Scientific Research Publishing

Journal of Computer and Communications, 2020, 8, 285-294
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Garbage Classification Detection Based on Improved YOLOV4

Qingqiang Chen[1,2], Qianghua Xiong[3]

[1]School of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China
[2]School of Electronic Engineering and Intelligentization, Dongguan University of Technology, Dongguan, China
[3]School of Computer, Dongguan University of Technology, Dongguan, China
Email: 1144580071@qq.com

How to cite this paper: Chen, Q.Q. and Xiong, Q.H. (2020) Garbage Classification Detection Based on Improved YOLOV4. *Journal of Computer and Communications*, **8**, 285-294.
https://doi.org/10.4236/jcc.2020.812023

Received: November 10, 2020
Accepted: December 24, 2020
Published: December 31, 2020

Abstract

As the rate of garbage generation gradually increases, the past garbage disposal methods will be eliminated, so the classification of garbage has become an inevitable choice. The multi-category classification of garbage and the accuracy of recognition have also become the focus of attention. Aiming at the problems of single category, few types of objects and low accuracy in existing garbage classification algorithms. This paper proposes to use the improved YOLOV4 network framework to detect 3 categories, a total of 15 objects, and find that the average accuracy is 64%, Frame per second 92f/s. It turns out that the improved YOLOV4 can better detect garbage categories and is suitable for embedded devices.

Keywords

YOLOV4, Computer Vision, Classification, Convolutional Network

1. Introduction

Traditional garbage disposal methods such as burial, incineration, chemical corrosion, etc. will cause great environmental pollution to soil and air, and at the same time require a lot of costs. Therefore, garbage classification has emerged, but traditional garbage classification cannot be achieved. The detection of various types of garbage has disadvantages such as low detection accuracy and slow detection speed. Therefore, in 2019, the country called on the nation to implement garbage classification, and the garbage category is no longer only recyclable and non-recyclable, but also increased hazardous garbage and kitchen garbage. Therefore, the accurate detection and classification of multiple types of garbage are of great significance to improving the environment and improving

the quality of life.

Some progress has been made in the detection of garbage classification. For example, Yue Xiaoming *et al.* proposed an Anchor-Free CenterNet network for garbage classification and detection [1]. Sang Shenggao *et al.* proposed to combine the Internet with garbage collection, and through people reporting online, the cost of garbage collection was reduced [2]. Chen Ningxin, Yang Jiarui, etc. proposed a dry and wet garbage sorting bin based on a binary classification strategy [3]. According to the analysis function of Abaqus software, Yang Mingwei and others designed a multi-function automatic sorting trash can, and processed the trash through voice input [4]. However, because the types of garbage are very abundant, and the current knowledge of category of garbage is not popular enough at present, it is very important to classify and detect garbage to avoid mishandling of garbage.

In recent years, with the development of deep learning, neural networks have been widely used in image detection. The existing neural network models can be divided into two types. One requires the extraction of candidate frames from the original image. From Alexnet's proposal to extract image candidate frames with a sliding window, to later R-CNN based on selective search, Fast R-CNN, Faster R-CNN [5], and R-FCN, etc. The other is to input the original image and directly return to the detection target through the convolutional network. The step of extracting the candidate frame is omitted, which speeds up the image detection, but at the same time reduces the detection accuracy. The main representatives are YOLO series and SSD [6]. According to this article, the improved YOLOV4 is an improved target detection method based on the YOLO series. Although the original YOLOV4 method has high accuracy, the amount of parameters and network models involved are very large, and it is not suitable for embedded devices. Therefore, the improved YOLOV4 reduces the network structure, reduces the amount of parameters, and improves the image detection efficiency while ensuring accurate target detection. This article trains on the improved YOLOV4 object detection framework and the pre-trained weights of the VOC data set, and detects 3 categories, 15 types of garbage, and a total of 22,000 images.

2. YOLOV4 Network Introduction

The YOLOV4 algorithm has been developed from YOLOV1. After YOLOV2, YOLOV3, YOLO network can generally be divided into three parts, namely the backbone network, neck network and Prediction. The backbone network is mainly responsible for the extraction of image features, but with the development of deep learning, it is found that although the more layers of the network, the richer the extracted feature information, but it will increase the cost of training. After reaching a certain number of layers, its training effect will decrease instead. Therefore, people now increasingly expect to use lightweight layer networks to replace complex and computationally intensive neural networks on the premise of ensuring effective feature extraction from images. The

Neck network can enhance image features, process and enhance the shallow features extracted from the backbone network, and merge the shallow features with the deep features to enhance the robustness of the network, thereby obtaining more effective features. The Head network classifies and regresses the features obtained by the backbone network and the Neck network.

