


Article

YOLOv5-OCDS: An Improved Garbage Detection Model Based on YOLOv5

Qiuhong Sun ^{1,†}, Xiaotian Zhang ^{1,†}, Yujia Li ¹ and Jingyang Wang ^{1,2,*} ¹ Hebei University of Science and Technology, Shijiazhuang 050018, China² Hebei Intelligent Internet of Things Technology Innovation Center, Shijiazhuang 050018, China

* Correspondence: jingyangw@hebust.edu.cn

† These authors contributed equally to this work.

Abstract: As the global population grows and urbanization accelerates, the garbage that is generated continues to increase. This waste causes serious pollution to the ecological environment, affecting the stability of the global environmental balance. Garbage detection technology can quickly and accurately identify, classify, and locate many kinds of garbage to realize the automatic disposal and efficient recycling of waste, and it can also promote the development of a circular economy. However, the existing garbage detection technology has some problems, such as low precision and a poor detection effect in complex environments. Although YOLOv5 has achieved good results in garbage detection, the detection results cannot meet the requirements in complex scenarios, so this paper proposes a garbage detection model, YOLOv5-OCDS, based on an improved YOLOv5. Replacing the partial convolution in the neck with Omni-Dimensional Dynamic Convolution (ODConv) improves the expressiveness of the model. The C3DCN structure is constructed, and parts of the C3 structures in the neck are replaced by C3DCN structures, allowing the model to better adapt to object deformation and target scale change. The decoupled head is used for classification and regression tasks so that the model can learn each class's characteristics and positioning information more intently, and flexibility and extensibility can be improved. The Soft Non-Maximum Suppression (Soft NMS) algorithm can better retain the target's information and effectively avoid the problem of repeated detection. The self-built garbage classification dataset is used for related experiments, and the mAP@50 of the YOLOv5-OCDS model is 5.3% higher than that of the YOLOv5s; the value of mAP@50:95 increases by 12.3%. In the experimental environment of this study, the model's Frames Per Second (FPS) was 61.7 f/s. In practical applications, when we use some old GPU, such as the GTX1060, it can still reach 50.3 f/s, so that real-time detection can be achieved. Thus, the improved model suits garbage detection tasks in complex environments.

Keywords: YOLOv5; ODConv; C3DCN; Soft-NMS; decoupled head

Citation: Sun, Q.; Zhang, X.; Li, Y.; Wang, J. YOLOv5-OCDS: An Improved Garbage Detection Model Based on YOLOv5. *Electronics* **2023**, *12*, 3403. <https://doi.org/10.3390/electronics12163403>

Academic Editor: Chiman Kwan

Received: 30 June 2023

Revised: 31 July 2023

Accepted: 8 August 2023

Published: 10 August 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Garbage refers to discarded materials that have no reuse value or are no longer needed. This paper's garbage research object is domestic garbage, which refers to the waste generated in daily life, such as tea residue, daily paper, defiled plastic, glassware, metal, etc. With the continuous improvement of industrialization and marketization, people's lifestyles and entertainment are diversifying. People's consumption level is becoming higher and higher, resulting in more and more garbage output. The traditional artificial garbage sorting technology has a series of problems, such as a low sorting efficiency, long consumption time, and large consumption of human and material resources, and the sorting environmental sanitation conditions are not up to standard, among other issues. Therefore, how to detect and classify garbage targets reasonably has become a practical problem that needs to be solved urgently. At present, deep learning technology has developed rapidly, especially in the fields of target detection [1] and image classification. Suppose that the relevant

advanced technologies in deep learning can be effectively used to solve the problem of garbage classification. In that case, the utilization of human resources will be greatly improved, effectively improving the efficiency of garbage sorting and contributing to the protection of the ecological environment.

The current object detection models are divided into two categories: traditional feature extraction methods and deep learning methods. The traditional method consists of image feature SIFT [2], HOG [3], SVM [4], and other classifiers. The traditional method has a good detection effect for obvious objects and simple backgrounds. However, in the face of complex situations such as irregular shape, large size variation range, serious occlusion, and various types of garbage, the traditional target detection model has low performance and poor robustness for the detection of interest targets and cannot meet the real-time requirements due to excessive human intervention in extracting image features.

