Shellfish Detection Based on Fusion Attention Mechanism in End-to-End Network

Guangyao Li[1,2], Zhenbo Li[1,2(✉)], Chuyue Zhang[1], Yaodong Li[3], and Jun Yue[4]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
{liguangyao, lizb, zcy980416}@cau.edu.cn
[2] National Innovation Center for Digital Fishery, China Agricultural University, Beijing 100083, China
[3] College of Engineering, China Agricultural University, Beijing 100083, China
lyd980324@cau.edu.cn
[4] School of Information and Electrical Engineering, Ludong University, Yantai 264025, China
yuejuncn@126.com

Abstract. Object detection has many difficulties and challenges in the agricultural field, mainly due to the lack of data and the complexity of the agricultural environment. Therefore, we built a shellfish dataset containing 3772 images in 7 categories, all of which were manually labeled and verified. In addition, based on the SSD model framework, we used the lightweight MobileNet-v2 classification network to replace the original VGG16 network, and introduced a residual attention mechanism between the classification network and the prediction convolution layer. This could not only lead to a better capture the local features of the images, but also meet the needs of real-time and mobile use. The experimental results show that the performance of our model on the shellfish dataset is better than the current mainstream target detection models. And the verification results achieved an accuracy of 95.38% and a detection speed of 33 ms per picture, indicating that the validity of the model we proposed.

Keywords: Shellfish detection · Attention mechanism · MobileNet-v2 · SSD

1 Introduction

Object detection is an important part of computer vision, and it is of great significance in smart agriculture and human-computer interaction. The traditional object detection algorithm based on manual extraction features has occupied a dominant position for a long time, but it faces the problem of severe redundancy, lack of pertinence and high time complexity of the region selection strategy window of the sliding window. RBG et al. designed the R-CNN [1] framework in 2014 using a combination of convolutional neural networks (CNN) and regional suggestion networks, and replaced the methods

Student Paper.

© Springer Nature Switzerland AG 2019
Z. Lin et al. (Eds.): PRCV 2019, LNCS 11859, pp. 516–527, 2019.
https://doi.org/10.1007/978-3-030-31726-3_44

liguangyao@cau.edu.cn

based on sliding windows and manual design features in traditional object detection. Since then, object detection has made a huge breakthrough.

There are two main targets for the current object detection. One is the method with Region Proposal as the main step, such as R-CNN, SPP-NET [2], Fast-RCNN [3], Faster-RCNN [4]. The other is based on end-to-end regression, such as Yolo [5], Yolo v3 [6] and other models. The former introduces a deep neural network into the object detection, achieving a qualitative leap in accuracy, but does not meet the real-time requirements in terms of speed. The latter uses the idea of regression to directly return the object frame and the object category at this position in multiple positions of the image, which greatly speeds up the detection, but it is difficult to improve in accuracy. The SSD [7] method proposed by Liu et al. combines the Anchor mechanism in Faster R-CNN with the regression idea in YOLO, and returns the regional features of each position of the whole graph through a multi-scale method. This method also achieves good accuracy and speed in the case of low image resolution. However, as the depth of the model deepens, the complexity of the model increases. For example, the number of layers in ResNet [8] has reached 152 layers. So these methods have high computational power requirements for hardware and are difficult to popularize. To this end, researchers have begun to study lightweight networks for better versatility.

In order to meet the needs of the equipment, some lightweight networks have been proposed, the most typical of which are MobileNet [9], SqueezeNet [10], ShuffleNet [11], Xception [12] and so on. Especially the MobileNet uses different convolution kernels for each input channel and uses a smaller convolution kernel of 1x1 to improve accuracy. These lightweight networks are a good solution to the problem of insufficient memory due to the deep depth of the model. However, these networks are faced with a large number of invalid feature maps during the training process. Then, the attention mechanism based on recurrent neural network has entered people's field of vision, which has enabled researchers to have new ideas in the direction of improving object detection efficiency. Wang et al. [13] proposed an Attention-based residual learning approach to improve the performance of stackable rescue attention modules. In addition, the first-ranked SENet [14] network in the ILSVRC 2017 classification project introduces the residual attention network and repeatedly adjusts the network model by weight function to re-weight the information features to achieve better image classification tasks. The introduction of these attention mechanisms further increases the efficiency of object detection.