2.1. Data Augmentation

Because the ability to collect data sets is limited, YOLOV4 will create new training samples from existing data sets. Therefore, it is necessary to perform data augmentation operations on existing data sets to improve the generalization ability of the training model. Therefore, in the current image processing, diversified data augmentation can maximize the use of the data set, which is the key to the performance breakthrough of the object detection framework. The photometric distortion of YOLOV4 is to adjust the brightness, contrast, hue, saturation and noise of the image. Geometric distortion is the addition of random zooming, clipping, flipping and reverse flipping to the image. In addition to geometric distortion and photometric distortion, YOLOV4 also uses A series of image occlusion technologies, including Random Erase, Cutout, Hide and Seek, Grid Mask, and image blending technologies CutMix and Mosaic. Mixed-up is also the most commonly used method for data augmentation. Mixed-up adds two images in different proportions to obtain frames with different confidence levels, which helps increase the number of samples during training. In addition, SELF-Adversarial Training (SAT) is also used to resist adversarial attacks to a certain extent. The principle is that CNN calculates LOSS and then performs back propagation to form the illusion that there is no target on the picture. Performing normal target detection on the modified image, SAT can improve the robustness of the model and enhance the generalization ability. Image augmentation effect is shown in **Figure 1**.

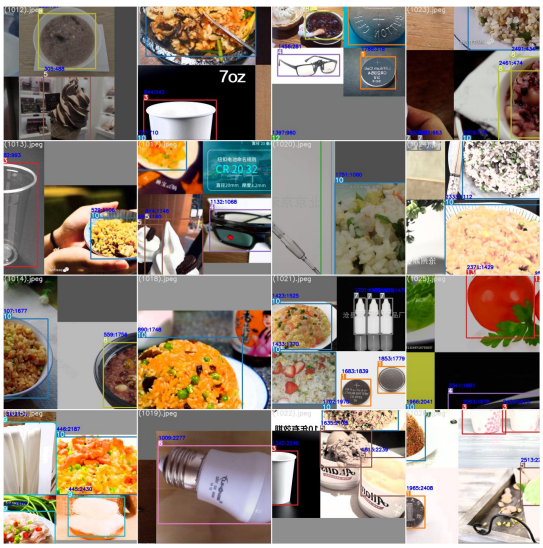

Figure 1. Data augmentation.

2.2. Backbone

The backbone network models of YOLO are ResNet-50, EfficientNet, ResNet-101, Darknet53, and some lightweight networks such as MobilenetV1, V2, ShuffleNet1, 2. The backbone network of YOLOV4 is CSPDarknet53 improved based on YOLOV3's Darknet53. CSP stands for Cross-Stage-Partial-connection. CSPDarknet53 uses DenseNet and CSP to enhance the learning ability of convolutional networks, reduce network model calculations and memory costs while maintaining accuracy. Darknet contains 5 residual modules, and CSPDarknet53 uses 1 × 1 convolution to divide the input feature map channel layer into two before each residual network, and adds CSP after each large residual module. The activation function in CSPDarknet53 is mainly the Mish function [7]. The mish function avoids the saturation caused by capping. Unlike the hard zero boundary of the RELU function, the mish function is compared with other functions. Its smooth performance can allow better information to be passed in. The neural network enhances the accuracy and generalization ability of the network, and prevents gradient disappearance and gradient explosion. The input image has three outputs after passing through CSPDarknet53.

2.3. Neck

Neck is mainly used to generate feature pyramids. The feature pyramid will enhance the model's detection of objects at different scales, so that it can recognize the same object of different sizes and scales. Before PANET came out, FPN had always been the State of the art of the feature aggregation layer of the object detection framework until the appearance of PANET (Path Aggregation Network). The Neck network in YOLOV4 uses SSP (Spatial pyramid pooling) and PANetP. SSP uses 4 different sizes of sliding kernels, 1 × 1, 5 × 5, 9 × 9, and 13 × 13 to convolve the candidate images, and then apply Multi-scale Maxpooling to get the same dimensions of feature maps [8]. SSP allows the spatial size of each candidate map to be preserved, and then connects feature maps of different core sizes as output, and the output is a fixed size feature map. The SSP network structure is shown in **Figure 2**. PANet is an innovation based on FPN and Mask RCNN. PANet proposes a more flexible ROIPooling (Region of interested Pooling) that can extract and integrate features at various scales, while FPN only extracts information from high-level feature layers.