In recent years, with the rapid development of deep learning, scholars from various countries have also carried out extensive research in the field of garbage detection. Zeng et al. [5] proposed a Multi-Scale Convolutional Neural Network (MSCNN) to classify hue–intensity–saturation (HIS) data pixels and generate binarized garbage segmentation maps for the hyperspectral image garbage detection problem which has a good performance in large-area garbage detection. Ma et al. [6] proposed an improved Faster R–CNN [7] garbage target detection model, and the experimental results showed that the average accuracy was improved by 8.26% compared with the traditional Faster R–CNN algorithm. Mikami et al. [8] used the SSD algorithm to detect garbage bags, and the average accuracy of garbage bag recognition was 62%. Liu et al. [9] proposed a garbage classification method based on the YOLOv2 [10] model that serves as a lightweight version of the YOLOv2 model. Still, the deficiencies are that the annotated garbage categories are not subdivided and could not distinguish the garbage types. Xu et al. [11] proposed a lightweight garbage target detection algorithm based on the YOLOv3 [12] algorithm, which can effectively detect garbage targets. Zhang et al. [13] proposed a YOLO-WASTE model based on the YOLOv4 network, using transfer learning for training, and achieved good results on self-built datasets.

The garbage detection model based on deep learning technology has significantly improved accuracy, speed, and robustness. However, most of the garbage image datasets used in the current research are single-target and few-target data or lack rich garbage categories, making them insufficient to meet the actual needs of different types and quantities of garbage piled up in real life. For the partially obscured, changeable shapes and various garbage targets in life, the existing models have some problems, such as false detection, missed detection, and inaccurate positioning frame, which affect the detection effect.

Aiming at making up for the shortcomings of the existing technical solutions, this study constructed a domestic garbage dataset including 38 categories with 17,057 images and proposed a domestic garbage detection model based on YOLOv5-OCDS to meet the needs of practical engineering applications.

The main application scenarios of the model are in places containing domestic garbage, such as highways, streets, parks, etc. Compared with the original YOLOv5 model, the detection effect of this model is better when dealing with complex situations such as objects being displayed incompletely or small targets. After determining the usage scenario of garbage, how to implement garbage detection is also crucial. Firstly, data collection is required. It is necessary to collect image datasets containing different types of waste. Each image must be annotated with corresponding bounding boxes and category information. Then, a model is built by selecting a suitable deep learning model, such as YOLOv5, and configuring it according to the number of categories and image size of the dataset. Then, model training is performed, and model parameters are optimized. Finally, a model evaluation is carried out. Test sets are used to evaluate the trained model, and the performance of the model is assessed through indices such as mAP. When applying this model in real life, it is necessary to export the trained weight file first and then load the exported model file into the memory of the computing device for reasoning. The camera takes pictures containing garbage, and the computing device uses the loaded model to reason the image

to be detected, that is, to identify the target object in the image and its corresponding position box. Finally, the detection results are visualized. For example, the model can be deployed in scenic spots, and images can be obtained in real time on cameras or monitoring equipment in scenic spots. The images can be input into the garbage detection model for analysis and prediction. According to the model's output results, the location and quantity of garbage in the scenic spot can be marked, thus helping health workers optimize the garbage cleaning plan and resource allocation.

There are four main contributions of this paper:

- (1) Turned part of the convolution in the neck of YOLOv5 into ODConv, which can dynamically convolve features at different scales to capture the feature representation of the target at different scales.
- (2) Built the C3DCN module and used it to replace part of the C3 module in the neck of YOLOv5. The C3DCN module is a module with channel attention and deformable convolution, which can enhance the model's receptive field and representation power. By replacing the C3 module with C3DCN, the feature extraction capability of the model for the target in the garbage classification dataset will be enhanced, and the model's accuracy will be improved.
- (3) Replaced the coupled head of the original YOLOv5 with a decoupled head. A decoupled head is an improved detection head structure that deals with the category prediction and position prediction of targets separately by decoupling classification and regression tasks. It can also better handle the differences between categories and reduce the mutual interference between categories. It makes the model more focused on learning the features and location information of each category and improves the robustness of target detection.
- (4) Replaced the NMS algorithm of the original YOLOv5 with the Soft-NMS algorithm. The number of overlapping boxes can be reduced during object detection processing, which helps to remove redundant detection results and improve the accuracy of object detection. When dealing with small targets, the traditional NMS algorithm tends to remove small targets in overlapping boxes, while Soft-NMS can retain these small targets and improve the detection ability of small targets.

