Garbage detection and classification based on improved YOLOV4

Dongwei Guo¹, Lufei Cheng¹*, Meng Zhang², Yingying Sun¹

¹College of Computer Science and Technology, Jilin University, Changchun, Jilin, 130012, China
²College of Automotive Engineering, Jilin University, Changchun, Jilin, 130022, China
*Corresponding author’s e-mail: chenglf19@mails.jlu.edu.cn

Abstract. Object detection is an important research task in computer vision. With environment problems becoming seriously, garbage detection has become a hot direction in object detection. Though great progress has been made in garbage detection, there still exists challenges for general-purpose detectors with no public datasets and low recognition accuracy rate. Based on YOLOV4 model, this paper proposes an improved YOLOV4 detector for garbage detection. Specifically, CBAM is added to the feature extraction network in order to better extract features in deep networks, and focal function is integrated into the loss function to improve class imbalance. When preparing the data, environment information and random information are added to pictures to simulate the distribution of garbage in real environment, and finally an urban household garbage dataset containing 47 classes of 45,910 images named TrashSet has been produced. Experiments results on TrashSet verify that our detector has a considerable performance, and mAP reaches 97.15%.

1. Introduction
Object detection is a basic computer vision task and various object detectors have been devised such as YOLO[1], SSD[2] and Faster R-CNN[3]. On general datasets such as COCO[4] and VOC2007[5], these general detectors have achieved exciting results. Garbage detection and classification is an important part of object detection. However, the characteristics of garbage detection task are different from those of ordinary object detection tasks: complicated background, irregular shape and arbitrary orientations. And most general object detectors do not pay attention to the difference between the specific scenes and the standard datasets scenes, cannot achieve the same results as general datasets. Therefore, it is of great practical significance and application value to devise a special object detector for garbage detection.

In this paper, we discuss the applicability of the improved YOLOV4[6] object detector in the detection garbage. There are four main contributions, Firstly, there is no public dataset of urban household garbage in the object detection field at present. Therefore, one of the key points and difficulties in our experiments is to construct an urban domestic garbage dataset. Secondly, because of the complicated background, irregular shape and arbitrary orientations of objects to be detected, we add CBAM[7] to the backbone network, used to abstract effective features in deep network. Thirdly, in order to solve the problem of class imbalance, focal function[8] is added to the loss function devising. Lastly, through the contrast analysis of ablation experiment, it is proved that adding CBAM to backbone network and focal function to the loss function can improve the detection effect.
2. Materials and Methods

Garbage detection has been a hot but difficult research direction in object detection with the characteristics of complex background, irregular shape, and arbitrary orientations of the objects to be detected. Considering the rationality and the applicability of neural network models in urban garbage detection scene, the following part will mainly tell from the neural network structure devising and its improvement.

Based on YOLOV4 model, we propose an improved YOLOV4 detector consisting of three parts: feature extraction network, feature pyramid network and YOLO Head detector. We first give an overview of our improved YOLOV4 model in figure 1.

![Improved YOLOV4 neural network structure.](image1)

**Figure 1. Improved YOLOV4 neural network structure.**

2.1. Devise and improvement of feature extraction network

YOLOV4 object detection model adopts CSPDarknet53 network as feature extraction network. As shown in figure 2, the network is composed of a series of residuals structures. The remnant networks adopted in this part help to fuse different information of shallow and deep objects and better express object characteristics. CSPDarknet53 uses multiple feature layers for object prediction. It is located at the middle, middle and lower levels, representing different sizes of objects respectively. Therefore, the CSPDarknet53 network has a good detection effect for different size objects.

![Feature extraction network CSPDarknet53 structure.](image2)

**Figure 2. Feature extraction network CSPDarknet53 structure.**

Attention mechanisms are critical in human perception[9]. The purpose of adding an attention mechanism to the network is to amplify important features and suppress unwanted features so that the network can learn as we think. In the urban household garbage object detection task, there are many problems, such as complicated background, irregular shape and arbitrary orientations and these problems will seriously interfere with the correct classification and prediction of the detector. Therefore,
adding the attention mechanism to the feature extraction network of detectors can make the network learning pay attention to meaningful information and ignore unimportant features.