At present, mainstream object detection algorithms have achieved certain results in academia and industry, and research in the agricultural field is still rare. As an important part of the agricultural economy, marine shellfish have many kinds and complicated characteristics. The traditional human-oriented operation is difficult to meet the market demand. Therefore, the research on shellfish object detection has great significance for the agricultural economy.

In the existing shellfish target recognition, most of them are based on the manual extraction of contour features, using the principle of graphics to identify. For example, Kun et al. [15] proposed an algorithm based on Gabor filter and two-dimensional principal component analysis to extract the characteristics of snail shellfish, which can be used to classify snail shells. Hiroki et al. [16] used a 1-MHz acoustic focus probe to detect shellfish in sediments and then acquired shellfish targets using two-dimensional acoustic imaging techniques. These traditional methods of detection are often

liguangyao@cau.edu.cn

inefficient and inaccurate, and studies based on CNNs have not yet been discovered. This is because of the complexity of the agricultural environment and there is currently no corresponding shellfish dataset.

In this paper, we use the basic framework of SSD, combined with MobileNet lightweight network, through the introduction of residual attention mechanism, continuous parameter tuning, and finally achieved good results in the detection accuracy and speed of the shellfish dataset. The shellfish detection effect process is shown in Fig. 1. Our main contributions are as follows:

(1) We built a shellfish dataset by means of our own shooting and web crawling, which included 7 types of 3772 photos. And all the pictures are manually labeled and verified.
(2) Based on the SSD framework, we adopt a lightweight MobileNet-v2 network and introduce a residual attention mechanism to propose a shellfish detection and recognition model. And our method achieves 95.38% accuracy and 32 ms/sheet speed, laying the foundation for intelligent research of shellfish.

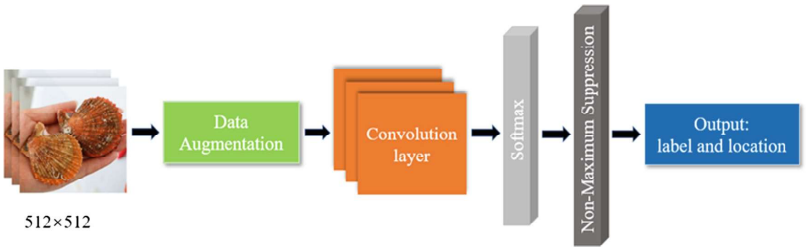

Fig. 1. Shellfish detection framework.

2 Methodology

The CNN can capture the characteristics of the image from the global receptive field to describe the image. However, it is quite difficult to learn a very powerful network. Therefore, the suggestion of attention mechanism can make up for this deficiency very well. In essence, it is to imitate the way humans observe objects, and can gather local features of targets in images to improve detection accuracy.

Therefore, the main idea of our model is based on the SSD model and based on the MobileNet classification network. A residual attention module is introduced between the MobileNet network and the prediction module to strengthen the selected interest area. The overall network framework is shown in Fig. 2. Our approach starts with the relationship between feature channels, explicitly models the interdependence between feature channels, and adopts a new "feature recalibration" strategy to automatically obtain the importance of each feature channel. Then, according to this importance level, the useful features are enhanced and the features that are not useful for the current task are suppressed, so that the feature channel adaptive calibration can be realized. This allows the entire network structure to focus not only on the overall information, but also on the local information. The schematic diagram of introducing the attention residual module is shown in Fig. 3.

