MobileNetV3 with CBAM for Bamboo Stick Counting

LIANGQUAN JIA1,2, YAWEN WANG2, YING ZANG2, QUANFENG LI2, HUANAN LENG4, ZHANCHUN XIAO3, WEI LONG2 AND LINHUA JIANG2

1Zhejiang University, Hangzhou 310058, China
2School of Information Engineering, Huzhou University, Huzhou 313000, China
3Anji Bata Robot, LTD, Huzhou 313300, China
4 Huzhou municipal bureau of planning and natural resources, Huzhou 313000, China

First Corresponding author: Wei Long (e-mail: lw@zju.edu.cn )
Second Corresponding author: Linhua Jiang (e-mail: 11594@zju.edu.cn )
Co-first authors: Liangquan Jia and Yawen Wang are contributed equally to this work, they are co-first authors of the paper.

This work was supported by Huzhou public welfare applied research Project (2021GZ30, 2021ZD2003, 2016GY11), Scientific Research Fund of Zhejiang Provincial Education Department (Y201941626). This work was partially supported by NSFC (No. 62175037) and Zhejiang Provincial Key R&D Program (No. 2020C02020).

ABSTRACT This study aims to solve the problems of inaccurate weighing of bamboo sticks and inefficient manual counting. To overcome this problem, an improved MobileNetV3 model and a counting algorithm suitable for bamboo sticks—combined with a spatial-temporal attention mechanism—are proposed in this paper. Inspired by the idea of EfficientNet, scaling coefficients are used to scale the MobileNetV3 network structure as a whole in terms of width and height. The optimal model for bamboo-stick recognition is screened initially, and then the algorithm uses the convolutional block attention module (CBAM) attention mechanism to replace the squeeze-and-excitation (SE) attention mechanism in the MobileNetV3 network structure to allow the network to extract features in the two dimensions of channel and space. Since the number of bamboo sticks in a single image is extremely dense—generally around 1000–3000—it is difficult to effectively count them with existing algorithms. The proposed algorithm divides the image into multiple equally sized blocks and then uses the boundary processing algorithm to merge the cut bamboo stick images and count the number of sticks. Experimental results show that the proposed algorithm can effectively perform near real-time detection on a mobile terminal and its accuracy can reach approximately 97%, which is in line with actual production applications.

INDEX TERMS MobileNetV3, Dense bamboo sticks, Attention Mechanism, Border object merger

I. INTRODUCTION

Bamboo has excellent characteristics, such as fast growth, easy regeneration, a low carbon footprint, environmental friendliness, and complete natural degradation. Therefore, its market prospects are very broad. Currently, there are two main ways to sell bamboo sticks, each with a drawback. The first is by weight; however, the bamboo’s wetness affects its weight. The second is by number; however, manual counting is time-consuming and expensive. With the widespread application of image object detection and recognition technology [1] on mobile devices [2, 3], the problems of manual counting can be solved [4]. The main difficulties of bamboo stick recognition algorithms applied to mobile terminals are the following:

(1) Counting bamboo sticks requires high recognition accuracy, and misidentification will cause unnecessary business losses.

(2) Bamboo sticks are tiny objects; therefore, each image contains a host of bamboo sticks. It is a considerable challenge to detect such dense, small objects.

By improving the lightweight MobileNetV3 [5] network, propose an image recognition method based on image object detection for the counting of dense bamboo sticks. The main contributions of the proposed method are the following.

(1) Inspired by the EfficientNet [6] network, we designed network structures with different depths and widths and obtained the network structure most suitable for bamboo stick counting.

(2) An attention mechanism, the convolutional block attention module (CBAM) [7], is used to focus on the channel and spatial dimensions of the feature map to further improve the recognition accuracy.

(3) The number of bamboo sticks in each image is 1000–3000, which is a large number.

Because of the relatively limited computing power of mobile phones, we split an inputted dense bamboo image into multiple small images according to the same ratio of width and height for detection, and then merge the detection results of the small images. Algorithms are used to deal with the problem of secondary detection of boundaries and effectively optimize the network structure. The method advanced in this paper is not only innovative in terms of the algorithm—which improves the recognition accuracy of the original model—but also meets practical demands.