The attention model is divided into channel and spatial attentional models. The channel attentional models aim to evaluate the importance of different channels, strengthen important channels in the process of learning and training, ignore minor channels and better abstract the features of inputs. The spatial attention models are intended to mimic the characteristics of the human eyes, focusing on pixel locations where features are prominent in the images. In this paper, the attentional model CBAM is adopted, which focuses on both channel and space.

CBAM called Convolutional Block Attention Model, is a simple but effective attention mechanism applied to feedforward neural network. CBAM infers the features passed by the middle feature maps along the two dimensions of channel and space, and multiplies the received attentional maps with the input characteristic maps to obtain the adaptive optimization features. In this paper, CBAM is added to the residual structure of backbone to extract the features better during training. Figure 3 shows the structure of CBAM.

Figure 3. The structure of CBAM.

2.2. Feature pyramid network
SPP[10] and PANet[11] networks are used for the YOLOV4 component of the feature pyramid network. Feature pyramid can combine deep / shallow features with multi-resolution prediction to improve the recognition accuracy.

2.3. Object detector head
YOLOV4 detector head is designed and combined with three parts of network outputs, loss calculation and forecast result analysis. Three feature layers of feature pyramid inputs are processed, and anchors are generated, then the loss is calculated by coding. Finally, the prediction results are decoded.

2.4. Devise and improvement of loss function
This paper uses the Pytorch neural network framework to devise the improved YOLOV4 object detector. The loss function is designed in three parts: location loss, confidence loss, classification loss. The loss function is shown in equation (1).

\[
Loss = \lambda_{loc} \times Loss_{loc} + \lambda_{conf} \times Loss_{conf} + \lambda_{cls} \times Loss_{cls}
\]

We set \(\lambda_{loc} = \lambda_{conf} = \lambda_{cls} = 1.0\).

Location loss is calculated using CIOU Loss[12] as shown in equation (2).

\[
Loss_{loc} = Loss_{cIoU} = 1 - IOU + \frac{\rho^2 + \omega^2}{c^2} + \alpha \nu
\]

The IOU represents the ratio of the intersection and union of a prediction box and a real one. \(\rho\) represents the Euclidean Distance between the centre of prediction and the true one and \(c\) represents the diagonal distance that contains the minimum closed area of both the forecast and real ones. Besides, we set

\[
\alpha = \frac{\nu}{1 - IOU + \nu}, \quad \nu = \frac{4}{\pi^2} \left( \arctan \frac{\omega^u}{h^u} - \arctan \frac{\omega}{h} \right)^2.
\]
The confidence loss consists of two parts, one is the existence of a real object, and the prediction results are compared with those of 1.0. The other is that there are no objects existence in the picture, and the result of comparing 0 with the value of confidence in the prediction results. The devise of confidence loss is shown in equation (3):

$$\text{Loss}_{\text{conf}} = \text{BCELoss}(\text{conf}, \text{mask}) \times \text{mask} + \text{BCELoss}(\text{conf}, \text{noobj\_mask}) \times \text{noobj\_mask}$$  \hspace{1cm} (3)

The classification loss is the result of comparing the prediction class with the real class against the actual object boxes, as is shown in equation (4).

$$\text{Loss}_{\text{cls}} = \text{BCELoss}(\text{pred\_cls}, \text{smooth\_labels})$$  \hspace{1cm} (4)

Where BCELoss represent the cross-entropy loss function, as shown in equation (5):

$$\text{BCELoss} = -y \times \log(y\prime) - (1-y) \times \log(1-y\prime)$$  \hspace{1cm} (5)

Of these, $y$ stands for real results and $y\prime$ stands for prediction results.

