above detection models does not fully consider the characteristics of the input data, and only uses a few layers of CNN. In addition, although the above methods have achieved good results in slit lamp and fundus images respectively, these two images can only see the fundus surface, cannot detect deep lesions, and the corresponding equipment is difficult to operate, which is not conducive to promotion. Therefore, we focus on B-ultrasound images that can detect deep tissue structure and do not require much expertise during operation. We study the characteristics of eye B-ultrasound images, select deep networks with richer feature expressions, better performance and higher accuracy, then carefully design a powerful model for cataract detection.

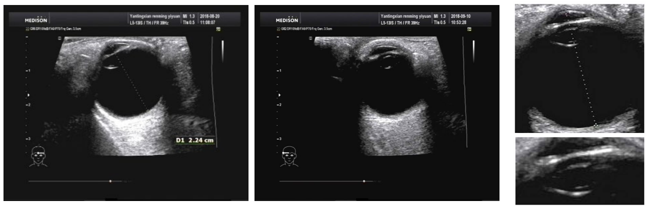


Fig. 1. B-Scan eye ultrasound images. Left two images are original images. Right is the eyeball and the lens in an original image.

For two reasons, it is difficult to classify cataracts in B-ultrasound images. First, the eyeball occupies only a small part of the B-ultrasound image, and the lens occupies only a small part of the eyeball (Fig. 1). There are many methods [15]–[18] to prove that object detection can separate the object of interest from the background and obtain the category and location of this object. In order to avoid most of the irrelevant background interference classification, we use the object detection method to cut out the eyeball. Second, the complexity and diversity of B-ultrasound images lead to high similarity between positive and negative samples, which means that the detection model needs good generalization ability. Ensemble learning can achieve better generalization performance than a single model. In [19], the ensemble method is used to integrate support vector machine and back propagation neural network for final fundus image classification. Furthermore, the abnormal areas of cataract are mainly in the posterior capsule and the center of the lens. It is necessary to focus on the lesions of cataract in the eyeball to further improve overall generalization, so as to better distinguish positive and negative samples. There are many attempts [20], [21] to show that attention mechanism can select more critical information related to the current task from the whole image. In particular, the cost of obtaining clinical samples is high. To solve the problem of small samples of training data, we use transfer learning to apply pre-trained models for cataract classification.

In this paper, we propose an attention-based multi-model ensemble method. First, cut out the eyeball in the eye B-ultrasound image through the object detection network. Then, ensemble three classification networks: DenseNet-161, ResNet-152 and ResNet-101, for the sake of obtaining the final classification result. Besides, each classification network is based on residual attention, which makes the network pay more attention to the lens. In our clinical B-scan eye ultrasound dataset, experimental results show that our ensemble attention model is more efficient than a single classification model. The contributions of this paper can be summarized as follows:

- An ensemble attention model for cataract detection is proposed, which integrates three best-performing general classification models by hard voting;
- According to the characteristics of cataract and eye B-ultrasound images, most of the irrelevant background outside the eyeball is reduced by the object detection network and the weights of the lens are increased by the residual attention module;
- Actual experimental results show that the classification accuracy of our ensemble attention model can arrive to 97.5%. We believe that our experimental research can be used as an important reference for the diagnosis of eye diseases based on the eye B-ultrasound image analysis.

II. RELATED WORK

Object Detection The purpose of object detection is to find all objects of interest in the image, including two subtasks: object location and object classification. At present, the mainstream object detection algorithms can be divided into two categories: (1) Two-Stage method, is also called region-based method. The first stage generates region proposals containing the approximate location information of the objects. The second stage then classifies and refines region proposals. The R-CNN series work [22]–[24] is the representative of this category. (2) One-Stage method, which does not need the region proposal stage but obtains both location and classification results at the same time by processing the image only once, also known as region-free method. YOLO [25]–[27] is the pioneering work of one-stage method. Compared with two-stage method, one-stage method has the advantage in speed. We choose YOLOv3 as our object detection network. YOLOv3 adopts the network structure of Darknet-53 and uses multi-scale feature maps with the advantage of the detection effect of small objects.

Image Classification Since the advent of the ImageNet dataset [28] and AlexNet [6], deep learning method has developed rapidly. Many classic CNNs have been born, which greatly improves the accuracy of image classification. Krizhevsky, Sutskever and Hinton proposed AlexNet, consisting of five convolution layers and three fully connected layers. For the first time, tricks such as relu activation function and dropout are used in CNN. In 2014, Simonyan and Zisserman [7] proposed VGGNet, including 16 layer and 19 layer versions. The network is deepened by repeatedly stacking 3×3 convolution layers and gradually doubling the number of convolution kernels. In the same year, Szegedy, Liu, Jia, Sermanet,

Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.

