ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 14, No. 4, 2019, *pp*.291-299

Insect real-time detection in complex environments based on improved YOLOV3

Juan Wang [1]

[1] College of Information and Science Technology, Jinan University,
Guangzhou, 51000, China

(Received November 02 2019, accepted December 28 2019)

Abstract. The combination of advanced computer vision technology and insect image recognition technology can be effectively applied to environmental monitoring, pest diagnosis, epidemiology and other fields. However, accurate location and classification of relatively small insects in complex scenes has always been a difficult problem for this technology. Although YOLOv3 combines deep features with shallow features to facilitate the detection of small objects, experiments have found that YOLOv3 has more undetected cases for small target insects in complex backgrounds. In order to solve this problem, this paper optimizes YOLOv3 algorithm. Firstly, the SE blocks are embedded into YOLOv3 network to learn global features and enhance the expression ability of feature maps, so that the network can detect more objects. Because YOLOv3 itself has a complex network structure and the amount of parameters and calculations are increased after embedding the SE block, so this paper also uses depthwise separable convolutions, which greatly reduces the amount of parameters and computation under the condition of little loss of accuracy, thus improving the detection speed. Training and testing on the insect dataset made in this paper, the original YOLOv3 runs at 33 f / s, and the mean Average Precision (mAP) is only 86.8%, While the improved YOLOv3 runs at 38 f / s, the mAP reaches 90.6%. The improved algorithm can detect more targets, reduce the omission factor and improve the detection speed.

Keywords: target detection, real-time detection, convolution neural network, YOLOv3.

1. Introduction

Accurate and timely identification of insects is the basis for monitoring agricultural pest information and crop pest control. Through remote image acquisition, photo detection and other means, combined with machine vision and image recognition of pest identification technology, will help to improve accuracy and efficiency, reduce losses caused by pests, thus promoting the implementation of precision agriculture, while also saving labor and time costs.

Classical insect image recognition technology usually requires three steps: image preprocessing, feature extraction, and classifier classification. For example, Kandalkar et al. [1] proposed a wavelet transform algorithm to extract features and use BP neural network to classify and identify agricultural pests. . Han Ruizhen [2] used gray level co-occurrence matrix, geometric invariant moment and other methods to extract features, and used support vector machine for classification and recognition. The recognition accuracy on the test set was 89.5%, and it took 1.5 seconds to identify an image. The process of image preprocessing is complicated and requires steps such as image graying, image denoising, image segmentation, etc. The result of target detection is directly related to the feature extraction algorithm used, while the manually designed feature extraction algorithm is not very robust to the diversity of target features. In recent years, researchers have begun to apply the relevant techniques of deep learning to insect image recognition. For example, Liang Wanjie et al. [3] applied convolution neural network to the identification of rice pests, reaching 89.14% identification accuracy. Cheng Xi et al. [4] used deep convolution neural network to classify and identify seven stored grain pests, and the test accuracy on Alexnet[5] and GoogLeNet[6] reached 97.61%. Liu et al. [7] introduced the convolution neural network, and the accuracy rate of identifying 12 kinds of rice field pests reached 93.2%, and the average time to identify an image was about 4 milliseconds.

YOLOv3 [8] is a target detection algorithm with relatively balanced speed and precision, which combines deep features with shallow features, retains fine-grained features and obtains more meaningful semantic information, so that small objects can be detected in real time. However, experiments have found that YOLOv3

[1] Corresponding author. E-mail address: satakiolo@163.com.

has more missed detection for small target insects in complex background. In order to solve this problem, based on the YOLOv3 network, this article optimizes the YOLOv3 algorithm for the self-made insect dataset. The SE blocks [9] are embedded in the YOLOv3 network to reduce the interference of light, background complexity, and high similarity between insects and the environment, and improve the mAP of target detection. On this basis, depthwise separable convolutions [10] are used to improve mAP while reducing the amount of computation and parameters, so as to achieve the purpose of accurate and real-time insect detection.

The main work of this article are as follows:

- To solve the problem of missed detection of target insects in complex background by YOLOv3 network, SE blocks are embedded into YOLOv3 network to obtain SE-resistant structures, which will introduce the original information into the deep layers to inhibit the degradation of information, then pool and expand the receptive field, fuse the shallow layers information and the deep layers information from multiple angles, so that the combined output contains multi-level information, can learn the global features, and enhance the expression ability of the feature map.
- To solve the problem that YOLOv3 itself has a complex network structure and increases the amount of parameters and computation after embedding SE block, the use of depthwise separable convolutions can greatly reduce the amount of parameters and computation under the condition of little loss of accuracy.