2. Related Work

In order to realize the reduction, reutilization, and harmless treatment of domestic waste, it is the general trend to promote sorting automation based on artificial intelligence. The premise of realizing sorting automation is to determine the location and identify the types of garbage, so garbage detection technology is very important.

Currently, the traditional garbage detection methods based on machine learning and image processing are mainly divided into three parts: region selection, feature extraction, and image classification [14]. The garbage detection and sorting device proposed by Salimi et al. [15] used the Haar cascade method to detect garbage on the ground for the first time and then combined the grayscale co-occurrence matrix and directional gradient histogram for a texture and shape analysis to obtain a set of features, which were input into Support Vector Machine (SVM) for classification. The accuracy rate can reach 73.49%. This kind of detection method needs to extract features manually and can obtain better detection results by adopting specific methods under specific conditions. Hu et al. [16] proposed a method of the infrared spectrum combined with machine learning, which realized the deep sorting of high-value utilization of domestic garbage.

In recent years, many scholars have applied deep learning to garbage detection and classification research. Ma et al. [17] proposed the enhanced SSD [18] algorithm L-SSD and used ResNet-101 [19] as the backbone structure. Compared to the SSD algorithm, it added a lightweight and efficient feature fusion module, strengthening the network detection performance. Feng et al. [20] proposed a model to improve the performance of Mask R-CNN [21] by using MobileNet as the backbone structure. The model had a small number of parameters and thus could be applied to embedded devices. Liu et al. [22]

proposed a cross-channel interactive attention mechanism ECA by improving the channel attention mechanism SENet [23] module and introduced it into the residual unit of YOLOv3. Compared with the original YOLOv3 model, the mAP@50 and mAP@50:95 of the improved method increased by 0.72% and 1.07%, respectively.

Pan [24] proposed a target detection model YOLOv3++ for garbage classification. By optimizing the backbone network and using transfer learning, the computational load and parameter number were reduced, and the efficient detection of 10 kinds of recyclable garbage was realized. Li et al. [25] proposed YOLOv3-2SMA, which removed the largest target detection layer and generated the new anchor boxes, improving the detection speed and making the accuracy as high as before. Iqbal et al. [26] used YOLOv4 with CSPDarkNet_tiny as the backbone, trained it with images collected by themselves, and successfully deployed it to TX2. Wang et al. [27] improved the YOLOv4 model by using the lightweight EfficientNet backbone networks and the deep separable convolutional layers, reducing the number of parameters in the network model. Patel et al. [28] compared classical models such as Faster R-CNN and YOLOv5m, optimized the parameters of YOLOv5m, and finally obtained the highest precision value of YOLOv5m. Yan et al. [29] proposed a garbage classification detection method that integrates the pruning strategy into the YOLOv5 model. This approach had somewhat alleviated the issues of time consumption and laborious garbage classification, but the number of pictures in their experimental dataset was small. In addition, the above studies all used self-built datasets, and the picture background was relatively simple. Lin et al. [30] proposed a Soft YOLOX model to detect the dense garbage in the manhole cover. By using the Soft-NMS algorithm to punish the detection box score, the missing detection phenomenon was avoided. However, the dataset used in this work is a specific garbage target in a particular scenario, so the categories are fewer. However, in real life, there are many kinds of garbage, the garbage environment is complex, and there are many small target pieces of garbage and garbage that is blocked, so this paper studies multi-target garbage detection under the complex environment.

3. Methodology

3.1. YOLOv5

The network structure of YOLOv5 series models is composed of three parts: feature extraction backbone, feature fusion neck, and detection head. The YOLOv5 model is divided into YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x successively from small to large. The model's depth and the number of output channels at each layer are controlled by the difference of *depth_multiple* and *width_multiple* parameters in the program. This paper uses YOLOv5s as the basic model, and the YOLOv5s is shown in Figure 1. The backbone consists of the CBS, C3, and SPPF modules. The CBS module first carries out two-dimensional convolution (Conv2d), conducts normalization processing of the BN layer secondly, and then passes the results to the SiLu part of the activation function, enhancing the feature transfer and information flow. The C3 module is composed of a series of convolutional layers and channel connection operations designed to improve the model's receptive field and strengthen the semantic expression ability of features. SPPF is a pooling operation of the input features at multiple scales separately, and then the pooling results at these scales are spliced together. In this way, the context information of different scales can be captured, and the perception ability of the model for different scales can be enhanced. The neck structure of YOLOv5 is used for feature extraction. It is located between the backbone network and the detection head. The neck structure is implemented by the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN). FPN enhances the perceptual capability of the model through multi-scale feature fusion. PAN fuses feature between the upper and lower feature maps to promote the transmission and integration of cross-scale information, helps the model better capture the features of different scale targets, and improves the detection performance. The head is used to detect the location and category of objects.