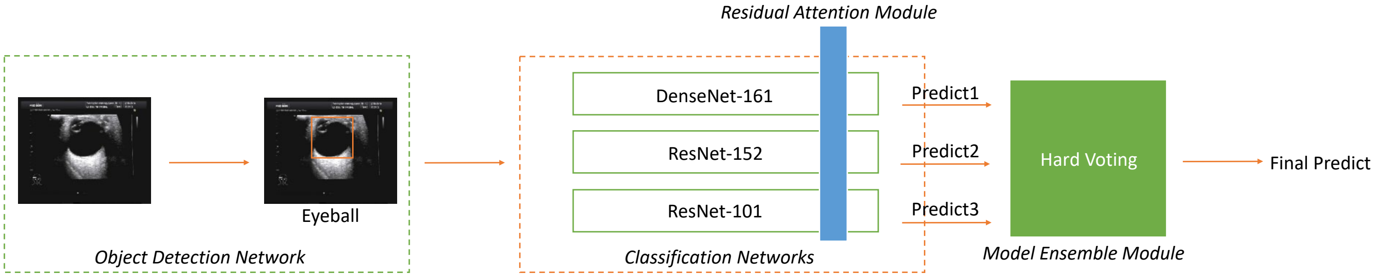


Fig. 2. Ensemble Attention Model

Reed, Anguelov et al. [8] proposed GoogLeNet, with a total of 22 layers, formed by stacking inception blocks. Inceptionv1 consists of four parts: 1×1 convolution, 3×3 convolution, 5×5 convolution and 3×3 maximum pooling. Inceptionv2 uses two 3×3 convolutions instead of a large 5×5 convolution, and also uses batch normalization. In 2015, He, Zhang, Ren and Sun [9] proposed ResNet, a 152 layer network trained by using residual blocks. It is worth mentioning that ResNet solves the degradation problem through skip connection, which provides the possibility for thousands of layers of network training. On the basis of ResNet, Huang, Liu, Van and Weinberger [29] proposed DenseNet using dense connection, where each layer of the network is connected with all the previous layers. Another difference is that DenseNet concats the feature maps of different layers in the channel dimension, so as to achieve feature reuse. Hu, Shen and Sun [30] proposed SeNet to obtain the importance of each channel through learning and perform feature recalibration.

Attention Mechanism Attention mechanism was first used in machine translation tasks [31] and later widely used in natural language processing, computer vision and other fields. Visual attention can quickly scan the global image to locate the target area that needs to be focused on, then pay more attention to the target area to obtain the required details. At present, CNN based visual attention methods are widely used in image classification tasks. Xiao, Xu, Yang, Zhang, Peng and Zhang [32] used two levels of attention for fine-grained classification. First, classifying birds at object-level and part-level respectively. Then, adding the obtained scores to output the final classification result. Zhao, Wu, Feng, Peng and Yan [33] proposed DVAN (Diversified Visual Attention Network) to improve the diversity of visual attention, thereby extracting the most discriminative features. Wang, Jiang, Qian, Yang, Li, Zhang et al. [34] added a new branch called soft mask branch to obtain attention weights beside the trunk branch. The trunk branch and the soft mask branch together form the residual attention module, which makes the network pay more attention to the main features. Hu, Shen and Sun [30] performed squeeze and excitation operations in the channel dimension to automatically learned the importance of each channel. Woo, Park, Lee, So and In [35] put forward CBAM (Convolutional Block Attention Module), which applies attention to both channel and spatial dimensions.

Ensemble Learning Ensemble learning is roughly divided into three categories: bagging [36], boosting [37] and stacking [38]. Bagging uses bootstrap to sample different random subsets of the entire dataset for training each model, and the final prediction result is obtained by voting on multiple models. There are two kinds of voting: hard voting and soft voting. Hard voting is also called majority voting, when the prediction labels of each model are different, the prediction label with the most occurrences is taken as the final classification result. Soft voting calculates the weighted average probability, and selects the label with the largest weighted average probability value as the final prediction result. Boosting iteratively trains each model. At each iteration, the next model is trained by modifying the dataset weights according to the prediction errors in the previous iteration. Stacking trains one model to combine other models. First, training the entire dataset using several different base models. Then, training a new meta model using the outputs of each base model as the input to obtain the final prediction. Hard voting is the most common algorithm in classification tasks. We use it as the ensemble method of our deep learning model.

III. OUR APPROACH

In this section, we will focus on our ensemble attention model, or EAM for short. EAM consists of four parts: object detection network, residual attention module, classification networks and model ensemble module (Fig. 2). The object detection network performs eyeball detection on the original ultrasound image, in order to solve the problem that the eyeball occupies only a small part of the original image and eliminate the strong echo interference in irrelevant backgrounds. The classification network extracts the features of the eyeball and outputs the preliminary prediction result. In each classification network, the attention module is added to make the network follow with interest the lens. The model ensemble module integrates multiply classification networks and outputs the final classification result. We now delve into each of the four parts.

A. *Input and Detector*

Original B-scan eye ultrasound images are saved in DCM format. DCM is the suffix name of DICOM file. DICOM (Digital Imaging and Communications in Medicine) is the standard format for medical image data storage and exchange. We use Python's pydicom package to convert the original

Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.