In this paper, the focal function is used to deal with the problem of class imbalance in the sample. Focal function reduces the weight of a simple negative sample in training and is an improved version of the cross-entropy loss function. In this paper, in the devise of the YOLOV4 loss function, location loss and confidence loss function are replaced by focal loss function, as shown in equation (6).

$$\text{FocalLoss} = \text{BCELoss} \times \alpha \times \gamma \times |y - \text{sigmoid}(y\prime)|^\gamma$$  \hspace{1cm} (6)

$\alpha$ is a class equilibrium factor and $\gamma$ is a background loss factor. The purpose of setting $\gamma > 1$ is to reduce the loss values of the loss function, such as complex background, and to make neural networks easier to focus on the hard parts to learn. Adjusting the class equilibrium factor $\alpha$ to adjust for class imbalance in the sample. After several experiments, the best results are obtained at $\alpha = 0.25, \gamma = 2$.

3. Results & Discussion

Trainings are implemented by Pytorch on a server with Nvidia Geforce RTX 2080Ti and 8G memory. We implement experiments on the self-made dataset named TrashSet to verify the applicability of our design scheme. Finally, we analyse and discuss the experimental results. In experiments, the statistical score thresholds of precision rate and recall rate are set to 0.5, and mPrecision and mRecall represent the average value of precision rate and recall rate in categories. Note that the duration test platform is an Alibaba Cloud lightweight application server, equipped with a 1-core CPU and 2G memory.

3.1. Experimental design

3.1.1. Urban household garbage dataset. Because there has been no publicly available dataset for the urban garbage study, the garbage images used in this study are captured by web crawlers and taken by SLR cameras. The data from web crawlers account for 80 percent of the original images, while the remaining 20 percent are taken by SLR cameras. The original trash images include 47 categories totalling 36,571 images. After simple screening and merging of these images, a total of 47 classes and 23,093 garbage pictures are obtained. The categories of garbage are shown in figure 4.

We process the garbage pictures obtained: set the length and width of each picture to a random number between 416 and 608; add 10 scenes and 120 backgrounds such as desktop, floor tiles, wallpaper, grass, land, and solid colours to each picture. Some small pictures are added to the background of the picture, and a random distance is used between the pictures. The pictures are also rotated at a random angle to simulate the placement of garbage in the life scene, and some pictures are added to the mosaic. After the garbage dataset is established, the open source tool Labelme is used to check the quality and label accuracy of each picture and mark it. The urban household garbage dataset named TrashSet, containing the training set of 45,910 pictures and the test set of 6,887 pictures is formed, and finally
the dataset is converted into the VOC format. Urban household garbage dataset TrashSet is shown in figure 5.

3.1.2. Evaluation indicators. We use precision rate, recall rate, mean average precision (mAP) and time complexity as the evaluation indicators of the experimental results. Precision rate indicates how many of the objects predicted to be positive are true positive objects. Recall rate indicates how many positive objects in the sample are predicted correctly. MAP can comprehensively evaluate the precision and recall of experimental results. In addition, time complexity is also one of the important evaluation indicators of this experiment.

3.1.3. Experiment methods. In the training phase of the improved YOLOV4 model, the cosine annealing decay method is used in learning rate. The learning rate will first rise and then fall. It rises linearly when rises, and when it falls, the simulated cosine function decreases. This process will be executed multiple times. The initial learning rate is set to 0.001, and the epochs is set to 100,000. The loss function declining curve is shown in figure 6, and the learning rate changeable curve is shown in figure 7.

3.2. Experiment results
In order to verify the applicability of our detector, we perform many ablation experiments for the improved part of this paper. The ablation experiments all use the same training dataset and test dataset, the same training batch and learning rate. The size of the pictures is adjusted to $418 \times 418$ before training, and the results of the ablation experiment are shown in figure 8.
Table 1 shows the precision, recall and AP of the improved YOLOV4 model in TrashSet dataset.