2. YOLOv3

YOLO algorithm is a regression-based target detection method proposed by Redmon et al. [11] in 2016. YOLOv3 has been developed in 2018. It can detect a variety of objects by only one forward operation, so Yolov3 series of algorithms have a fast detection speed. YOLOv3 borrows the ideas of ResNet [12], introduces a plurality of residual network modules and uses a multi-scale prediction method to improve the defects of YOLOv2 in small target recognition, so YOLOv3 still maintains the fast detection speed of YOLOv2[13], and the recognition accuracy rate is greatly improved, especially in the detection and recognition of small targets, the accuracy rate is greatly improved.

YOLOv3 designed the basic model of classification network Darknet53. Darknet53 include convolution layers and Residual layers. The convolution layers are obtained by integrating convolution layers with better performance from various mainstream network structures, and each convolution is followed by Batch Normalization and linkyRelu activation operations. The Residual layers uses ResNet's Residual structure for reference and are mainly composed of 1×1convolutions and 3×3convolutions. Using this structure can make the network structure deeper and avoid gradient disappearance and gradient explosion while extracting deeper features [12]. YOLOv3 also introduced the idea of anchor boxes[14]. For COCO dataset and VOC dataset, it uses three scales for prediction. These three different scales come from the output of convolution layers at different levels, and each scale has three anchor boxes. This method borrows from the idea of FPN [15]: compared with prediction using shallow feature maps directly and up-sampling prediction with deep feature maps, the latter helps retain fine-grained features and obtain more meaningful semantic information, and is more conducive to detecting small objects.

Fig. 1 shows the network structure of YOLOv3. YOLOv3 first converts the input image into a size of 416×416, extracts image features through Darknet53, and then uses a feature pyramid structure similar to FPN network to output a feature map of three scales. Among the output feature maps, the 13×13 size feature map is responsible for detecting large objects, the 26×26 size feature map is responsible for detecting medium objects, and the 52×52 size feature map is responsible for detecting smaller objects.

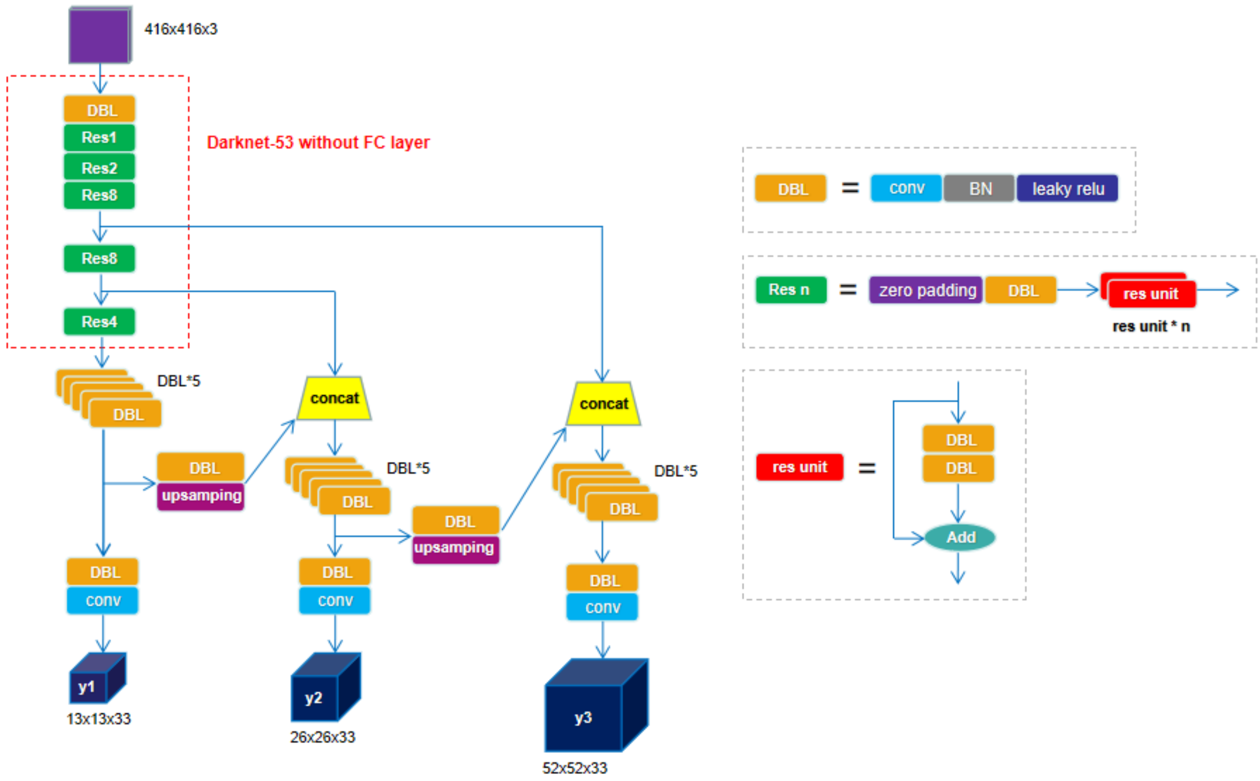


Fig. 1: YOLOv3 network structure

3. Improvements based on YOLOv3

In order to detect small target insects in complex environment, this paper proposes two improvements to YOLOv3 in improving accuracy and speed.