liguangyao@cau.edu.cn

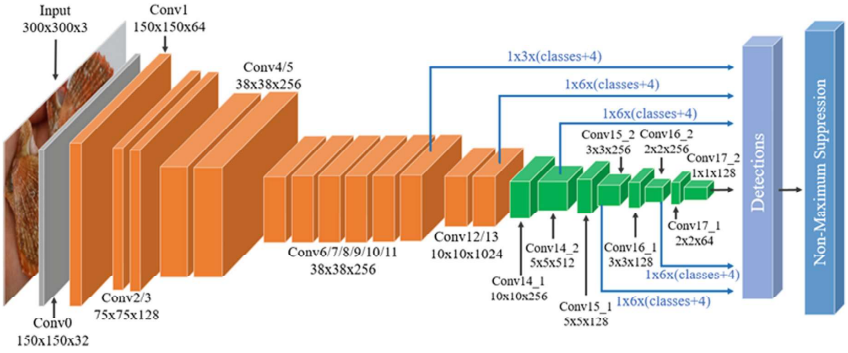

Fig. 2. MobileNet-v2-SSD framework.

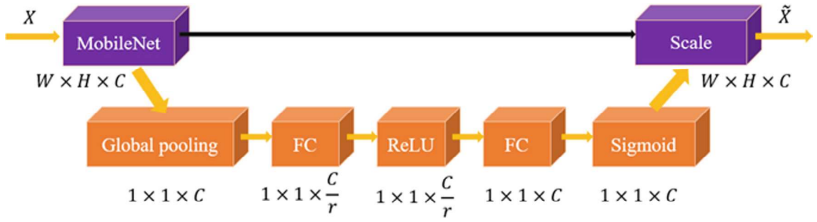

Fig. 3. Framework drawing introducing attention mechanism.

The above figure is interpreted as assuming that the original feature map is $W \times H \times C$, and then globally pooled through $W \times H$ pooled window to get the feature map of $1 \times 1 \times C$. Then use two fully connected layers and one sigmoid layer to output the result. In order to better fit the complex correlation between the channels, when the first fully connected layer is used, the number of neurons C is divided by r for dimensionality reduction, and the second fully connected layer is added to the dimension to return to F. feature. In addition, due to the correlation between the channels, the final output is $1 \times 1 \times C$, where sigmoid is used instead of SoftMax. And our model uses default boxes and loss functions similar to the SSD framework.

Default Boxes. To deal with objectives of different sizes and shapes in images, it is necessary to set prior frames of corresponding scale and proportion according to the network's feature map. The setting of a prior frame is mainly consisting of three parameters: scale, ratio and default box. The specific formula is as follows:

To calculate scale, the size of Default Box in each Feature Map could be calculated as follows:

$$S_k = S_{\min} + \frac{S_{\max} - S_{\min}}{m - 1}(k - 1), \ k \in [1, m] \tag{1}$$

In the formula, the value of $S_{\min}$ is 0.2 and the value of $S_{\max}$ is 0.95, which means the level of the lowest layer is 0.2 and that of the highest layer is 0.5. Different values of

liguangyao@cau.edu.cn

$a_r = 1,\ 2,\ 3,\ 1/2,\ 1/3$ are respectively applied to calculate the height and width of the Default Box. The height and width could be calculated as follows:

$$w_k^a = S_k\sqrt{a_r}, \qquad h_k^a = S_k\sqrt{a_r} \tag{2}$$

In addition, for the case of Ratio = 1, a specific Default Box is added with a scale of $s'k = \sqrt{s_k s_{k+1}}$. As a result, there are 6 different Default Boxes. Set the center of each Default Box to $((i+0.5)/|f_k|, (j+0.5)/|f_k|)$. And $|f_k|$ represents the size of the first feature map $i, j \in [0, |f_k|)$.