Table 1. Improved YOLOV4 network performance.

| Class                      | Precision (%) | Recall (%) | AP (%) | Class                      | Precision (%) | Recall (%) | AP (%) |
|----------------------------|---------------|------------|--------|----------------------------|---------------|------------|--------|
| Package                    | 94.34         | 97.28      | 98.34  | Newspaper and flyer        | 97.93         | 97.43      | 98.60  |
| Shells                     | 95.92         | 98.75      | 99.18  | Vegetable leaves           | 98.32         | 97.56      | 98.64  |
| Glass products             | 94.23         | 97.27      | 98.46  | Plugs and wires            | 92.90         | 96.10      | 97.88  |
| Portable chargers          | 94.47         | 97.44      | 98.43  | Flowerpots and bowls       | 94.39         | 95.04      | 96.11  |
| Bones                      | 90.61         | 96.61      | 97.05  | Cosmetic bottles           | 90.57         | 93.77      | 96.12  |
| eggshells                  | 92.99         | 96.70      | 97.58  | Metal food cans            | 94.27         | 94.66      | 96.66  |
| Waste lamp tubes and bulbs | 93.56         | 93.56      | 93.83  | The wire bottles           | 93.97         | 95.41      | 97.44  |
| Batteries                  | 95.73         | 96.70      | 97.87  | Used clothing              | 95.32         | 96.37      | 97.49  |
| Pots                       | 94.87         | 97.55      | 98.14  | Tetra Pak                  | 96.10         | 96.00      | 97.57  |
| expired drugs              | 90.60         | 90.60      | 90.87  | Plush toys                 | 95.53         | 96.19      | 97.89  |
| Pesticide container        | 94.17         | 90.22      | 93.10  | Diapers                    | 90.51         | 88.83      | 91.25  |
| Foam plastics              | 94.61         | 94.77      | 95.68  | Plastic bottles            | 93.43         | 96.12      | 97.75  |
| Leather shoes              | 93.45         | 97.31      | 98.11  | Plastic toys               | 97.25         | 96.26      | 97.94  |
| Ointment                   | 96.41         | 96.95      | 98.16  | Plastic bowls              | 97.80         | 98.59      | 99.26  |
| Leftover                   | 93.76         | 89.23      | 92.51  | Toilet paper               | 91.03         | 93.12      | 92.90  |
| Books                      | 95.60         | 96.66      | 98.00  | Stain plastic              | 96.57         | 97.01      | 98.24  |
| Fruit peel                 | 95.38         | 95.25      | 97.10  | Western food cakes         | 95.38         | 94.36      | 95.98  |
| Fruit pulp                 | 96.55         | 97.39      | 98.51  | Toothpicks                 | 93.14         | 96.14      | 97.84  |
| Mercury thermometers       | 93.93         | 97.10      | 98.50  | Cigarette ends             | 95.90         | 97.97      | 98.53  |
| Disposable tableware       | 96.78         | 97.94      | 98.54  | Cartons                    | 96.62         | 98.83      | 99.27  |
| Coat hangers               | 96.34         | 96.50      | 97.52  | Paper cups                 | 94.27         | 97.26      | 98.20  |
| Cans                       | 95.39         | 97.06      | 98.08  | Chopsticks                 | 96.84         | 97.42      | 98.36  |
| Fishbone                   | 93.60         | 95.48      | 96.33  | Pillows                    | 95.14         | 95.79      | 97.36  |
| Chopping blocks            | 96.81         | 97.74      | 98.97  |                            |               |            |        |
3.3. Discussion

Through the ablation experiments, we can draw the following points:

- **Effect of CBAM.** It can be convinced in figure 8 that the evaluation indicators have been improved to varying degrees after adding CBAM. We find mPrecision increasing by 0.14%, mRecall increasing by 0.87%, mAP increasing by 0.31%, and the time detection cost increasing by 0.8s. As discussed above, CBAM is effective to suppress the noise and highlight the information of objects both from channels and spaces.