3.1. SE blocks embedded design

Hu et al. [9] proposed a novel network structure unit called sequeeze-and-exception(SE) block in 2017. The structure consists of three parts: Squeeze, Excitaion and Reweight. By explicitly modeling the correlation between channels to improve the network's representation ability, it can enhance the receptive field and balance the confidence of the feature map at different locations, so that the network can learn Global characteristics. The SE block uses a mechanism that allows the network to recalibrate features. This mechanism automatically obtains the importance of each feature channel through learning, and then promotes useful features and suppresses features that are of little use to the current task according to this importance. Darknet53, which is used for feature extraction, is the key for YOLOv3 network to accurately predict the results. In this paper, SE block is embedded into the Residual layers of Darknet53 network to obtain the SE-Residual substructure, thus expanding the perception range of feature map to global information, and making the effective feature map have significant weight, while the invalid or ineffective feature map have small weight.

The SE block is first a Squeeze operation, which performs global average pooling on the input feature maps to obtain a feature map with a size of $1 \times 1 \times C$.

It will change each two-dimensional feature channel into a real number, which has a global receptive field to some extent, and makes the layers close to the input also get a global receptive field. Then the Excitation operation, through two fully connected layers to model the correlation between the channels, and then obtain the normalized weights between 0 and 1 through the sigmoid function. The operation of reducing the dimensions first and then increasing the dimensions has more nonlinearity than directly using a Fully Connected layer, which can better fit the complex correlations between the channels, and greatly reduces the amount of parameters and calculations. Finally, the Reweight operation considers the weight of the output of Excitation as the importance of each feature channel after feature selection, and then multiplies the previous feature map by channel through multiplication to complete the recalibration of the original feature in the channel dimension.

As shown in Fig. 2, the SE module is embedded into the Residual layers of YOLOv3. This article chooses to recalibrate the Residual features on the branch before Add to obtain SE-Residual. Liu et al. [16] chose to recalibrate the features on the main branch after Add. However, due to the scale operation of 0~1 on the main branch, gradient dissipation is easy to occur near the input layer when BP optimization is carried out on deeper networks, which makes the model difficult to optimize.

Under the multi-scale training strategy, the mAP of the self-made insect dataset under the original YOLOv3 is 86.8%, while the YOLOv3 based on SE block proposed in this paper can improve the mAP to 91.0%. Relatively speaking, the parameter quantity increases from 61M to 62M, and the detection speed decreases.

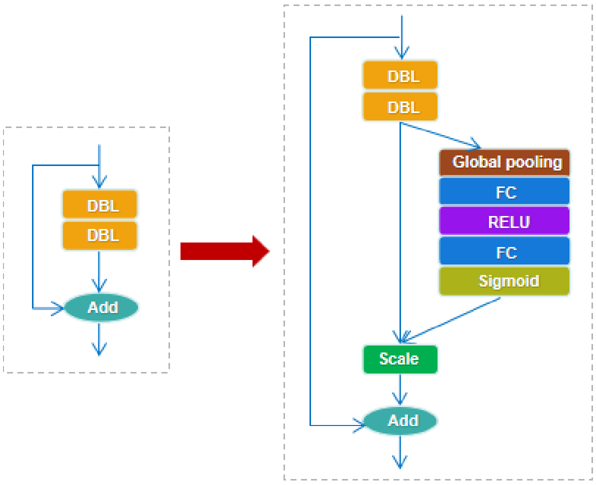


Fig. 2: SE-Residual Module

3.2. Depthwise Separable Convolutions

Because YOLOv3 itself has a complex network structure, and embedding SE block into YOLOv3 to improve mAP will bring more parameters and computation, this paper uses depthwise separable convolutions instead of ordinary convolution in YOLOv3, which greatly reduces the parameters and computation without much loss of accuracy. Depthwise separable convolutions(DepthSepConv) is first introduced by F.Chollet [10], and consists of two parts: depthwise convolutions and pointwise convolutions. It first performs convolution on each channel separately, and then mixes these outputs by point-by-point convolution, which achieves the separation of channels and regions. Depthwise separable convolutions requires much fewer parameters and less computation.