II. RELATED WORK
As a one-time consumable, bamboo sticks have a huge market demand. In response to this situation, Xu [8] proposed an automatic counting machine for bamboo sticks, which counts by detecting the pulse signals of the bamboo sticks passing by the photoelectric sensor on the conveyor belt. However, there is a problem that a lot of equipment needs to be purchased and the real-time counting effect is poor. In order to be able to use the recognition system in mobile devices, it is necessary to further reduce the number of parameter calculations and computational complexity, and improve the lightweight network model. Z.Qin et al. [9] proposed a fast down-sampling MobileNet (FD-MobileNet). By performing 32 down-sampling, the improved model is half of the original MobileNet, which greatly reduces the computational cost and improves the real-time recognition effect. H. Chen et al. [10] adopted a depth multiplier combined with fractional maximum pooling or maximum pooling to improve the depth separable convolution of MobileNet, which reduces the amount of calculation and improves the accuracy. D.Sinha et al. [11] proposed three MobileNet architectures, using Drop activation to replace the Relu activation function, and introducing random erasure regularization technology to replace Dropout, which not only reduces the amount of calculation, but also improves the accuracy. Wang W et al. [12] used dense blocks in the MobileNet network to reduce model parameters and computational complexity by setting a smaller growth rate.

Small object detection has always been a problem in traditional image algorithms and convolutional neural network models. Traditional image detection algorithms have shortcomings such as poor model generalization and complex manual operation for small object detection. With the development of deep learning in recent years, the efficiency of small object detection has been significantly improved [13]. In order to solve the problem of small object detection and recognition, Li et al. [14] proposed to use a generative adversarial network to reduce the difference between small objects and large objects and improve the detection accuracy of small objects. Lin et al. [15] proposed that the feature pyramid network played a key role in small object detection. Singh et al. [16] proposed the idea of image pyramid scale normalized SNIP, which adopted multi-size image input and selected the size of PriorBox to significantly improve the detection results of small objects. Lang et al. [17] proposed a dense convolutional network that horizontally connects pyramid and small object anchor points. Chen et al. [18] proposed a method to improve training based on data provided by feedback during training, which splices multiple small object images into a new image for training.

Some scholars have found that the attention mechanism also plays a key role in small object detection. X. Ran et al. [19] used color information to guide visual attention to the most interesting areas in the image data and proposed an object detection method based on the visual attention mechanism. Z.Jian et al. [20] used the SKBlock structure to expand the field of shallow feature maps and use a self-attention mechanism to improve its ability to recognize small objects. Zhang et al. [21] proposed to construct SEASAM attention module by utilizing the channel and spatial attention in MobileNetV3. Nie et al. [22] proposed a triple attention module. Jiang et al. [23] proposed a small object detection method combining feature fusion and spatial attention. Li et al. [24] proposed an adaptive attention mechanism. Wang et al. [25] proposed a feature stream pyramid network, which designed a feature stream based on the original FPN model and combined it with the CBAM attention module to significantly improve the accuracy of small object detection. Lim et al. [26] proposed a method of combining context and attention mechanisms.

III. PROPOSED METHOD
Aiming at the task of dense bamboo stick detection, this paper proposes a bamboo stick detection method that improves the lightweight network model MobileNetV3. The sparse bamboo stick images are input into the model for training, and the model is optimized according to the experimental results. Finally, a model suitable for bamboo stick detection is obtained. The dense bamboo stick image is divided into small pieces by an algorithm, and the small pieces are input into the model for detection. Finally, the detection results of the small image are merged into the original image size by the algorithm and the number of bamboo sticks is counted. The overall identification process is shown in Figure 1.

This paper selects the lightweight network model MobileNetV3 as the basic network and uses a smaller scale scaling factor to change the channel and depth coefficients of the MobileNetV3 network. Through experimental comparison, it is finally determined that the empirical value of the Bneck channel scaling factor is 0.7, and the depth scaling factor is 1, not only reducing the parameter operation, but also improving its accuracy. In addition, the SE module of the original MobileNetV3 network is replaced with a CBAM module, and
the attention maps formed in the channel and space dimensions are superimposed, which further improves the detection accuracy. The improved model is shown in Figure 2.