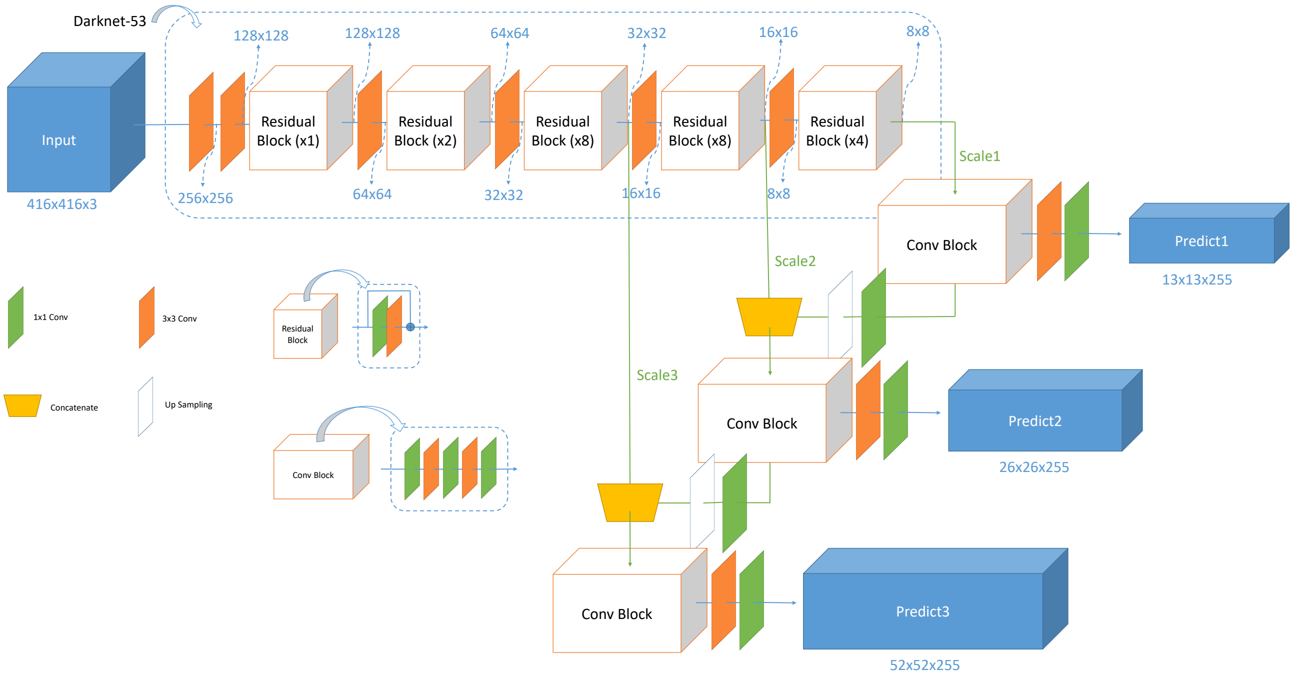


Fig. 3. YOLOv3

DICOM image into PNG format with the size of 720×576. We then fed these PNG images into YOLOv3.

YOLOv3 is one of the representative methods of object detection. During training, the input of YOLOv3 is the image and the original label $(class, x_{min}, y_{min}, x_{max}, y_{max})$ of the image. The label consists of the class of the object to be detected and the coordinates of the upper left and lower right corners of the region of interest containing the object. We only have one class eyeball. YOLOv3 first performs feature extraction on the image through the new backbone network darknet-53. Darknet-53 uses the structure of the residual network, but it is deeper, faster and better. YOLOv3 uses three different scale feature maps to predict, with the sizes of 13×13, 26×26 and 52×52. The small-size feature map detects large-scale objects and the large-size feature map detects small-scale objects, thereby improving the effect of small object detection. A $N \times N$ feature map is divided into $N \times N$ grid cells. If the central coordinate of the ground truth is located in a grid, the probability of this grid containing the object is 1, i.e., the confidence is 1, and the confidence of other grids is 0. Each grid cell predicts 3 anchor boxes. The size of anchor boxes are obtained by K-means clustering. Different scale feature maps correspond to different size anchor boxes. There are 9 anchor boxes in total.

$$\begin{gathered} N \times N \times N_{anchor} \times (N_{position} + N_{score} + N_{class}), \\ where\ N \in \{13, 26, 52\}, N_{anchor} = 3 \end{gathered} \tag{1}$$

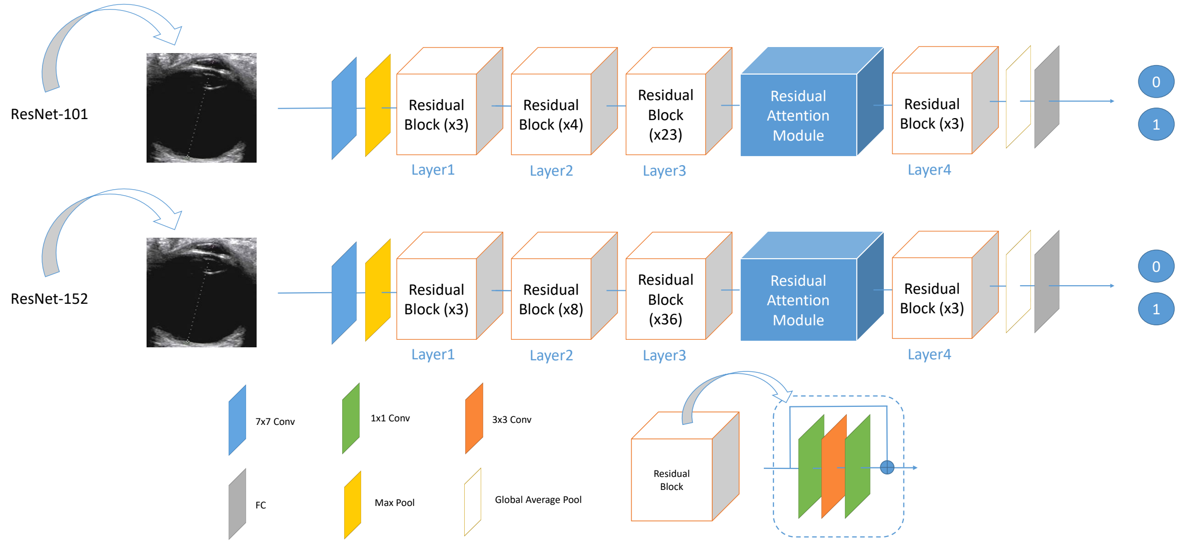
The output of each anchor box is bounding box position, confidence score and conditional class probability. The bounding box position (b_x, b_y, b_w, b_h) is the central coordinate (b_x, b_y), width b_w and height b_h of the bounding box. The confidence score reflects whether the object is contained or not and the accuracy of the bounding box position when the object is contained. The conditional class probability is the probability of the object class. The output of YOLOv3 (Eq. (1)) consists of three parts, so the loss is the weighted sum of these three parts, where the bounding box position uses MSE (Mean Squared Error) loss and the latter two use cross entropy loss. Since eyeball detection has only one class, the detection is less difficult. Eyeball detection by YOLOv3 is not only fast, but also effective. We detect the eyeball through YOLOv3, then use eyeball images as the input of classification networks.