Fig. 3 is a schematic diagram of depthwise separable convolution. Assuming that the input feature map size is $H \times W \times C$, the output feature map size is $H \times W \times N$, and the convolution kernel size is 3×3, then the calculation amount of ordinary convolution is: $H \times W \times C \times N \times 3 \times 3$. Depthwise refers to dividing the $H \times W \times C$ input into C groups, and then performing 3×3 convolution on each group.This is equivalent to collecting the spatial features of each channel. The calculation of depthwise is:$H \times W \times C \times 3 \times 3$. Pointwise refers to doing Nordinary 1×1 convolutions on the input of $H \times W \times C$.This is equivalent to collecting the features of each point. The calculation of pointwise is:$H \times W \times C \times N$.The splitting of Depthwise and Pointwise is equivalent to compressing the calculation of ordinary convolution into:

$$\frac{depthwise + pointwise}{conv} = \frac{H \times W \times C \times 3 \times 3 + H \times W \times C \times N}{H \times W \times C \times N \times 3 \times 3} = \frac{1}{N} + \frac{1}{3 \times 3} \tag{1}$$

After the experiment, it was found that the parameter quantity was reduced from 62.5M to 18.6M after the depthwise separable convolution was used to replace the ordinary convolution for Darknet53 embedded in SE block and FPN structures, but mAP was reduced from 91.0% to 83.6%, even lower than 86.8% of the original YOLOv3. After analysis, backbone for feature extraction is the key to accurate prediction of YOLOv3 network. Depthwise separable convolutions inevitably reduce the accuracy while increasing the speed. Therefore, only the blue dashed box in Fig. 4 is used in depthwise separable convolution in this paper, which reduces the amount of network parameters and computation while not affecting the network to extract features. The parameter quantity obtained by this method is 45.9M, which is 16.6M lower than 62.5M, and mAP is 90.6%, which is only 0.4% lower than 91.0%.

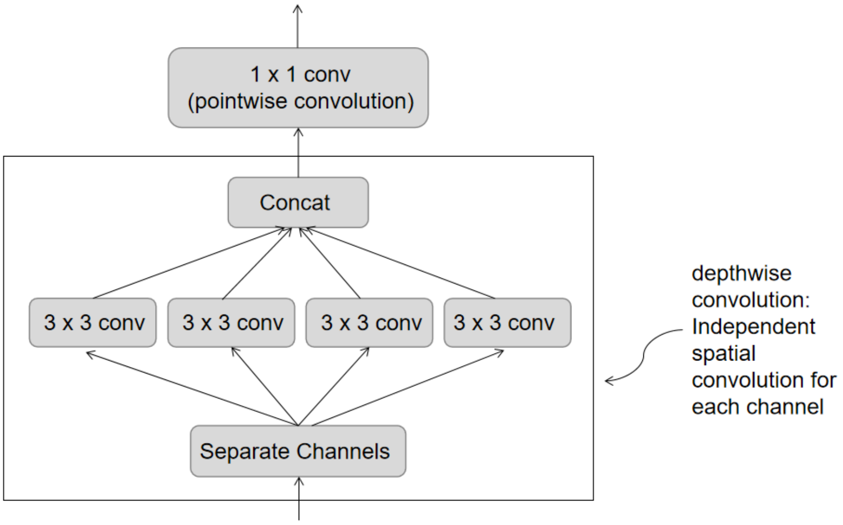


Fig. 3: Depthwise Separable Convolution

The resulting YOLOv3-SD network structure is shown in Fig. 4.

4. Experimental results and analysis

4.1. Experimental configuration and evaluation index

Experimental equipment configuration: CPU is Intel(R) Core(TM) i7-8700, GPU is NVIDIA GeForce GTX 1080Ti, operating system is Ubuntu 16.04LTS, and deep learning framework is Keras.

Configuration of network parameters: batchsize is 8, gradient optimization uses the Adam algorithm, the initial learning rate is 0.001, the momentum is 0.9, the attenuation coefficient is 0.0005 and Early Stopping strategy is used in training, which is a regularization method to avoid over-fitting of the network.

Evaluation index: This paper uses mAP, transmission rate and processing time to evaluate the proposed network model. Among them, mAP is the performance measurement standard for predicting the target location and category. AP measures the quality of the learned model in each category, and the mAP measures the quality of the learned model in all categories. Transmission rate refers to the number of images N processed by the algorithm per second, and its unit is frame per second (f / s). Processing time refers to the time required by the algorithm to process each image t = 1 / N, and its unit is ms. Generally speaking, the processing speed of the algorithm reaches 30 f / s can be considered as real-time.