**A. THE NETWORK STRUCTURE**

Past experimental experience has shown that for the improvement of convolutional neural networks, the focus is on the three dimensions of network depth, network width, and resolution. EfficientNet uses a series of fixed scale scaling factors to unify the dimensions of the network. Increasing the depth of the network can obtain richer features, but if the network depth is too deep, it will face the problem of gradient disappearance. Increasing the width of the network enables higher fine-grained features, but it will increase the computational overhead and storage cost. This paper adopts the idea of a series of fixed scale scaling coefficients to adjust the width and depth of the model. The channel value is obtained by multiplying the channel dimension by the multiple factors, and the integer multiple of 8 with the smallest difference from the channel value is taken as the final channel value. The depth dimension is multiplied by a multiple and rounded up to get the depth value. After many comparison experiments, the channel values of the Bneck structure are finally determined to be 8, 16, 32, 56, 80, and 112, which are 0.7 times that of the original Bneck structure, and the depth remains unchanged. The improved Bneck composite zoom model is shown in Figure 3.

**B. CBAM ATTENTION MECHANISM**

The Bneck structure of the MobileNetV3 network uses the Squeeze and extraction modules to solve the loss problem caused by the different importance of different channels of the Feature map during the convolution pooling process, but the SE module pays attention to the channel dimension of the feature map and ignores the spatial dimension of the target information. The CBAM module will form an attention map in the two dimensions of channel and space in turn, and perform element-wise multiplication operations on the attention map and the input feature map of the respective dimensions. The extracted target features are more comprehensive and have higher accuracy. Compared with the SE module, the channel attention mechanism module of CBAM [27] has more parallel Global Max Pooling layers, and different pooling operations can extract richer high-level features [28, 29]. The Bneck structure in the bamboo stick recognition model performs dimension upgrade and deep convolution on the number of input channels and inputs the feature F obtained by the deep convolution into the channel attention module of CBAM to obtain the channel feature. The feature F obtained by the deep convolution and the channel feature are multiplied bit by bit to obtain the feature F', which is input to the spatial attention module to obtain the spatial attention map.
feature. The channel feature $F'$ and the spatial feature are multiplied bit by bit to obtain the final feature $F''$, which is subjected to linear point-by-point convolution. Figure 4 shows the structure diagram after adding CBAM to the Bneck structure of MobileNetV3.

**C. SEMANTIC GUIDANCE STRUCTURE**

![Diagram](image)

**FIGURE 5. Dense bamboo stick identification process.**

The counting process of dense bamboo sticks is shown in Figure 5. Before inputting the trained model, the dense bamboo stick data with labels needs to be processed. According to the width and height of the image to be sliced, the required slice size for partition into $N$ equal parts is calculated.

Formula (1) is used to calculate the slice width $W_1$ and height $H_1$, where $W$ is the width and height of the original image, $N$ is the number of divisions. The height is the same.

$$W_1 = \frac{W}{N}$$

Sliding the slice from left to right and from top to bottom to cut the image. During the slicing process, the 9 kinds of situations that appear in the label box and the slice (the label box is in the upper left corner, lower left corner, upper right corner, lower right corner, right top, right bottom, right-left, right and center of the slice) are analyzed. Figure 6 shows the position of the label box at the upper left corner of the slice.

**FIGURE 6. The positional relationship between the label box and the slice.**

Determine whether the upper left corner and lower right corner of the object point are in the slice area. For example, if the upper left corner and lower right corner of the object point are in the slice area at the same time, it is judged as the center position, the other positions are the same. Then recalculate the $x_{min}$, $y_{min}$, $x_{max}$ and $y_{max}$ of the label box on the cut image. If the object point is at the center of the new image after cutting, the new image $x_{min}$ is the original image $x_{min}$ minus the slice $x_{min}$, the new image $x_{max}$ is the new image $x_{max}$ plus the original Figure $x_{max}$ minus $x_{min}$, the other points are the same. Discard the box with too small cutting. When performing the last slice of each row and column, the width and height of the original image are used to subtract the slice size, the overflow part is adjusted. The non-maximum suppression algorithm is used to deal with the repeated calculation of the object in the repeatedly cut region.

The cut image and its corresponding label file are input into the model for detection. For the problem that a bamboo stick is repeatedly detected at the edge of the cut image, analyze the $x_{min}$, $y_{min}$, $x_{max}$ and $y_{max}$ sizes of the edge detection frame of the cut image. Compared with the size of the normal image detection frame, it is concluded that the edge detection frame with $(x_{max}-x_{min}) < 10$ or $(y_{max}-y_{min}) < 10$ needs to be eliminated in the detection process.