Loss Function. The loss function is calculated similar to the loss function in Fast RCNN. The total loss function is the weighted sum of localization loss (LOC) and confidence loss (CONF). The formula is as follows:

$$L(x, c, l, g) = \frac{1}{N}\left(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)\right) \tag{3}$$

In this formula, N is the number of default boxes matching the ground truth box. $x = \{0, 1\}$ is an indicator parameter. And the weighting coefficient α is set to 1 by cross-validation. For CONF, the idea of SoftMax loss is used, defined as follows:

$$L_{conf}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^p \log(\hat{c}_i^p) - \sum_{i \in Neg} \log\left(\hat{c}_i^0\right) \tag{4}$$

Where x_{ij}^p is the identifier of whether the i^{th} default box matches the j^{th} normal data of the category p, and the value is in $\{0, 1\}$. c_i^p is the output of the softmax of the category confidence of the i^{th} default box, $\hat{c}_i^p$ is the confidence level of the background class of the i^{th} default box. *Pos* and *Neg* respectively identify the positive sample set and the negative sample set.

For LOC, Smooth L1 loss mechanism is used, defined as follows:

$$L(x, l, g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^k smooth_{L1}\left(l_i^m - \hat{g}_j^m\right) \tag{5}$$

Because $x_{ij}^k = \{0, 1\}$, the position error could only be calculated for positive samples. It is worth noting that the ground truth g is first encoded to obtain $\hat{g}$, because the predicted value l is also the encoded value. g represent the ground truth box. l indicates the offset of the predicted box from the default box.

3 Experiments

We use dual NVIDIA 1080Ti graphics cards as the computing platform on the Ubuntu 18.04LTS system. This experiment is based on the basic framework of SSD and incorporates lightweight networking and attention mechanisms such as MobileNet. The

liguangyao@cau.edu.cn

comparison models we used include Faster R-CNN, Yolo v3, SSD-MobileNet-v1, SSD-inception-v2, and so on. The comparison experiment employs the Object Detection API framework, which provides a nice API interface and rich varieties of object detection models. 2726 original training sets and 565 test sets are used in the experiment, and they are from the same data source. There is no intersection between the data sets and they are relatively independent.

3.1 Dataset

The shellfish dataset contains 3772 pictures in 7 categories. The pictures taken in the field account for about 75%, and those from the network account for about 25%. The 7 categories are scallop, clam, oyster, oncomelania, mussel, razor clam and conch, and the distribution is shown in Fig. 4. Due to the limitations of machine performance, all images are reset to around 500*500 pixels, and the targets in the images are manually labeled according to the Pascal VOC standard. Then the dataset is randomly divided into training set, verification set and test set according to the ratio of 8:1:1. The results of the division are shown in Table 1.

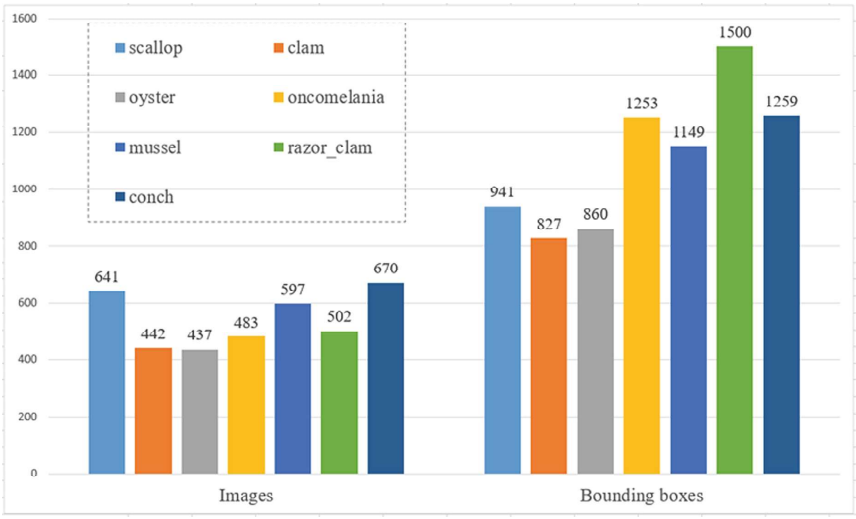

Fig. 4. Dataset type and label distribution.