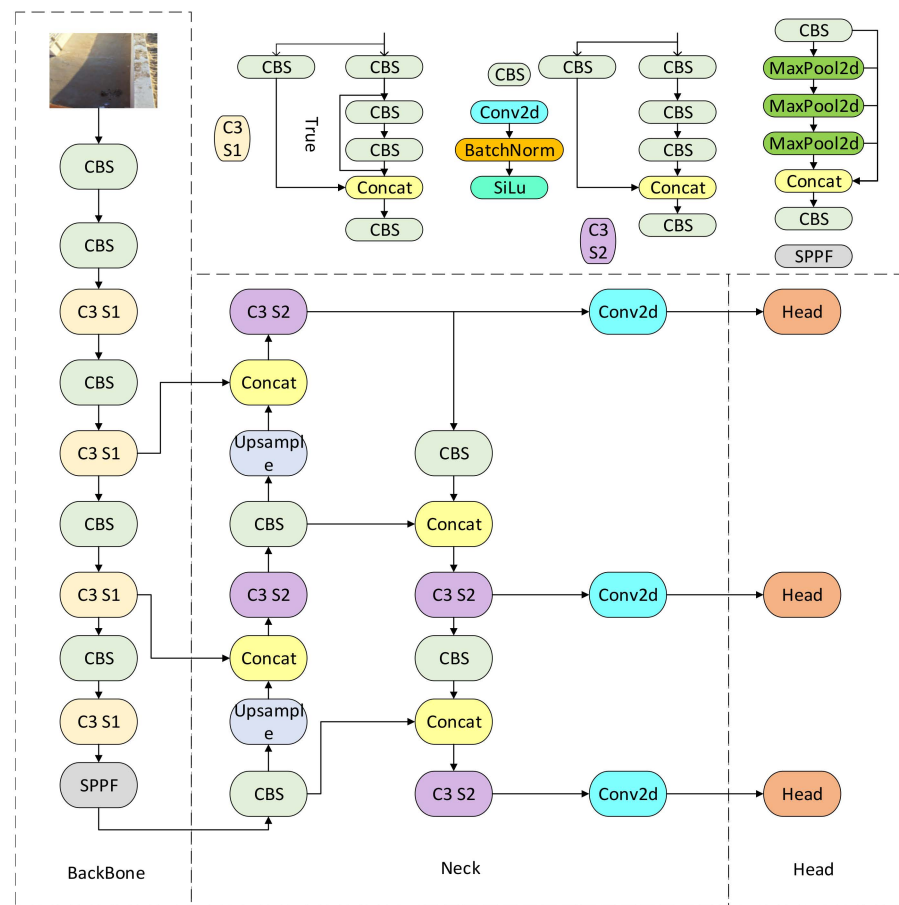


Figure 1. YOLOv5 structure diagram.

3.2. ODConv

The convolution kernel of ordinary convolutional neural networks is static. The existing dynamic convolution realizes the attention weighting of the convolution to the input data through the linear combination of the convolution kernel weights, which can significantly improve the accuracy while maintaining high-speed inference. The person who proposed ODConv [31] believes that the existing dynamic convolutions (CondConv and DyConv) only pay attention to the dynamics of the number of conv-kernels while ignoring the spatial, input-channel, and output-channel dynamics.