4.2. Insect Dataset

Deep learning model has higher requirements on the sample data. In fact, the sample size determines the performance of the model to a considerable extent.Therefore, we use data enhancement [17] to increase the data volume of the sample.In this paper, we use the methods of adjusting saturation, exposure, color tone, horizontal flip and increasing noise to enhance the data of the images collected by cameras and the internet, and manually mark the positions of insects in the images by LabelImg software, thus 4394 insect images are obtained, of which 3600 images are randomly selected as training dataset, 400 as verification dataset and 394 as test dataset to detect the training effect.

4.3. Comparison of generating schemes for anchor boxes

Anchor boxes are some initial candidate boxes with fixed width and height that assist in predicting target boundaries and are designed based on different datasets. Its number is set manually, and its setting will affect the accuracy and speed of target detection.

Although YOLOv3 has achieved satisfactory results on COCO dataset, it is not suitable for self-made insect target dataset. If the original anchor box parameters are directly used for training to generate weight files, it is found that during the test, it is easy to misrecognize the target and the recognition rate is low. In order to adapt to the target box in the self-made insect dataset, this paper uses k-means algorithm to carry out dimensional clustering analysis on the statistical rules of the target boxes in the dataset, and obtains the number and wide-height scale of anchor boxes suitable for the self-made dataset, so as to achieve the optimal recognition effect.