We set the number of rows and columns according to the number of divisions, paste the detection results from left to right and top to bottom according to the image labels, process the overlapping images through an algorithm, and finally get the original size image, and calculate all bounding boxes. Count and display to the center of the screen, the counting result of the dense bamboo stick image is shown in Figure 7.

**IV. TRAINING**

**A. THE NETWORK STRUCTURE**

The experimental data in this paper comes from the bamboo sticks provided by farmers who sell bamboo in Anji. We randomly grab less than 100 bamboo sticks and bundle them together. The heights of 5cm, 10cm, 15cm, and 20cm were taken from the front and left and right inclination to take pictures, screen clear and effective experimental data, and then
use labeling software to label them. The sparse bamboo stick samples collected were 600. Basic image processing techniques such as cropping and flipping process some data to obtain 700 samples to form a sparse bamboo stick sample set, and then randomly divide the image data into a training set and a validation set according to a ratio of 5:1. The sparse bamboo stick sample set is shown in Figure 8 shown.

In the model testing process, the dense bamboo stick images taken from the front of each bundle of bamboo sticks with a diameter of about 15cm were used. Since the diameter of each bamboo stick varies, the range of each bundle of bamboo sticks is about 1000-3000. Figure 9 shows sample data of dense bamboo stick images.

B. HYPERPARAMETER SETTING

The method in this paper is based on the MobileNetV3 network in the open-source TensorFlow Object Detection API Model 2.0, the experimental environment is Linux, the TITAN RTX GPU is used for training. When training based on the improved MobileNetV3 bamboo stick recognition system, some parameters are set as follows: Batch_size is 64, the learning rate of 2000 steps before training is 0.001, the learning rate afterward is 0.004, the IOU threshold is 0.5.

C. EXPERIMENTS AND RESULTS

The MobileNetV3 model is directly trained with dense bamboo stick images. After multiple down-samplings, the images are aggregated into a point on a deep feature map, which makes the model indistinguishable, and the experimental effect is extremely poor. The average precision (AP) only reaches 1%, and training cannot be performed. Therefore, we propose the use of sparse and clear bamboo stick images to train the original MobileNetV3 model, which obtains a suitable bamboo-stick detection model. Then, a cutting of dense bamboo stick images experiment was conducted to select the optimal number of cut copies. Finally, the optimal bamboo-stick detection model was obtained by optimizing the original model.

The rotated sparse bamboo-stick images were unified into a single image with a size of 320 × 320 pixels, and the data volume was further expanded through random clipping and horizontal flipping during model training to improve the model’s generalization ability. Herein, we compare the recall rate, average accuracy, and detection quantity of each dense bamboo-stick image by cutting the dense bamboo skewers into 25 and 36 equal parts after boundary optimization. The experimental results show that the recall rate of 25 slices is improved by 10.24% compared with 25 slices, and the average accuracy increased by 8.59%. In addition, considering the subsequent merge boundary and error between the detection and actual numbers, it is determined that the dense bamboo-stick image is divided into 36 equal parts, as shown in in Table 1.
This approach combines the idea of calculating a series of scaling coefficients with EfficientNet, and concludes that a series of scaling coefficients of the Bneck structure in the MobileNetV3 model are as follows: width coefficient 0.6 combined with depth coefficient 1, width coefficient 0.7 combined with depth coefficient 1, width coefficient 0.7 combined with depth coefficient 0.7. The idea of dichotomy is then adopted to select coefficients from the two aspects of width and depth, and the recall rate and average precision of the worst and best dense bamboo stick images collected are compared. The experimental results are shown in Figure 10, in which c is the width of the model and d its depth. When the Bneck structure scaling factor selects a width factor of 0.7 combined with a depth factor of 1, in the case of poor image quality, the recall rate is increased by 3.75%, and the accuracy is increased by 3.67%. In the case of good image quality and good original model recognition accuracy (b), the recall rate and average accuracy are improved by approximately 0.5%; this is enough to prove the effectiveness of the Bneck scale scaling factor in bamboo-stick detection and recognition proposed in this paper.