The conventional convolution layer has a static convolution kernel, which is suitable for all input samples. The dynamic convolution layer is different from the conventional convolution layer. It uses the linear combination of n convolution kernels and the attention mechanism for dynamic weighting, making the convolution operation dependent on the input. The dynamic convolution operations can be defined as follows:

$$y = (\alpha_{w1}W_1 + \dots + \alpha_{wn}W_n) * x \tag{1}$$

where x is the input data expressed in (h, w, c_{in}) format; y is the output data expressed in (h, w, c_{out}) format; W_i denotes the i -th convolution kernel (where the data format of W_i is $W_i^m \in \mathbb{R}^{k \times k \times c_{in}, m = 1, \dots, c_{out}}$); and $\alpha_{wi} \in \mathbb{R}$ is the attention scalar weighted to W_i , calculated by the attention function from the input data, with $*$ representing the convolution operation.

ODConv’s definition of dynamic convolution is shown in Formula (2), where α_{wi} represents the attention to the convolution kernel, α_{si} represents the attention to the $k \times k$ convolution kernel space, α_{ci} represents the attention to the input channel, and α_{fi} represents the attention to the output channel. Moreover, \odot stands for multiplication along

different dimensions of kernel space. Here, the implementation of each attention is slightly different. As shown in Figure 2, multiplying different attentions along dimensions such as space, channel, filter, and kernel will obtain a better performance to capture rich contextual information. Therefore, ODConv can capture the structure and features in the image better. More importantly, ODConv, with fewer convolution kernels, can attain a comparable or even superior performance compared to DyConv. In Figure 2, * stands for convolution operation.

$$y = \left(\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \dots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n \right) * x \quad (2)$$

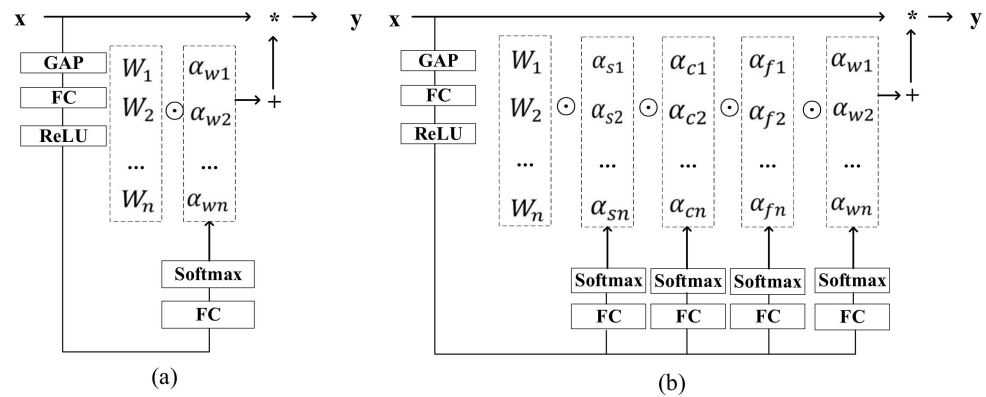


Figure 2. (a) DyConv and (b) ODConv.

3.3. C3DCN Module

Deformable convolution v2 [32] is an improved convolution operation aimed at improving the modeling ability of the convolutional neural network for target deformation and pose variations. Deformable convolution v2 is an extension of the traditional fixed convolution kernel. By introducing learnable migration parameters to adjust the position of convolution sampling points, it can adapt to the targets' spatial deformation and attitude-pose changes.

Deformable convolution can learn the features of the deformed objects well, so that the region of interest can be modeled more accurately. Moreover, it can adapt to object deformation better than ordinary convolutional networks. Compared with the traditional fixed convolutional kernel, the deformable convolutional network has stronger expressibility and flexibility.

It can adjust the received input feature's position and the input features' amplitude (importance). Through deformable offset and sampling, it can adjust the position and deform the input features according to the learned parameters. By adjusting the deformable offset's value, the input feature's position can be fine-tuned to align the target object or feature of interest more precisely.

Given a convolution kernel with K sampling positions, w_k and p_k represent the weight and preset offset of the k_{th} location, respectively; for example, when $K = 9$ and $p_k \in \{(-1, -1), (-1, 0), \dots, (1, 1)\}$ represents a 3×3 convolution kernel with a dilation rate of 1, let $x(p)$ and $y(p)$ denote the features at position, p , in the input feature map, x , and output feature graph, y , respectively. The deformable convolution can be defined as follows:

$$y(p) = \sum_{k=1}^K w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \quad (3)$$

where p is the real pixel coordinate, Δp_k and Δm_k are learnable offset and regulation parameters at the k_{th} position, the regulation parameters are $\Delta m_k \in [0, 1]$, and Δp_k is any value. Compared with the previous version of deformable convolution, the offset and the weights of the sampling points are learned.