B. Residual Attention

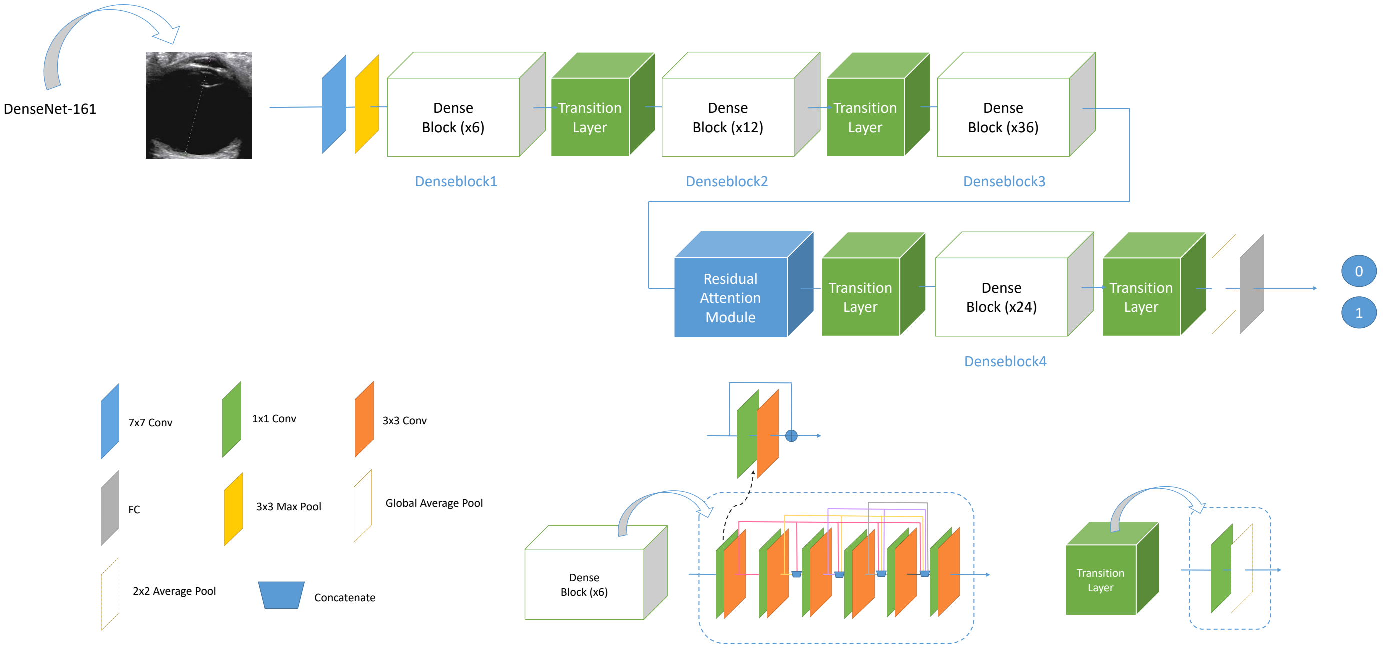
Wang, Jiang, Qian, Yang, Li, Zhang et al. [34] proposed RAN (Residual Attention Network) for image classification. RAN is composed of stacked residual attention modules that enhance features of important areas and suppress meaningless information in other areas. The residual attention module consists of two branches: trunk branch and mask branch.

Trunk branch is a common CNN that is responsible for feature extraction and can be any state of the art. The output of the input x of trunk branch is denoted as $T(x)$. Mask branch generates attention weights by using a bottom-up top-down structure. It is first down sample the input through a series of convolution and pooling operations, then the extracted global high-level feature map with attention is up sampled to generate a soft weighted mask $M(x)$. $M(x)$ and $T(x)$ are the same size. Skip connections are also added to capture information from different scales.

Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.



(a) Attention-Based ResNet



(b) Attention-Based DenseNet

Fig. 4. Classification Networks

The output feature map of trunk branch $T(x)$ and the soft weighted mask $M(x)$ is element-wise multiplied to obtain a weighted attention map. Then, attention residual learning is put forward: after obtaining the weighted attention map, it performs an element-wise addition with $T(x)$, which makes the distinctive features of the original trunk branch feature map more distinctive. The final output of the attention module is as follows:

$$H(x) = (1 + M(x)) \times T(x), where\ M(x) \in [0,1] \tag{2}$$

C. Transferable Models

In order to solve the problem that there are few training samples of our eye B-ultrasound images, we use models pre-trained on the ImageNet dataset which includs more than 1,000 classes of objects and 10 million pictures. If we do not use transfer learning, but directly start training from scratch with our small samples, the effect is poor and it is easy to overfit. And ImageNet pre-trained models can learn more generalized feature expressions, thus we transfer them to the cataract detection task.