In this paper, the CBAM spatial-temporal attention module is integrated into the Bneck structure of the MobileNetV3 model, and feature extraction is carried out in channel and space dimensions. By comparing the original MobileNetV3 model (designated model 1), MobileNetV3 with a scaling factor of 0.7 combined with a depth factor of 1 (2), MobileNetV3 combined with CBAM model (3), MobileNetV2 model (4), and ResNet101 model (5). This paper proposes comparisons of the MobileNetV3 model combining the CBAM and a width scaling factor of 0.7 combined with a depth factor of 1 (6). The experimental results shown in Figure 11 show that the MobileNetV3 model combined with CBAM and scaling coefficients has a higher recall rate and average precision than other models, and the number of bamboo sticks detected is closer to the actual number.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
in this paper reaches 1.329/s.

Comparing the results of counting dense bamboo-stick images obtained using models 1–6, this paper proposes that the Bneck structure scale coefficients are selected as a width coefficient of 0.7 and depth coefficient of 1. Combined with the CBAM and compared with the original MobileNetV3 model, the results are improved by 4.52% under the conditions of shooting light interference and arbitrary binding shape. The difference between the detected and actual quantities is within 1%. When the dense pictures were clear and the bundles relatively neat, test results reached 97.38% accuracy. The experimental results are shown in Table 3, c represents the width of the model, and d represents the depth of the model. The results of segmentation detection in this paper are shown in Figure 12. The results of dense bamboo sticks are shown in Figure 13.

| TABLE 3. Test results of different models. |
|------------------------------------------|
| Model         | Recall/% | Ap%  | Actual Number | Detected Number |
|---------------|----------|------|---------------|-----------------|
| MobileNetV3   | 92.23    | 91.71| 1840          | 1782            |
| C*0.6+d*1+SE  | 87.04    | 86.14| 1840          | 1767            |
| C*0.7+d*0.7+SE| 92.61    | 91.94| 1840          | 1823            |
| C*0.7+d*1+SE  | 95.98    | 95.38| 1840          | 1912            |
| C*1+D*1+CBAM  | 97.11    | 96.29| 1840          | 1974            |
| MobileNetV2   | 86.66    | 86.52| 1840          | 1639            |
| ResNet101     | 88.81    | 88.39| 1840          | 1717            |
| C*0.7+d*1+CBAM| 94.05    | 93.46| 1840          | 1879            |
| MobileNetV3   | 93.06    | 92.79| 1760          | 1725            |
| C*0.6+d*1+SE  | 85.28    | 84.74| 1760          | 1624            |
| C*0.7+d*0.7+SE| 92.89    | 92.61| 1760          | 1718            |
| C*0.7+d*1+SE  | 94.36    | 94.04| 1760          | 1765            |
| C*1+D*1+CBAM  | 94.87    | 94.55| 1760          | 1783            |
| MobileNetV2   | 86.97    | 86.87| 1760          | 1556            |
| ResNet101     | 91.99    | 91.74| 1760          | 1663            |
| C*0.7+d*1+CBAM| 95.26    | 94.94| 1760          | 1760            |
| MobileNetV3   | 94.87    | 94.03| 1755          | 1847            |
| C*0.6+d*1+SE  | 95.81    | 95.04| 1755          | 1860            |
| C*0.7+d*0.7+SE| 96.25    | 95.59| 1755          | 1885            |
| C*0.7+d*1+SE  | 97.3     | 96.53| 1755          | 1881            |
| C*1+D*1+CBAM  | 97.24    | 96.44| 1755          | 1885            |
| MobileNetV2   | 87.7     | 87.31| 1755          | 1628            |
| ResNet101     | 93.05    | 92.49| 1755          | 1757            |
| C*0.7+d*1+CBAM| 97.57    | 96.73| 1755          | 1863            |
| MobileNetV3   | 91.44    | 90.12| 1800          | 1806            |
| C*0.6+d*1+SE  | 90.3     | 89.12| 1800          | 1797            |
| C*0.7+d*0.7+SE| 92.74    | 91.79| 1800          | 1823            |
| C*0.7+d*1+SE  | 93.33    | 92.39| 1800          | 1841            |
| C*1+D*1+CBAM  | 94.36    | 93.28| 1800          | 1858            |
| MobileNetV2   | 87.53    | 86.98| 1800          | 1664            |
| ResNet101     | 92.25    | 91.51| 1800          | 1782            |

V. CONCLUSIONS

In this paper, a lightweight network model is constructed for the implementation of dense bamboo-stick counting on mobile terminals. The divide-and-conquer concept is adopted to solve the problem that it is difficult to directly extract effective features from dense bamboo-stick images. Experimental verification shows that it is very clear that the number of model detections is close to the actual number. Herein, we propose a Bneck scaling factor suitable for the bamboo-stick detection task that not only reduces the amount of the model’s parameter calculations but also further improves the accuracy of bamboo-stick detection. The proposed algorithm integrates a spatial-temporal attention mechanism to replace the original model channel attention mechanism, extracts more effective features, and further reduces the counting error of dense bamboo sticks, providing new ideas for future research directions.