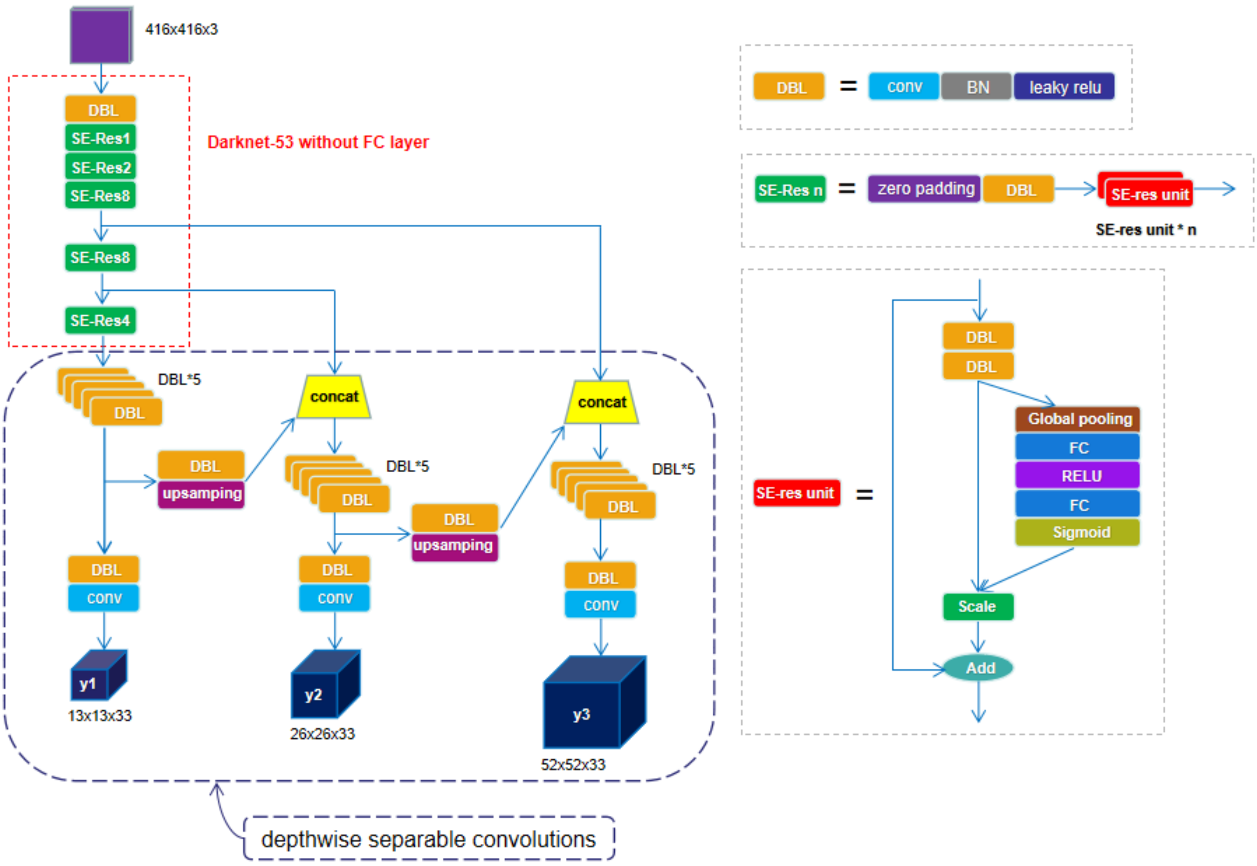


Fig. 4: Improved YOLOv3 network structure

The k-means algorithm is used to perform dimensional clustering analysis on the width and height of the target boxes. The change of the object function when different K values are taken is shown in Fig. 5. In the experiment, the elbow method was used to select the K value. From the figure, it can be seen that the k value corresponding to the elbow is 9. Therefore, for the clustering of self-made insect dataset, the best clustering number should be 9. When k = 9, the clustering result is shown in Fig. 6. The coordinates of the clustering center in the clustering results are taken as the wide and high dimensions of anchor boxes, and different colors in the figure correspond to the target boxes in different regions.

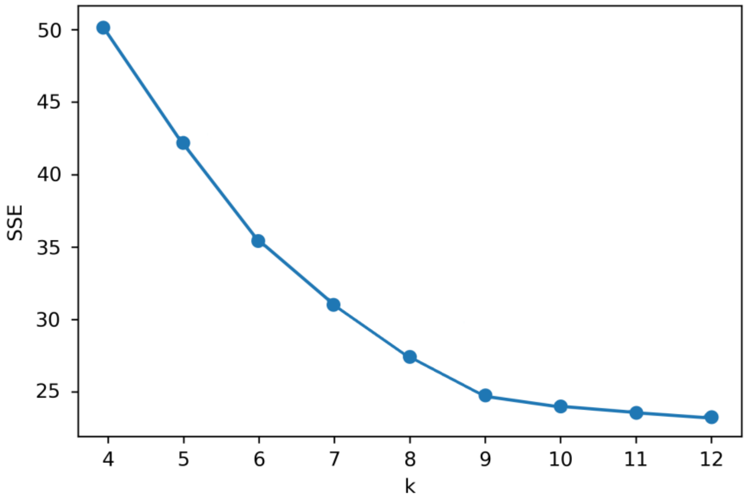


Fig. 5: Variation curve of k value

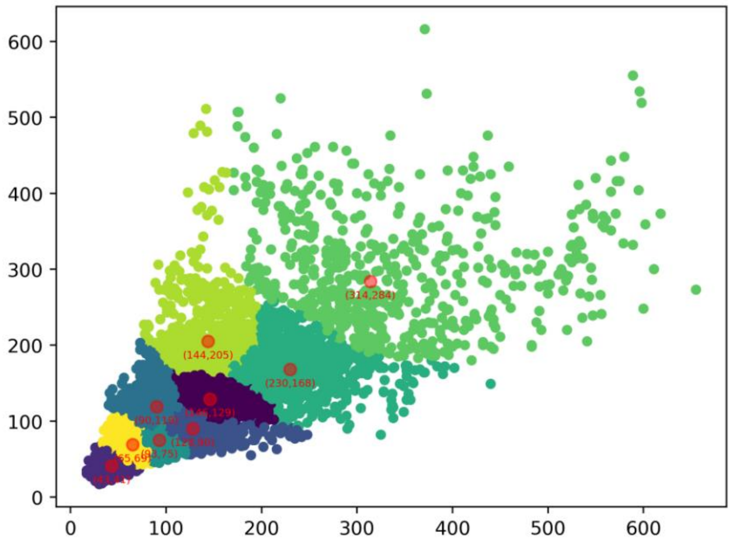


Fig. 6: Cluster distribution of anchor boxes

In this paper, k-means clustering analysis is performed on the target frames in the self-made insect dataset to obtain the optimal anchor number and width-height scale. Because the dataset contains some large-resolution pictures, some large cluster center coordinates will be generated. The Average IOU and mAP are shown in Table. 1.

Generation scheme	Generated anchor boxes	Avg IOU/%	mAP/%
YOLOv3	(10,13),(16,30),(33,23), (30,61),(62,45),(59,119), (116,90),(156,198),(373,326)	67.41	86.8
Improved YOLOv3	(43,41),(65,69),(90,119), (93,75),(128,90),(144,205), (146,129),(230,168),(314,284)	76.75	87.6

Table. 1: Anchor boxes Generation Scheme Comparison Table

The Average IOU of anchor boxes and real boxes in YOLOv3 is 67.41%, and the Average IOU of anchor boxes and real boxes after re-clustering with k-means is 76.75%, which improves the degree of coincidence between anchor boxes and real boxes in the dataset, making it easier to learn insect detection tasks.