2.4. Prediction

Prediction is mainly used in the final detection part. It applies anchor boxes on the feature map and generates the final output vector with class probabilities, object scores and bounding boxes. The feature maps of different scales output by PANet are spliced, and after convolution operation, 3 heads of different scales can be obtained. The sizes of the heads are (76 × 76 * 3 * (4 + 1 + classes)), (38 × 38 × 3 * (4 + 1 + classes)), (19 × 19 × 3 * (4 + 1 + class_num)), 4 is the coordinate value of the calibration box, 1 is the confidence level of the calibration box,

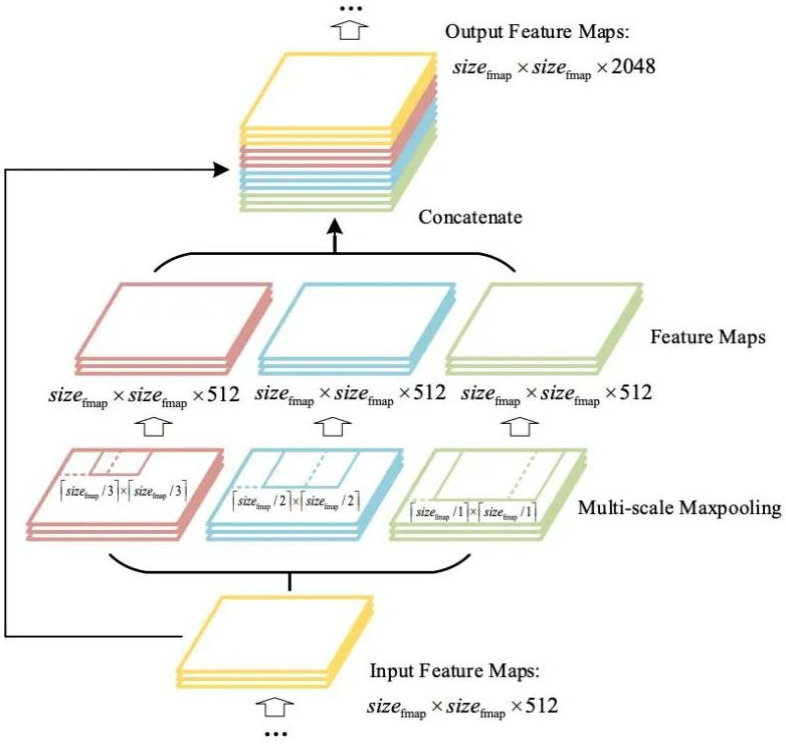

Figure 2. SSP network.

classes is the number of categories, each head contains 3 Bounding boxes, and 3 heads of different scales can be used to detect objects of different sizes. The Prediction of YOLOV4 is consistent with YOLOV3 [9]. The YOLOV4 network structure is shown in **Figure 3**.

2.5. Loss Function

The calculation of the loss function of the YOLO series is based on the two parts of Bounding Box Regression Loss and Classification Loss. The output of YOLOV4 uses Ciou-Loss as the Bounding Box regression loss, and Diou_nms as the Classification Loss. The calculation method of Bounding Box ranges from Smooth L1 Loss, IOU Loss, GIOU Loss to today's CIOU Loss. CIOU Loss takes into account the overlap area between the prediction box and the target box, the distance between the center points, and the aspect ratio, which improves speed and precision of prediction box regression. The Diou NMS is used to filter out prediction boxes with higher confidence, so as to keep the prediction boxes with the highest confidence. The formula is as follows.

$$\begin{aligned} loss &= \lambda_{cord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(x_{ij} - \overline{x}_{ij} \right)^2 + \left(y_{ij} - \overline{y}_{ij} \right)^2 \right] \\ &+ \lambda_{cord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(\sqrt{w_{ij}} - \sqrt{\overline{w}_{ij}} \right)^2 + \left(\sqrt{h_{ij}} - \sqrt{\overline{h}_{ij}} \right)^2 \right] \\ &+ \lambda_{cord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(c_{ij} - \overline{c}_{ij} \right)^2 \right] + \lambda_{cord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(p_{ij} - \overline{p}_{ij}(c) \right)^2 \right] \end{aligned} \quad (1)$$