Compared with existing bamboo-stick counting methods, the proposed method greatly improves the real-time efficiency of dense bamboo-stick counting, and it has the advantages of low cost, simple maintenance, and convenient operation compared with large-scale, photoelectric-sensor-based counting equipment. It provides a new technology for promoting the automation of and building intelligence into the bamboo industry. In future work, it will be necessary to further optimize the errors that are difficult to eliminate in the cutting edge detection of dense bamboo-stick images, and further improve the accuracy of dense bamboo-stick recognition. In planned follow-up work, we will collect more samples in actual production and apply the research results to actual bamboo stick production.
REFERENCES

[1] S. V. Naik, S. K. Majjigudda, S. Naik, S. M. Dandin, U. Kulkarni, S. Meena, S. V. Gurfalahour, and P. Benagi, "Survey on Comparative Study of Pruning Mechanism on MobileNetV3 Model," IEEE 2021 International Conference on Intelligent Technologies (CONIT), pp. 1-8, June 2021.

[2] A. Banerjee, P. Washington, C. Mutlu, A. Kline, and D. P. Wall, “Training and Profiling a Pediatric Emotion Recognition Classifier on Mobile Devices,” arXiv preprint arXiv:2108.11754, 2021.

[3] Q. Wang, J. Ke, J. Greaves, G. Chu, G. Bender, L. Sbaiz, A. Go, A. Howard, M.-H. Yang, and J. Gilbert, “Multi-path Neural Networks for On-device Multi-domain Visual Classification,” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3019-3028, 2021.

[4] S. Qian, C. Ning, and Y. Hu, “MobileNetV3 for Image Classification,” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 490-497, April 2021.

[5] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, and V. Vasudevan, "Searching for mobilenetv3." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314-1324, 2019.

[6] M. Tan, and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks." Proceedings of the 36th International Conference on Machine Learning, pp. 6105-6114, 2019.

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

[8] H.Z.XU, "The Design of the Bamboo Stick Automatic Counting Machine." Equipment Manufacturing Technology, no. 2, pp. 87-89, 2016.

[9] Z. Qin, Z. Zhang, X. Chen, C. Wang, and Y. Peng, "Fé-d-mobilnet: Improved mobilnet with a fast downsampling strategy," In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1363-1367, Oct. 2018.

[10] H.-Y. Chen, and C.-Y. Su, "An enhanced hybrid MobileNet.,” 2018 9th International Conference on Awareness Science and Technology (iCAST), pp. 308-312, Nov.2018.

[11] D. Sinha, and M. El-Sharkawy, "Thin mobilnet: An enhanced mobilnet architecture." 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0280-0285, 2019.

[12] W. Wang, Y. Li, T. Zou, X. Wang, J. You, and Y. Luo, “A novel image classification approach via dense-MobileNet models.,” Mobile Information Systems, vol. 2020, Jan.2020.

[13] K. Tong, Y. Wu, and F. Zhou, “Recent advances in small object detection based on deep learning: A review,” Image and Vision Computing, vol. 97, pp. 103910, 2020.

[14] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, “Perceptual generative adversarial networks for small object detection,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1222-1230, 2017.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117-2125, 2017.

[16] B. Singh, and L. S. Davis, "An analysis of scale invariance in object detection snip." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3578-3587, 2018.

[17] Z. Liang, J. Shao, D. Zhang, and L. Gao, "Small object detection using deep feature pyramid networks." Pacific Rim Conference on Multimedia, pp. 554-564, Sep.2018.

[18] Y. Chen, P. Zhang, Z. Li, Y. Li, X. Zhang, G. Meng, S. Xiang, J. Sun, and J. Jia, “Stritcher: Feedback-driven data provider for object detection,” arXiv e-prints, pp. arXiv:2004.12432, 2020.

[19] X. Ran, and L. Ren, “Search aid system based on machine vision and its visual attention model for rescue target detection.” 2010 Second WRI Global Congress on Intelligent Systems, Vol. 2, pp. 149-152, 2010.