- **Effect of focal loss.** As seen in figure 8 that the effect of the whole experiments have been greatly improved. We find mPrecision increasing by 2.62%, mRecall increasing by 9.73%, mAP increasing by 5.20%. At the same time, time cost has not changed compared to before the focal function is added. After analysis, we find the following two reasons: Firstly, the dataset TrashSet has class imbalance problems, such as 1,251 newspaper flyer, 1,674 vegetables and tea, 98 leftovers, 74 diapers and 31 expired drugs in the original garbage pictures. The focal function has a class equilibrium factor $\alpha$, which makes the network give higher weight to the category with fewer original images, and can effectively make up for the effect caused by the large difference in the number of samples in different categories. Secondly, in the characteristic diagram of one-stage object detection algorithm, there are a large number of background boxes which are unrelated to the objects, i.e. negative samples, and the samples containing the targets are only in a small part of the region. The unbalanced positive and negative sample ratio will greatly affect the convergence of loss function. The negative and positive background loss function factor $\gamma$ is set up in focal loss function, and we set $\gamma > 1$, which greatly reduces the impact of negative samples such as background frames on the loss functions, so that the network can be concentrated into the difficult samples.

- **Effect of CBAM and focal loss.** Figure 8 shows that the experiments perform best when we add CBAM and focal function: mAP reaches 97.15%. But, compared with before adding CBAM, the time cost increases by 0.8s. From table 1, we can see that the precision rate, recall rate and mAP of most classes of garbage have reached more than 95%, but the recognition accuracy of similar-shaped garbage such as pesticide containers and plastic bottles is still not high. It is also the direction we will continue to study.

4. Conclusion

We have presented an improved YOLOV4 detector and a dataset for urban household garbage detection. We study the characteristics of complicated background, irregular shape and arbitrary orientations of the objects to be detected. Besides, we propose an improved YOLOV4 detector model, which makes adaptive improvements to the urban household garbage scene. By constructing TrashSet, a dataset of urban household garbage containing 47 categories, adjusting the feature layer, adding CBAM to the feature extraction network, adjusting anchor boxes through the k-means algorithm, and adding the focal function to the loss function, strengthens the ability of our detector to detect garbage. Experimental results show that the improved YOLOV4 object detector can meet the needs of real-time detection; the precision rate, recall rate and mAP of most classes of garbage have reached more than 95%, but the recognition accuracy of similar-shaped garbage such as pesticide containers and plastic bottles is still not high. It is also the direction we will continue to study.

References

[1] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, (2016) You Only Look Once: Unified, Real-Time Object Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle. pp. 779-788.
[2] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., & Fu, C. Y., et al. (2016). SSD: Single Shot MultiBox Detector. In: European Conference on Computer Vision. Amsterdam. pp. 779-788.

[3] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(6): 1137-1149.

[4] Lin, T. Y., Maire, M., Belongie, S., Hays, J., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In: European Conference on Computer Vision. Zurich. pp. 740–755.

[5] Everingham, M., Eslami, S., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: a retrospective. International Journal of Computer Vision, 111(1): 98-136.

[6] Bochkovskiy, A., Wang, C. Y., & Liao, H. (2020). Yolov4: optimal speed and accuracy of object detection. https://arxiv.org/abs/2004.10934.

[7] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: convolutional block attention module. In: European Conference on Computer Vision. Munich. pp(11211). 3-19.

[8] Lin, T. Y., Goyal, P., Girshick, R., He, K., & P Dollár. (2017). Focal loss for dense object detection. 16th IEEE International Conference on Computer Vision (ICCV). Venice. pp. 2999-3007.

[9] Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews Neuroscience, 3(3), 201-215.

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(9), 1904-1916.

[11] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City. pp. 8759-8768.

[12] Zheng, Z., Wang, P., Liu, W., J Li, & Ren, D. (2020). Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In: AAAI Conference on Artificial Intelligence. New York.