Table 1. .

Total number	Total spices	Train	Validation	Test
3772	7	2726	481	565

liguangyao@cau.edu.cn

3.2 Data Augmentation

The size and quality of the dataset are critical in deep learning algorithms. On the one hand, a large amount of data could enable the deep learning network to be suitable for complex functions. On the other hand, it could accurately extract the high-level semantic features of data samples. Therefore, in order to make this algorithm more robust to the input in different sizes and shapes, a variety of data augmentation is performed on the images in the dataset, including horizontal flip, random crop and color distortion, randomly sample a patch. Each of the training images randomly samples multiple patches, and the smallest jaccard overlap between objects is: 0.1, 0.3, 0.5, 0.7, and 0.9. This data enhancement operation could increase the number of training samples, and construct more targets of different shapes and sizes, thereby guiding the network to learn more robust features.

3.3 Training Procedure

Before training, the parameters of our model were initialized by a pre-trained model which had been trained on VOC2007 dataset. At the beginning of the training, we randomly weight the image features and use a Gaussian distribution to initialize the weight matrix. The deviation term is a standard normal distribution of 0.0005, and the initial learning rate of the weight is set to 0.0001. In order to improve the training efficiency and make the model converge faster, the Adam gradient descent algorithm is used to train the optimization model. To save training time and speed up the convergence, the experiment uses migration learning to train the deep learning model. First, the parameters of the trained MobileNet classification network are loaded. And then the last classification layer is removed. Finally, the remaining parameter values are assigned to the corresponding parameters in the SSD model.

In training, the values of batch-size, impulse, weight attenuation coefficient, maximum iteration number and initial learning rate are set to 24, 0.9, 0.002, 20000, 0.004, respectively. Then the model is saved once every 5000 iterations, and finally the model with the highest precision is selected. Meanwhile, the Hard-negative mining [17] strategy is used in the training process to enhance the ability of the model to discriminate false positives.

4 Results and Analysis

All of the tests were conducted on our own shellfish dataset. During the evaluation, for each prediction box, the category and the confidence value are first determined according to the category confidence, and the prediction box is filtered out. The prediction boxes with lower confidence threshold is then filtered out. Decode the left prediction frame and get its true positional parameter according to the default box. After decoding, the prediction boxes are arranged in descending order according to its confidence, after which only about 400 of them could be reserved. Finally, the NMS algorithm is performed to filter the prediction boxes with relatively large overlap, and the remaining prediction boxes are the detection result. In order to test the network

liguangyao@cau.edu.cn

performance of the integration of attention mechanisms, the result is compared with those by current mainstream object detection methods. Table 2 shows the results of different model methods for shellfish object detection.

Table 2. Comparison of the test results of each model in the Shellfish test library.

Models	Method	Speed (ms)	mAP
Faster RCNN	Resnet50	89	92.59%
	Inception_v2	58	93.09%
Yolo v3	Darknet-53	38	90.94%
SSD	VGG16	40	94.38%
	Inception-v2	42	94.14%
	Mobilenet-v1	30	92.74%
	Mobilenet-v2	31	93.77%
Our model		**33**	**95.38%**

From Table 2 we can see that our model achieved the accuracy of 95.38% on the shellfish dataset. In addition, the object detection based on the end-to-end series of Yolo v3 and SSD series is considerably faster than the region-based Faster R-CNN method. The detection speed of our model is comparable to the lightweight network MobileNet in SSD. And the detection precision scatter plot is shown in Fig. 5.