[20] Z. Jian, Z. Yonghui, Y. Yan, L. Ruonan, and W. Xueyao, "MobileNet-SSD with adaptive expansion of receptive field."2020 IEEE 3rd International Conference of Safe Production and Informatization (ICSPSI), pp. 177-181, 2021.

[21] X. Zhang, N. Li, and R. Zhang, "An Improved Lightweight Network MobileNetv3 Based YOLOv3 for Pedestrian Detection." 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 114-118, 2021.

[22] J. Nie, Y. Pang, S. Zhao, J. Han, and X. Li, “Efficient Selective Context Network for Accurate Object Detection,” IEEE Transactions on Circuits and Systems for Video Technology, Vol. 31, no. 9, pp. 3456 - 3468, Nov.2020.

[23] D. Jiang, B. Sun, S. Su, Z. Zuo, P. Wu, and X. Tan, “FASSD: A feature fusion and spatial attention-based single shot detector for small object detection,” Electronics, vol. 9, no. 9, pp. 1536, 2020.

[24] W. Li, K. Liu, L. Zhang, and F. Cheng, “Object detection based on an adaptive attention mechanism,” Scientific Reports, vol. 10, no. 1, pp. 1-13, 2020.

[25] J. Wang, Y. Wang, Y. Wu, K. Zhang, and Q. Wang, “FRPN: A Feature-Refining Pyramid Network for Object Detection of Remote Sensing Images,” IEEE Geoscience and Remote Sensing Letters, pp. 1-5, 2020.
[26] J.-S. Lim, M. Astrid, H.-J. Yoon, and S.-I. Lee, “Small object detection using context and attention.” In 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), pp. 181-186. Available: https://ieeexplore.ieee.org/abstract/document/9415217

[27] H. Fu, G. Song, and Y. Wang, “Improved YOLOv4 Marine Target Detection Combined with CBAM,” Symmetry, vol. 13, no. 4, pp. 623, 2021.

[28] M. Canayaz, “C+ EffxNet: A novel hybrid approach for COVID-19 diagnosis on CT images based on CBAM and EfficientNet,” Chaos, Solitons & Fractals, vol. 151, pp. 111310, 2021.

[29] X. Lu, E. Y. Chang, C.-n. Hsu, J. Du, and A. Gentili, “Multi-Classification Study of the Tuberculosis with 3D CBAM-ResNet and EfficientNet.” CLEF2021 Working Notes. CEUR Workshop Proceedings, Bucharest, Romania, CEUR-WS, Sep. 2021.

LIANGQUAN JIA, born in Oct. 1983, he is an AI engineer of School of Information Engineering of HuZhou University, and doing Postdoctoral work in Zhejiang University. he is currently working on the research of Object detection, Semantic segmentation and image processing.

YAWEN WANG, born in July. 1998, she is a graduate student of School of Information Engineering of Huzhou University. She is currently working on the research of Object detection and image processing.

YING ZANG, born in Dec. 1981, she is an AI engineer of School of Information Engineering of HuZhou University. She is a doctoral student at the University of Chinese Academy of Sciences. She is currently working on the research of Object detection, Semantic segmentation and image processing.

QUANLING LI, born in Feb. 1996, he is a graduate student of School of Information Engineering of HuZhou University. he is currently working on the research of object detection, image restoration and image style conversion.

HUANAN LENG, born in Mar. 1984, he is a forestry engineer. He has been engaged in ecological related work for a long time. He has been involved in ecological work such as afforestation, mine restoration and water environment treatment, and has a more comprehensive ecological concept and grass-roots practical experience.

ZHANCHUN XIAO, born in May. 1979, he is an AI engineer of Anji Bata Robot Ltd, and doing R & D director work in Zhejiang Province. he is currently working on the research of Object detection, Semantic segmentation and image processing.

WEI LONG, born in Aug. 1978, he is a lecturer in School of Information Engineering of Huzhou University. His currently interests are include in the research of Object detection, Semantic segmentation and Deep learning.

LINHUA JIANG, PhD. Special appointed professor, school of information engineering, HuZhou University. Director of the AI Research Group. His current research fields include artificial intelligence, signal and image processing, computer networks, smart Internet-of-Things research, etc. Prof JIANG published more than one hundred scientific papers and conference articles.