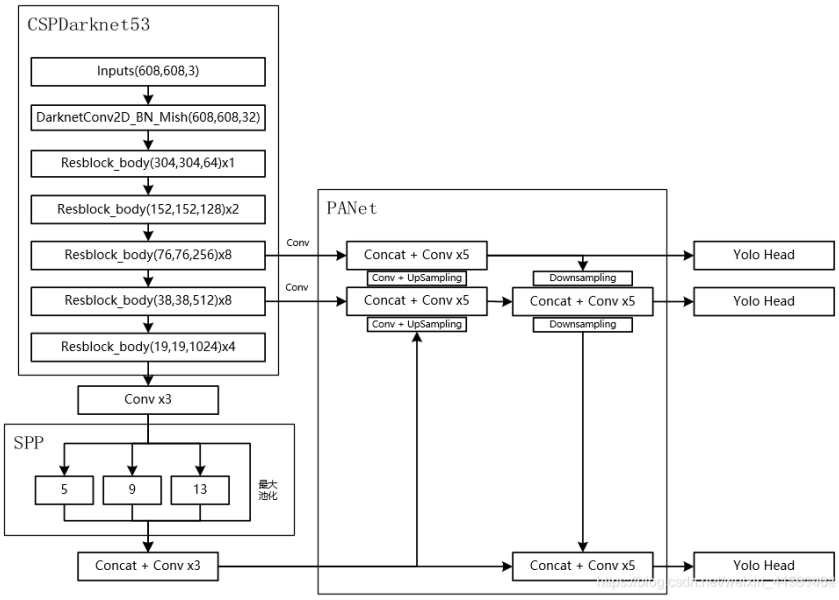

Figure 3. YOLOV4 network structure.

$$IOU = \frac{Prediction \cup GroundTruth}{Prediction \cap GroundTruth} \tag{2}$$

$$CIOU = IoU - \left(\frac{\rho^2(b, b^g t)}{c^2}\right) - \alpha\upsilon \tag{3}$$

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{4}$$

$$\alpha = \frac{\upsilon}{(1 - IOU) + \upsilon} \tag{5}$$

A_C: The minimum closure area of the prediction box and the ground truth box, U: The sum of the area of the prediction box and the real ground truth. Prediction means prediction box, Ground Truth means ground truth box, ∂: The newly added weight coefficient based on GIOU, v: The similarity of the aspect ratio between the predicted frame and the real frame, c: is the length of the diagonal of the box, ρ^2 means the square of the distance between the center of the real frame and the predicted frame.

3. YOLOV4 Improvements

Since the original YOLOV4 detects 80 categories on the VOC data set which has many categories, and in this garbage classification, only 15 types of garbage categories need to be detected, and the backbone network parameters of the original YOLOV4 reach 52.49 M, and the number of layers is 334 layers, too many floating point operations, although it can increase the accuracy of garbage classification and detection, it is not suitable for the application of embedded devices. Therefore, the original YOLOV4 backbone network extraction part can be modified to reduce the memory size calculated by the model to meet the requirements of embedded devices.

3.1. MobilenetV3

Due to the excessive amount of backbone network parameters and floating point operations [10], part of the MobilenetV3 network is added to the YOLOV4 backbone network, *i.e.*, a spatial attention module and a channel attention module are added to the backbone network. The MaxPooling layer in the backbone network directly integrates image information, which will cause part of the information to be lost. Therefore, the spatial attention module performs corresponding spatial transformations on the spatial domain information in the image to extract key information. The channel attention module adds a coefficient to the channel dimension of the output feature information, and changes the weight of the corresponding channel by adjusting the size of the coefficient, thereby changing the importance of the channel. Therefore, the channel attention module and the spatial attention module are added to the input image after a certain convolutional network, which can reduce the amount of parameters and floating point operations of the backbone network. The structure of the attention mechanism model is shown in **Figure 4**.

The middle part of **Figure 4** is the SENet network, which contains three parts, namely the squeeze function, the incentive function and the scale function. The squeeze function is a global pooling layer, which adds up all the eigenvalues in the channel to average. The excitation function obtains a coefficient between 0 - 1 for each channel through the sigmoid function, and adjusts the size of the coefficient through training. The scale function multiplies the values on different channels by the corresponding weights, which can strengthen the attention to the key channels (**Figure 5**).