Because deep networks can learn more abstract semantic features, the effect is better. ResNet solves the degradation problem, so it can train deeper networks. DenseNet is based on ResNet but can obtain better performance with fewer parameters. Among DenseNet and ResNet models of all layers pre-trained on ImageNet, DenseNet-161, ResNet-152 and ResNet-

Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.

101 have the highest accuracy on the verification set, so we select these three best-performing models as our classification networks. In order to adapt DenseNet-161, ResNet-152 and ResNet-101 to cataract classification, we replace the last 1000d fully connected layer with a 2d fully connected layer. Then, start from the last layer, fine-tune layer by layer. The input size of each network is $224 \times 224 \times 3$.

ResNet-101 and ResNet-152 both consist of five convolutional blocks, a global average pooling layer and a fully connected layer. The first convolutional block just has one convolution layer of 7×7 kernel size. The second convolutional block does a 3×3 maximum pooling, then stacks the residual blocks, and the subsequent convolutional blocks are all simply stacked with the residual blocks. The difference is that the number of the residual blocks stacked by ResNet-101 in the second, third, fourth and fifth convolutional block is 3, 4, 23 and 3, while ResNet-152 is 3, 8, 36 and 3. The residual block is composed of three convolution layers of 1×1, 3×3, 1×1 kernel size, and performs identity mapping between input and output. The first 1×1 convolution is used to reduce the dimension, and the second 1×1 convolution is used to increase the dimension, which greatly saves computing time. DenseNet-161 has 161 learnable layers: one convolution layer of 7×7 kernel size, a 3×3 maximum pooling layer, four denseblocks, three transition layers, a 7×7 global average pooling layer and a 1000d fully connected layer. Four denseblocks are composed of 1×1 convolution and 3×3 convolution with 6, 12, 36 and 24 layers each. In denseblock, each layer has the same feature map size, and takes the output of all previous layers as input. The feature map of each layer is also transmitted to all subsequent layers in a cascade manner. Transition layer connects two adjacent denseblocks for down sampling, consists of a 1×1 convolution and a 2×2 avgrage pooling.

ResNet is connected through element-wise addition, while DenseNet is connected through concat with all previous layers in the channel dimension. The outputs of layer i of ResNet and DenseNet are shown in Eq. (3) and Eq. (4), where $x_0,\ x_1,\ \cdots,\ x_{i-1}$ are the outputs of layer $0,\ 1,\ \cdots,\ i-1$.

$$\text{ResNet: } x_i = H_i(x_{i-1}) + x_{i-1} \tag{3}$$

$$\text{DenseNet: } x_i = H_i([x_0, x_1, \cdots, x_{i-1}]) \tag{4}$$

Because the diagnosis of cataract is based on B-scan ultrasound results of the lens, the ophthalmologist determines if a patient has cataract by identifying whether there is strong echo near the posterior capsule of the lens or within the lens. The classification network itself has a certain degree of attention. However, for cataract classification, additional attention is needed to make the classification network be more attentive to the characteristics of abnormal areas of the lens.

Thus we use classification networks that add the additional attention module. The attention module is residual attention mentioned in the previous section. The specific structure is shown in Fig. 5. We use the output of layer 3 in ResNet as the input of the attention module, then use the output of the attention module as the input of next layer 4 (Fig. 4(a)). The output of denseblock 3 in DenseNet is used as the input of the attention module and the output of the attention module is used as the input of transition 3 (Fig. 4(b)). The input size of the residual attention module is both 14×14.

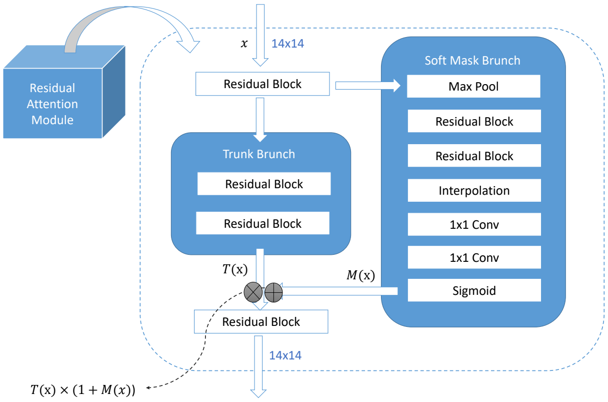


Fig. 5. Residual Attention Module

D. Ensemble Network

Compared with a single model, model ensemble can integrate the advantages of several different models to get better feature representation and higher accuracy. We simply integrate our attention-based DenseNet-161, ResNet-152 and ResNet-101 together by hard voting. Since we use three classification networks, it is only possible to become the final prediction result when one class gets two or three votes. c_0 and c_1 represent two classes with label 0 and 1, i.e., $c_0 = 0,\ c_1 = 1$. Suppose $p_1,\ p_2$ and p_3 are the prediction results of these three networks respectively, where $p_1,\ p_2,\ p_3 \in \{0, 1\}$, then the final classification result P is as follows:

$$a_{ij} = \begin{cases} 1, & p_j = c_i \\ 0, & p_j \neq c_i \end{cases} ,\ i \in \{0, 1\},\ j \in \{1, 2, 3\} \tag{5}$$

$$a_i = \sum_{j=1,2,3} a_{ij} \quad ,\ i \in \{0, 1\} \tag{6}$$

$$P = \begin{cases} 0, & a_0 > a_1 \\ 1, & a_0 < a_1 \end{cases} \tag{7}$$

In order to train the proposed ensemble attention model, we adjust the size of eyeball ultrasound images cut out by YOLOv3 to 224×224, as the input of our pre-trained attention-based DenseNet-161, ResNet-152 and ResNet-101. Data augmentation is performed on input images, including padding according to the longest edge, random horizontal flip, random rotation, color jitter and normalization. We divide 80% of the training data as training set and 20% as validation set, then fine-tune the whole network end-to-end. The loss function is cross-entropy loss. After a lengthy parameter search, we use a mini-batch stochastic gradient descent with batch size of 8, momentum of 0.9, weight decay value of 0.0001. We set learning rate to 0.001. The maximum epoch is 150. Learning rate decay rate is 0.1, decaying occurs when the step is 30%, 60%, or 90% of the total epoch.