4.4. Performance Comparison of Algorithms

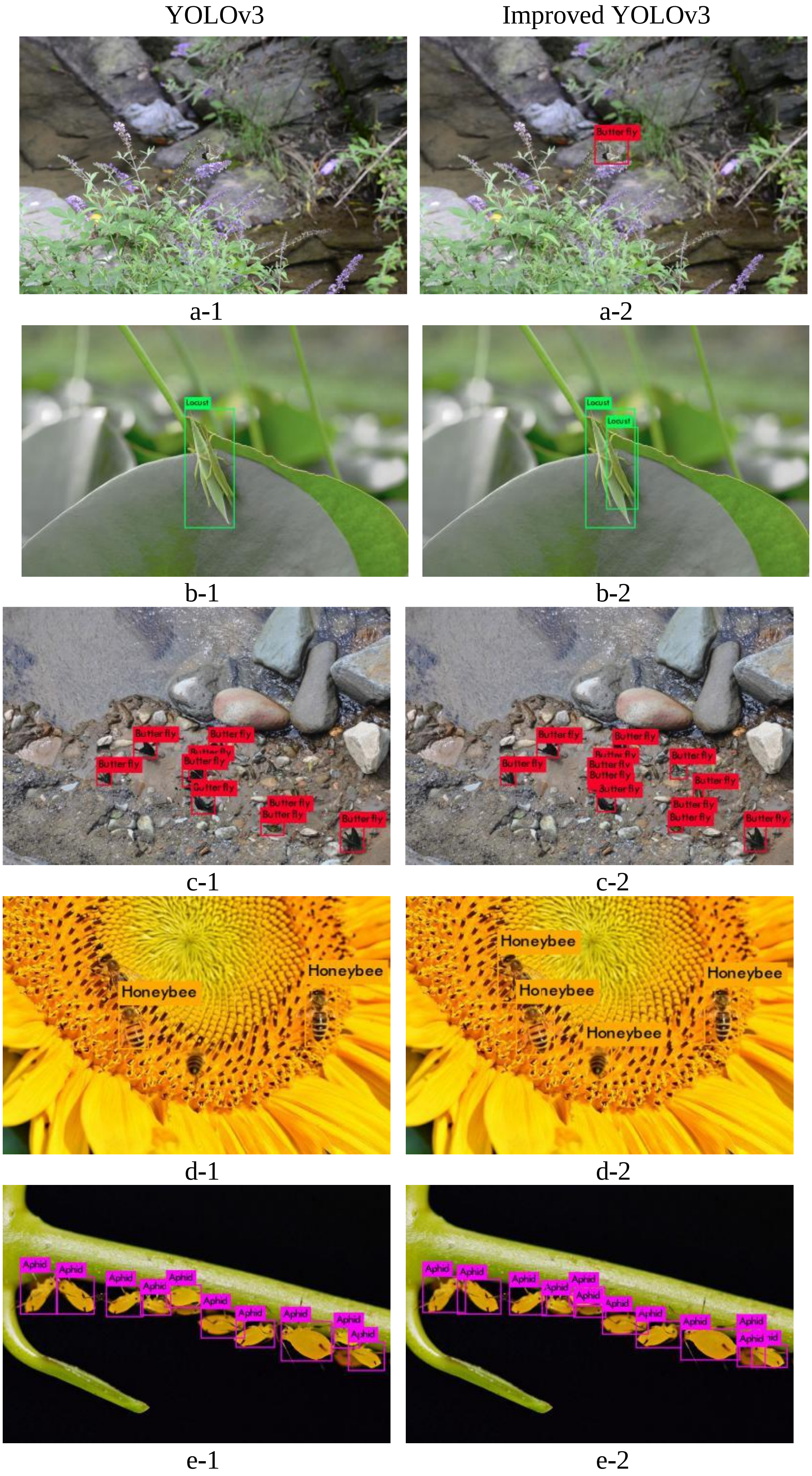
This paper trains and tests the proposed improved YOLOv3 network on 416×416 images, and compares it with YOLOv3. The performance comparison is shown in Table. 2.

	YOLOv3			Improved YOLOv3
New anchor boxes		√	√	√
SE block			√	√
DepthSepConv				√
Parameter quantity/(M)	61.6	61.6	62.5	45.9
mAP/%	86.8	87.6	91.0	90.6
Transmission rate/(f / s)	33.0	33.0	27.0	38.0
Processing time/(ms)	30.3	30.3	37.0	26.3

Table. 2: YOLOv3 vs. improved YOLOv3

As can be seen from the table, for the self-made insect dataset, the mAP of YOLOv3 using the original anchor boxes is 86.8%, while the mAP of anchor boxes regenerated using the K-means clustering algorithm is 87.6%.