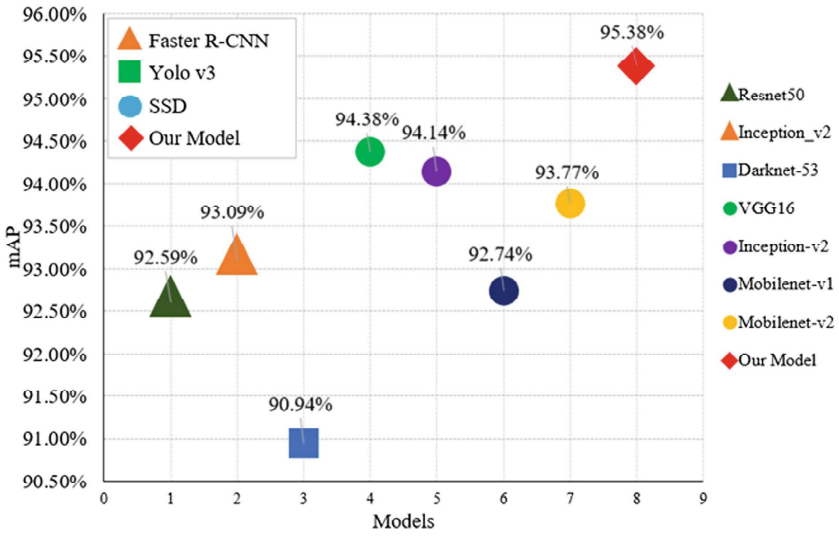

Fig. 5. Shellfish detection accuracy scatter plot.

We can see from Fig. 5 that our model is superior to other models. In order to further better demonstrate the detection performance of the model on each type of shellfish, we give the detection accuracy of each type of shellfish, as shown in Table 3.

liguangyao@cau.edu.cn

Table 3. .

Models	Method	conch	oncomelania	mussel	oyster	razor clam	scallop	clam
Faster RCNN	Resnet50	**98.00**	88.91	87.85	97.67	**98.00**	88.78	88.91
	Inception-v2	96.26	91.36	89.85	**97.87**	95.27	**90.12**	90.91
Yolo v3	Darknet-53	92.00	90.51	91.93	91.83	92.	86.33	91.97
SSD 300*300	VGG16	95.36	92.34	95.07	96.31	96.79	89.33	95.43
	Inception-v2	95.70	93.88	95.49	96.84	96.52	87.07	93.25
	Mobilenet-v1	94.82	93.90	93.85	92.59	96.49	86.83	90.7
	Mobilenet-v2	95.31	94.07	95.22	93.18	96.43	86.41	95.78
Our model		96.75	**95.00**	**96.27**	97.35	96.98	89.42	**95.89**

The detailed experimental results show that the proposed method of the model is better than the original model, indicating that our improved model improves the penetration capability of the network. The average mAP of our method in shellfish object detection is higher than other methods, although the mAP in some types is not as good as the existing models. For example, the detection accuracy of conch and Razor clam is not as good as Faster R-CNN resnet-50, and the detection accuracy of oyster and scallop is not as good as Faster R-CNN Inception-v2. But this does not affect it's overall mAP. We used the time required to test a single image to evaluate the speed in the experiment. The shorter the time, the faster the detection speed. It could be seen from Table 3 that our method is comparable to SSD MobileNet in speed and is superior in accuracy. Object detection based on the SSD framework could ignore the input size of the image, which means in actual applications, the user can use different camera models to take samples. Replacing the original Vgg16 network with MobileNet lightweight network makes the whole model more portable, laying the foundation for mobile applications such as microcontrollers or mobile phones. The introduction of the attention mechanism is more able to express the local features of the target and improve the detection accuracy. A more intuitive comparison is shown in Fig. 6.