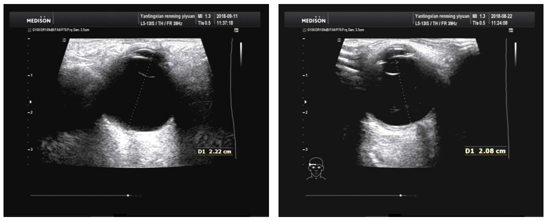
Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.

IV. EXPERIMENTS

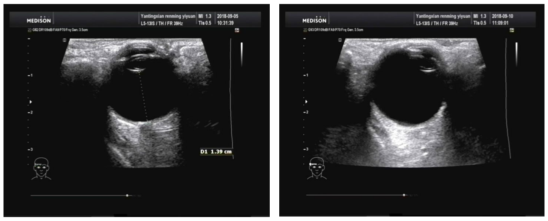
We first introduce our eye B-ultrasound image dataset and list the evaluation metrics. We then perform ablation experiments on EAM to verify the necessity of object detection, residual attention and ensemble learning. We also visualize certain layers of EAM during the training process to prove the rationality of residual attention as well as the accuracy of feature extraction and experimental results. At last we compare EAM with the baseline methods to demonstrate the superiority of our model.

A. Dataset and Metrics

Our B-scan eye ultrasound image dataset consists of normal eyes and cataract eyes. The B-ultrasound images of these two kinds of eyes are shown in Fig. 6. There are 1,894 ultrasound images of normal eyes and 3,615 of cataract eyes. Due to the reality of clinical ophthalmic ultrasound diagnosis in hospitals, images of cataract eyes account for the majority, while there are fewer images of normal eyes. In order to balance the data between positive and negative samples, we use all valid images in normal eyes after data cleaning as positive samples, a total of 1,877, and randomly select 1,896 valid images from cataract eyes as negative samples.



Normal eye



Cataract eye

Fig. 6. Normal and cataract eye images

For the purpose of data augmentation, we process 1,877 normal eye and 1,896 cataract eye B-ultrasound images in two ways, one is to manually cut out the eyeball, the other is to automatic detect the eyeball through YOLOv3. A total of 3,747 normal eyeball images with label 0 and 3,787 cataract eyeball images with label 1 are obtained by these two ways. We randomly select 3,413 of 0 as training set and 334 as test set, 3,441 of 1 as training set and 346 as test set (TABLE I).

TABLE I
TRAINING AND TEST SETS

	Label	Train	Test	Total
Normal eye	*0*	*3,413(1,877+1,870)*	*334*	*3,747*
Cataract eye	*1*	*3,441(1,896+1,891)*	*346*	*3,787*

We evaluate based on five metrics: accuracy, precision, recall, F1-measure and ROC (Receiver Operating Characteristic) curve. F1-measure is a comprehensive evaluation index given by precision and recall. ROC curve is often used to evaluate the quality of binary classification model. AUC is the area under ROC curve. The larger the AUC area, the better the model.

B. Ablation Study

We train EAM without OD (object detection), RA (residual attention) and EL (ensemble learning) respectively. In EAM without EL, there are three cases: only using a single DenseNet-161 (w/o EL-1), only using a single ResNet-152 (w/o EL-2) and only using a single ResNet-101 (w/o EL-3). The results are shown below (TABLE II). It can be seen that full EAM has the highest accuracy, precision, recall and F1-measure, thus proving the usefulness and necessity of OD, RA and EL. We also draw the ROC curve of each model and compare their AUC area (Fig. 7). The ROC curve of full EAM is above the other five curves. It shows that YOLOv3, the attention module, hard voting respectively eliminate redundant irrelevant backgrounds in eye B-ultrasound images, make the classification network pay more attention to the lens lesion area, combine the advantages of multiple classification networks, and finally make EAM achieve the optimal classification effect.

TABLE II
RESULTS OF ABLATION EXPERIMENT

EAM	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
w/o OD	82.7381	80.1075	87.6471	83.7079
w/o RA	97.0588	97.3837	96.8208	97.1014
w/o EL-1	96.7647	97.3684	96.2428	96.8023
w/o EL-2	96.4706	97.0760	95.9538	96.5116
w/o EL-3	96.3235	96.2536	96.5318	96.3925
full	**97.5**	**97.9592**	**97.1098**	**97.5327**