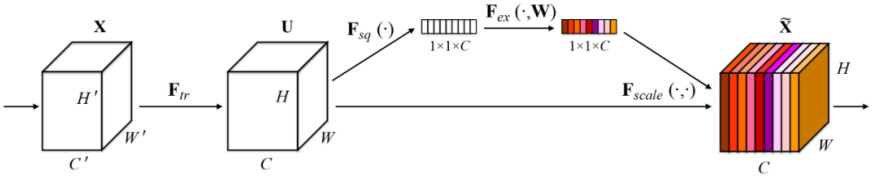

Figure 4. Attention mechanism.

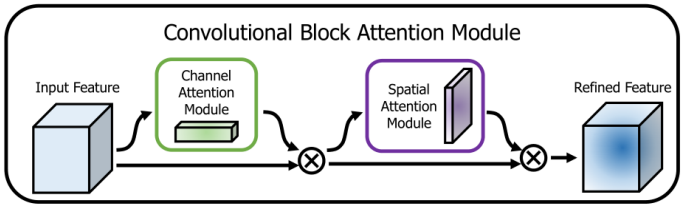

Figure 5. Attention mechanism model.

3.2. Rescurive-FPN

Since changing the backbone network in YOLOV4 to Mobilenetv3 will reduce the acquired image feature information, PANNet is converted to Rescurive-FPN in the Neck network, because PANNe is a simple two-way fusion of the feature

maps output by the backbone extraction network. Although the feature information on the image is retained to a certain extent, due to the garbage classification, there is rich and similar image information, such as kitchen garbage. Therefore, Rescurive-FPN is used to fuse the characteristic information output by traditional moral FPN and then input it to Backbone for a second cycle. In this way, the characteristic information of the backbone network is retained as much as possible (**Figure 6**).

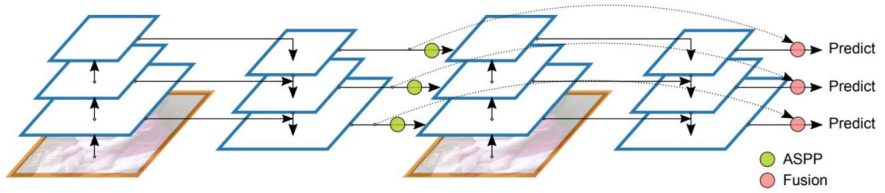

Figure 6. Rescurive-FPN.

4. Experiment and Analysis

4.1. Experiment Environment

In this training, a total of 22, 000 garbage images were collected through crawlers and actual collection of images, which were divided into 3 categories, namely hazardous garbage, kitchen garbage, and other garbage. There are 5 types of garbage in each category. The 15 types of garbage are batteries (1223), button batteries (1450), correction tapes (894), disposable cups (1500), glasses (1866), ice cream (1450), light bulbs (1168), medical bottles (1635), porridge (1422), Noodles (1976), rice (1670), gloves (1340), thermometer (1388), tomatoes (1233), toothbrush (1785). Use 15400 images as the training set and 6600 images as the test set, with a ratio of (7:3). The experiment is based on the Window environment, using NVIDIA GeForce GTX 1080Ti training set and test set. Set the number of iterations to 1200, the Batch size to 16, the weight to the V0C data set pre-training weight, the allocated GPU memory size to 22.0 G, and the LabelImg as the labeling software.

4.2. Result

Using 22,000 15 types of garbage objects for training and 6600 images for testing, it is found that when the number of iterations is between 0 and 200, the CIOU Loss decreases significantly, and the accuracy improves significantly. From 700 to 1200, its training performance gradually stabilized, and the MAP value eventually reached 64%. The improved YOLOV4 training loss value is shown in **Figure 7**. The trained model is used to detect the garbage category. The detection result is shown in **Figure 8**. From the figure, it can be seen that the model not only detects the garbage object and the confidence value, but also detects the garbage category.