When only the SE blocks are embedded, the parameter amount is increased by 0.9M compared to YOLOv3, and the detection speed is 27 f / s. If all 3x3 convolutions are replaced by depthwise separable convolutions, the parameter amount is 18.6M, but the mAP is only 83.6 %, even lower than the original YOLOv3's mAP. The improved YOLOv3 used in this paper obtains 90.6% mAP, which is 3.8% higher than YOLOv3, and realizes real-time detection at a speed of 38 f / s , which reduced the parameter amount by 15.7M compared to YOLOv3.


YOLOv3 Improved YOLOv3

a-1 a-2

b-1 b-2

c-1 c-2

d-1 d-2

e-1 e-2

Fig. 7: Comparison of test results

In view of the problems in insect dataset, such as small size of insects, vague insect targets, easy overlap or occlusion, extremely similar insect colors and background colors, etc., using the improved network has a more accurate recognition effect, which can detect more insect targets and greatly reduces the omission factor. The comparison of the detection results is shown in the following figure. The left is the detection result of YOLOv3, and the right is the detection result of improved YOLOv3. As can be seen from the figure, YOLOv3 failed to detect all insect targets due to the extremely similar insect color and background color and the phenomenon of overlapping occlusion, but could detect all insect targets after using the improved YOLOv3 algorithm.

5. Conclusion

Based on the fact that YOLOv3 has a large number of missed detection problems for target insects in complex background, this paper proposes an improved YOLOv3 algorithm. The algorithm uses K-means algorithm to re-cluster anchor boxes, and embeds SE block in YOLOv3, which enhances the receptive field of the network and makes the feature information learned by the network more comprehensive. The depthwise separable convolutions are used to reduce the amount of parameters and computation, thus improving the detection speed on the premise of ensuring the accuracy. Compared with the self-made insect dataset, the experimental results verify that the improved YOLOv3 algorithm has better robustness to insect targets, while the running speed reaches 38 f / s, the mAP reaches 90.6%, which can detect more insect targets at the same time and reduce the omission factor.

References

[1] Kandalkar G, Deorankar A V, Chatur P N, Classification of Agricultural Pests Using DWT and Back Propagation Neural Networks, International Journal of Computer Science & Information Technology, 2014, 5(3): 4034-4037.

[2] Ruizhen Han, Research on Fast Detection and Identification of Agricultural Pests Based on Machine Vision, Zhejiang University, 2014.

[3] Wanjie Liang, Hongxin Cao, Identification of Rice Pests Based on Convolution Neural Network, Jiangsu Agricultural Sciences, 2017, 45(20): 241-243.

[4] Xi Cheng, Yunzhi Wu, Youhua Zhang, et al. Image Recognition of Stored Grain Pests Based on Depth Convolution Neural Network, Chinese Agricultural Science Bulletin, 2018, 34(1): 154-158.

[5] Krizhevsky A, Sutskever J, Hinton G E, Imagenet classification with deep convolutional neural networks, In Advances in neural information processing systems. pp. 1097-1105, 2012.

[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1-9,2015.

[7] Liu Z, Gao J, Yang G, et al. Localization and Classification of Paddy Field Pests using a Saliency Map and Deep Convolutional Neural Network, Sci Rep, 2016, 6: 20410.

[8] Redmon J, Farhadi A, YOLOv3: An Incremental Improvement, arXiv preprint arXiv: 1804.02767, 2018.

[9] Jie Hu, Li Shen,et al. Squeeze-and-Excitation Networks, arXiv preprint arXiv: 1709.01503, 2017.

[10] F.Chollet, Xception: Deep learning with depthwise separable convolutions, In CVPR, 2017.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, Real Time Object Detection. In Computer Vision and Pattern Recognition. 2016: 779-788.

[12] 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.

[13] Redmon J, Farhadi A, YOLO9000:Better,Faster,Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6517-6525, 2017.

[14] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. International Conference on Neural Information Processing Systems. Cambridge, USA: MIT, 91-99, 2015.

[15] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature Pyramid Networks for Object Detection, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117-2125, 2017.

[16] Xueping Liu, Yuqian Li, Li Liu,et al. Improved YOLOV3 target recognition algorithm embedded in SENet structure, Computer Engineering, 2019, 45(11), 243-248.

[17] Li Y, Liu X Y, Zhang H Q, et al, Optical remote sensing image retrieval based on convolutional neural networks, Optics and Precision Engineering, 2018, 26(1): 200-207.