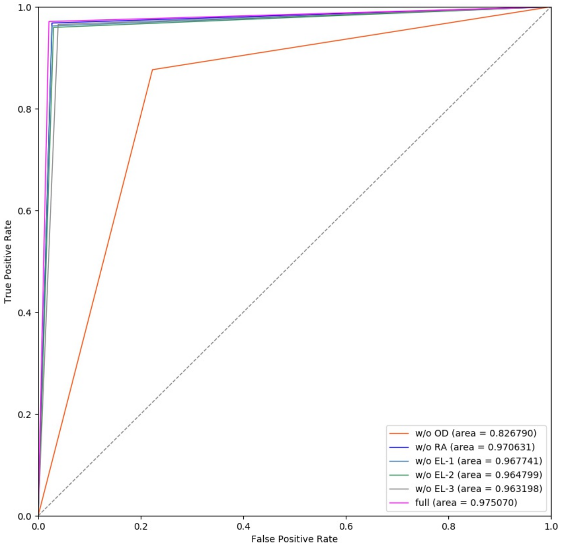
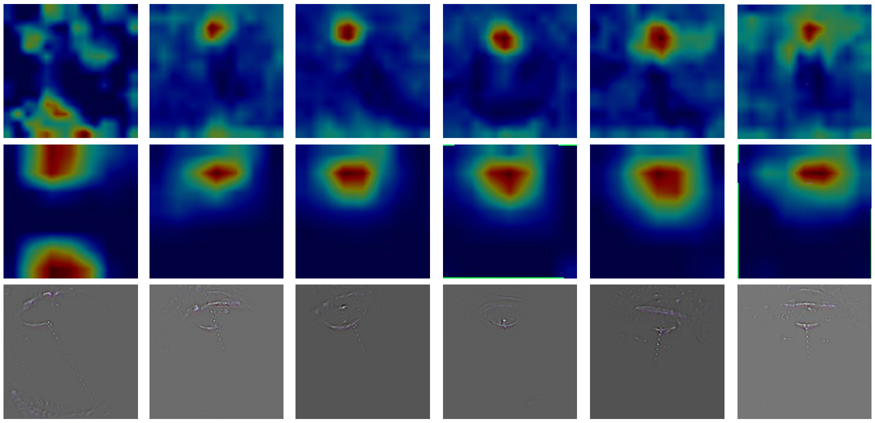
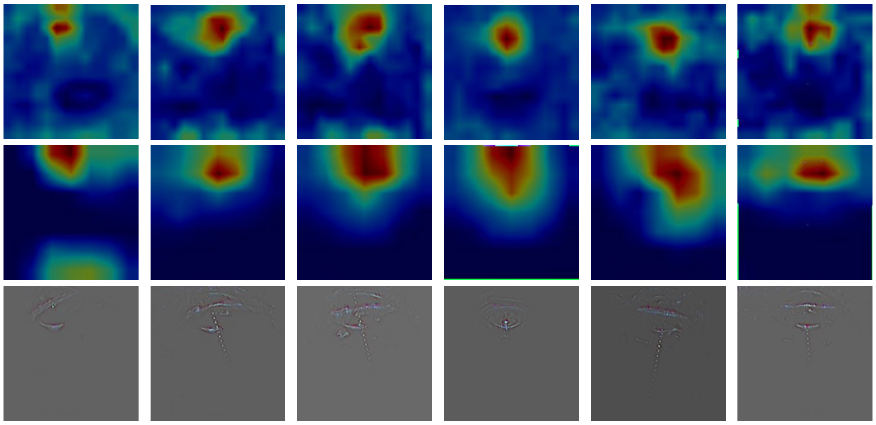


Fig. 7. ROC curve

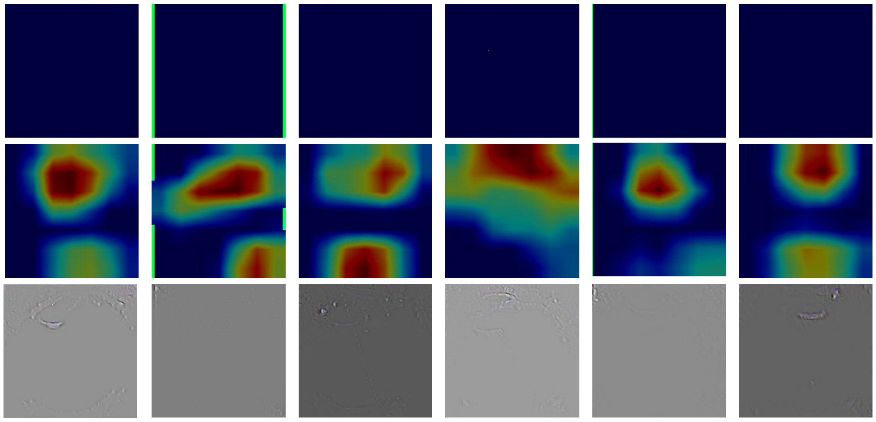
Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.



(a) Attention-Based ResNet-101



(b) Attention-Based ResNet-152



(c) Attention-Based DenseNet-161

Fig. 8. Visualization. First two lines are Grad-CAMs before and after adding the attention module. Last line is guided backpropagation activation maps.

C. *Attention Visualization*

In order to further verify the validity of the attention module, we visualize certain layers during the training process. For ResNet-101 and ResNet-152, we visualize the output of layer 3 before adding the attention module and the output of layer 4 after adding the attention module. For DenseNet-161, we also visualize the feature maps of denseblock 3 and denseblock 4 before and after adding the attention module. We use two visualization methods: Grad-CAM [39], [40] and guided backpropagation. The visualization results are shown in Fig. 8. It can be seen that after adding the attention module, the model highlights the discriminative areas that can be used to diagnose cataract. Take ResNet-101 as an example (Fig. 8(a)). In the first column, before adding the attention module, the lens did not receive significant attention. After adding, the lens attention weight increases. The lens area turns red. In the second, third and fourth columns, after adding the attention module, the anterior and posterior capsules and the central

Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.

area of the lens receive more attention. The red area expands to the entire lens. In the fifth and sixth columns, after adding, the irrelevant background attracts less attention. The noise near the lens is eliminated, the green area turns blue.