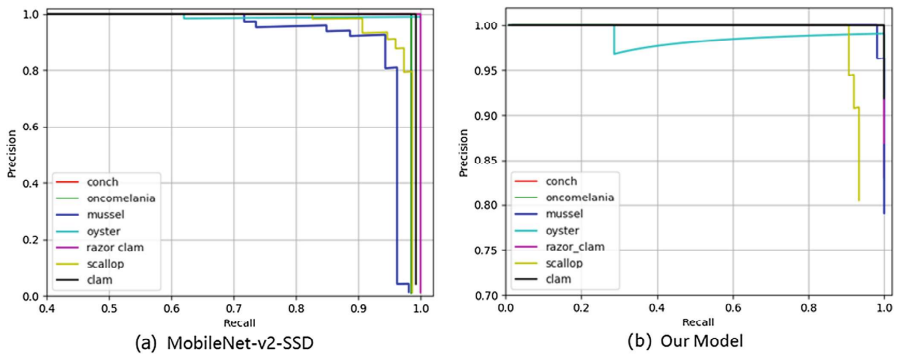

Fig. 6. MobileNet-SSD and our model's precision-recall curve.

liguangyao@cau.edu.cn

It could be seen from Fig. 6 that in the shellfish object detection, the effect of the Precision-Recall graph using only the MobileNet original network is not as good as the effect of adding the residual attention mechanism. To better demonstrate the difference between the better performances in the above two figures, the start values of the recall coordinates in Fig. 6(a) and the precision coordinates in the Fig. 6(b) are respectively set to 0.4 and 0.7. Part of the detection effect is shown in Fig. 7.

Fig. 7. Detection examples on Shellfish test data with our model.

From the pictures of the test we can see that most of the shellfish can still be detected, and can well mark the position of the shellfish target in the picture. However,

liguangyao@cau.edu.cn

when the target object is too dense or too large, the false detection rate will increase. This is because the current work is only a macro classification, and it is not classified from fine-grained. This is a work we will do in the later period, which is to detect and classify the characteristics of the smaller particles of the shellfish data.

5 Conclusions and Future Works

In this paper, we mainly do two aspects of work. On the one hand, we made a dataset containing seven kinds of shellfish, and performed a series of object detection tasks on it. On the other hand, based on the MobileNet-v2-SSD framework, we proposed a shellfish object detection model using the residual attention mechanism. It could better show features in the local detail section, and could simplify the training process of the object detection model and shorten the training time. We replaced the original VGG16 network in the SSD with a lightweight MobileNet-v2 classification network and introduced the attention residual mechanism between MobileNet-v2 and the predictive convolutional layer. This can not only better capture the local features of the image, but also meet the needs of real-time and mobile use. Our experimental results show that compared with the current mainstream object detection model, the model with residual attention mechanism has higher accuracy and speed for target recognition. Meanwhile, our work also has some shortcomings, such as high error rate in intensive targets, large targets and similarly shaped shellfish targets. Our future work will further expand the variety and number of shellfish datasets and study the fine-grained classification of images to enable them to be used for more visualization tasks.

Acknowledgements. This work was supported by National Key Research and Development Program of China under Grant 2018YFD0701003.

References

1. Girshick, R., Donahue, J., Darrell, T., Malik, J.: 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)
2. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 1904–1916 (2015)
3. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
4. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
5. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: 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)
6. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint. arXiv:1804.02767 (2018)

liguangyao@cau.edu.cn

7. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint. arXiv:1704.04861 (2017)
10. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. arXiv preprint. arXiv:1602.07360 (2016)
11. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
12. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
13. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
15. Wang, H., Zhang, K.: Research on shellfish image classification algorithm based on Gabor features. J. New Industrialization **6**, 59–62 (2016). (in Chinese)
16. Suganuma, H., Asada, A., Uehara, Y., Mizuno, K., Yamamuro, M., Okamoto, K.: Detection of shellfish in the sediment by 1-MHz ultrasound: focusing on weak scatter and incident angle. In: 2016 Techno-Ocean (Techno-Ocean), pp. 35–40. IEEE (2016)
17. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–769 (2016)

liguangyao@cau.edu.cn