Compare the improved YOLOV4 with YOLOV4 and YOLOV3, and the specific training parameters are shown in **Table 1**. It can be seen from the experimental results that the three models, YOLOV3, YOLOV4 and Improved YOLOV4,

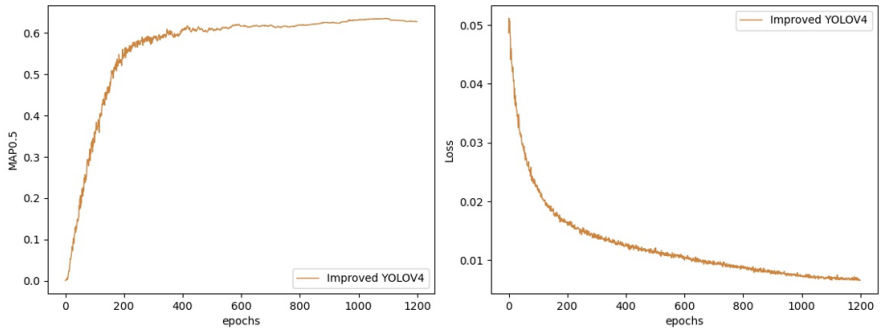

Figure 7. Improved YOLOV4 Loss and MAP.

Figure 8. Garbage category detection.

Table 1. Training result.

Models	Parameter	Model size	MAP	Ciou loss	FPS
YOLOV3	62.57 M	239 MB	53%	0.01481	72
YOLOV4	52.49 M	53.2 MB	68%	0.01121	84
Improved YOLOV4	38.26 M	32.2 MB	64%	8.0196e−3	92

can detect and classify garbage more accurately, but the parameters and model size of YOLOV3 are relatively large, the FPS is low, and embedded devices cannot be used. Since the number of VOC categories is much larger than that of garbage categories, the MAP value of training garbage classification using YOLOV4 has increased, with a MAP value of 68%. Compared with YOLOV4, the improved YOLOV4 has reduced the parameter amount by 14.23 M, the model size is 38.26 MB, the FPS is 92 fps, and the CIOU Loss can be reduced to 8.0196e−3.

5. Conclusion

Compared with the original version of YOLOV4, the accuracy of the improved YOLOV4 remains almost unchanged, at 64%, and the improved YOLOV4 has a higher FPS value than YOLOV4, which is 92 f/s. Therefore, using the improved

YOLOV4 can achieve more garbage category classification, and the model can be applied to embedded species to achieve real-time detection.

Acknowledgements

First of all, I would like to thank my paper supervisor, Professor Yang Lei from Dongguan University of Technology. Teacher Yang made instructive comments and recommendations on the research direction of my thesis, gave me careful advice on the difficulties and doubts I encountered in the process of writing the paper, and put forward many helpful suggestions for improvement. In addition, I would also like to thank my friends and classmates for their great support and help in writing the paper, which gave me great inspiration. I would also like to thank the authors in the references. Through their research articles, I have a good starting point for the research topic. Finally, thank you all for your hard work. I sincerely thank my family, friends, and classmates for their encouragement and support to successfully complete this paper.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Yue, X.M., Li, J., Hou, Y.X. and Lin, Z.C. (2020) Garbage Classification and Detection Method Based on Center Net. *Industrail Control Computer*, **33**, 78-79+82.

[2] Sang, S.G. (2020-11-23) "Internet + Door-to-Door Recycling of Waste" Helps Sorting Municipal Waste. Xijiang Daily.

[3] Chen, N.X., Yang, J.R., Dong, Y.H., Peng, X.F., Li, D.S. and Li, C.W. (2020) Innovative Design of Multifunctional Smart Trash Bin. *Green Packaging*, No. 11, 87-89.

[4] Yang, M.W., Cui, Y.Z., Zhang, Y.F., Tang, Y.W., Gao, Y.J. and Liu, H.Y. (2020) The Design and Abaqus Software Analysis of a New Type of Intelligent Multi-Function Automatic Sorting Bin. *Use and Maintenance of Agricultural Machinery*, No. 11, 25-26.

[5] Girshick, R. (2015) Fast R-CNN. *IEEE International Conference on Computer Vision*, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[6] Chen, H.J., Wang, Q.Q., Yang, G.W., *et al.* (2019) SSD Target Detection Algorithm Based on Multi-Scale Convolution Feature Fusion. *Computer Science and Exploration*, **13**, 1049-1061.

[7] Redmon, J., Divvala, S., Girshick, R., *et al.* (2016) You Only Look Once: Unified, Realtime Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779-788. https://doi.org/10.1109/CVPR.2016.91

[8] He, K.M., Zhang, X.Y., Ren, S.Q., *et al.* (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, 770-778. https://doi.org/10.1109/CVPR.2016.90

[9] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. *IEEE Conference on Computer Vision and Pattern Recognition*, 89-95.

[10] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.