D. Comoarison to Baseline Methods

TABLE III shows the results of five baseline methods training from scratch. The accuracy of EAM is 97.5%, which is much higher than that of each baseline model. The precision, recall and F1-measure of EAM are also much higher than those of the other five baseline models. It can be known that the baseline model shows poor generalization ability and is not suitable for specific tasks in the field of medical imaging. Different models need to be designed according to the characteristics of diseases and medical images for different tasks. EAM is specially designed for the cataract detection task of eye B- ultrasound images, which is obviously superior in cataract classification.

TABLE III
RESULTS OF BASELINE MODELS

Baseline	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
VGG-16	50.8824	50.8824	1	67.4464
VGG-19	50.8824	50.8824	1	67.4464
ResNet-18	85.4412	83.4688	89.0173	86.1538
ResNet-34	87.9412	87.9310	88.4393	88.1844
ResNet-50	85.2941	87.5	82.948	85.1632
EAM	**97.5**	**97.9592**	**97.1098**	**97.5327**

V. CONCLUSION

Compared with slit lamp and fundus images, B-scan eye ultrasound images have a larger range and contain a great deal of information that interferes with cataract detection. For professional ophthalmologists, the lens is the most important diagnostic indicator for determining whether a patient has cataract. Based on the above reality, we propose the ensemble attention model. EAM is composed of an object detection network, three classification networks that incorporate residual attention modules, and a model ensemble module. The attention module makes EAM be more attentive to lens lesions. The model ensemble module is equivalent to synthesizing the diagnosis of multiple professional ophthalmologists for more reliable classification results. We evaluate EAM on our B-scan eye ultrasound image dataset. The results show that EAM focuses on the abnormal areas of cataract in the eyeball and achieves better generalization performance than a single classification model. In the future, we will extend the model to three classes for cataract grading: no cataract, mild cataract and severe cataract. In addition, we will try more effective data preprocessing and model ensemble methods to achieve further improvements in accuracy.

ACKNOWLEDGMENT

This work is supported by National Natural Science Fund for Distinguished Young Scholars (Grant No. 61625204) and partially supported by Sichuan Key R & D Plan of China (Grant No. 2020YFG0741).

REFERENCES

[1] Z. Bentaleb-Machkour, E. Jouffroy, M. Rabilloud, J.-D. Grange, and L. Kodjikian, "Comparison of central macular thickness measured by three oct models and study of interoperator variability," *The Scientific World Journal*, vol. 2012, 2012.

[2] Y.-G. Kim, S.-H. Baek, S. W. Moon, H.-K. Lee, and U. S. Kim, "Analysis of spectral domain optical coherence tomography findings in occult macular dystrophy," *Acta ophthalmologica*, vol. 89, no. 1, pp. e52–e56, 2011.

[3] T. Bello and C. Adeoti, "Ultrasonic assessment in pre-operative cataract patients.," *The Nigerian postgraduate medical journal*, vol. 13, no. 4, pp. 326–328, 2006.

[4] L. Xiaomin, W. Ruilan, C. Li, *et al.*, "The diagnosis in 148 cases with cataract by b-mode ultrasonography," *Journal of Clinical Ultrasound in Medicine*, no. 1, p. 15, 2003.

[5] L. Dan and Z. G. Liu, "Senile cataract's ultrasonography diagnose analysis," *Guangxi Medical Journal*, 2001.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[10] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *International Conference on Information Processing in Medical Imaging*, pp. 588–599, Springer, 2015.

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

[12] X. Gao, S. Lin, and T. Y. Wong, "Automatic feature learning to grade nuclear cataracts based on deep learning," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 11, pp. 2693–2701, 2015.

[13] D. Kim, T. J. Jun, D. Kim, and Y. Eom, "Tournament based ranking cnn for the cataract grading," *arXiv preprint arXiv:1807.02657*, 2018.

[14] L. Zhang, J. Li, H. Han, B. Liu, J. Yang, Q. Wang, *et al.*, "Automatic cataract detection and grading using deep convolutional neural network," in *2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC)*, pp. 60–65, IEEE, 2017.

[15] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5325–5334, 2015.

[16] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv preprint arXiv:1509.04874*, 2015.

[17] P. Hu and D. Ramanan, "Finding tiny faces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 951–959, 2017.

[18] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 650–657, IEEE, 2017.

[19] J.-J. Yang, J. Li, R. Shen, Y. Zeng, J. He, J. Bi, Y. Li, Q. Zhang, L. Peng, and Q. Wang, "Exploiting ensemble learning for automatic cataract detection and grading," *Computer methods and programs in biomedicine*, vol. 124, pp. 45–57, 2016.

[20] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, pp. 2204–2212, 2014.

[21] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4438–4446, 2017.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.

[23] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[24] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, pp. 91–99, 2015.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[26] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[27] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[30] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

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

[32] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 842–850, 2015.

[33] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, “Diversified visual attention networks for fine-grained object classification,” *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.

[34] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2017.

[35] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[36] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[37] Y. Freund, R. E. Schapire, *et al.*, “Experiments with a new boosting algorithm,” in *icml*, vol. 96, pp. 148–156, Citeseer, 1996.

[38] D. H. Wolpert, “Stacked generalization,” *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[39] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Authorized licensed use limited to: Guangdong Univ of Tech. Downloaded on March 13,2021 at 04:08:17 UTC from IEEE Xplore. Restrictions apply.