The adjustable design of the RoI pooling layer is also similar. Given the region of interest of the input, RoI pooling divides it into K spatial cells, such as 7×7 . Within each cell, the size of the sampling kernel can be set, such as 2×2 , to obtain the output of each cell. Similarly, let Δp_k and Δm_k be learnable offsets and weights at the k_{th} position. The output features can be calculated by using Formula (4).

$$y(k) = \sum_{j=1}^{n_k} x(p_{kj} + \Delta p_k) \cdot \Delta m_k / n_k \tag{4}$$

where p_{kj} denotes the j_{th} cell in the k_{th} sampling region, and n_k denotes the number of cells in the sampling region.

3.4. Decoupled Head

The design idea of coupled head is to couple the classification of the target and the regression task together, making both classification and regression prediction by sharing features. The classification and regression branches in the coupled detection head are interdependent, sharing the underlying feature representation and making predictions through a joint network structure. This design can reduce the computation of the network by sharing features and make better use of the relationship between classification and regression tasks. Common designs of the coupled head include the use of multiple fully connected layers or convolutional layers for classification and regression prediction at the same time. Although the coupled head is more compact and can achieve efficient detection with limited computing resources, it is not flexible enough for different types of targets and task requirements.

The decoupled head [33] breaks down the object detection task into two independent subtasks: classification and regression. As shown in Figure 3, in the decoupled head, after a 1×1 convolution, independent classification and regression tasks are carried out, which are responsible for predicting the category and location information of the target, respectively. This design makes the model learn the target’s classification and location features more flexibly to improve the detection performance. The decoupled head generally has more flexibility and can better adapt to different target features and task requirements.

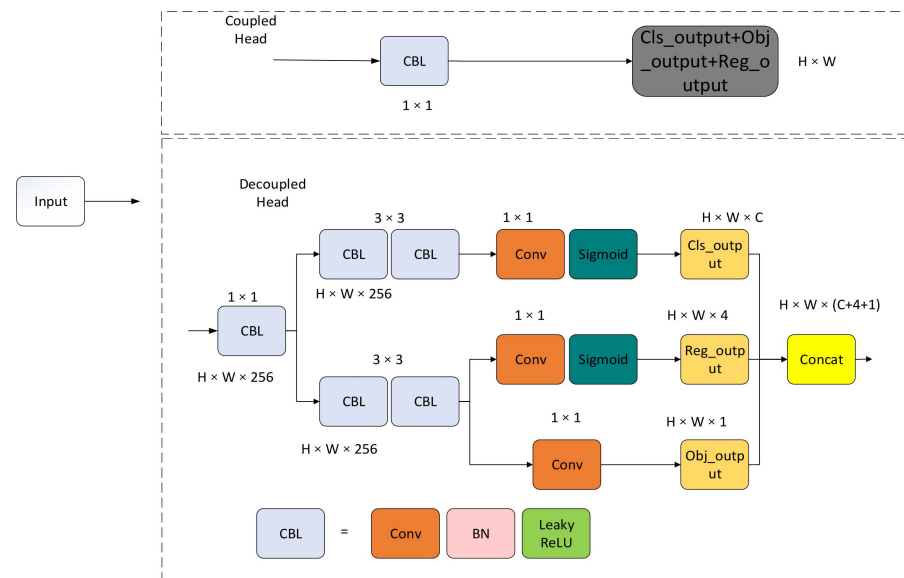


Figure 3. Comparison of coupled head and decoupled head.

3.5. Soft-NMS

In this paper, Soft-NMS [34] is selected as the postprocessing algorithm of the network model. NMS is a common technique used to suppress redundant bounding boxes in target

detection tasks, thereby improving the accuracy and stability of detection results. NMS judges the similarity between bounding boxes by intersection over union (IoU) and selects the bounding box with the highest score as the final detection result. In YOLOv5, the traditional NMS algorithm is used for postprocessing by default, which can effectively suppress bounding boxes with large overlaps, thus improving the recall rate and accuracy of detection results. The rescore function of the traditional NMS algorithm is shown in Formula (5).

$$s_i = \begin{cases} s_i, iou(M, b_i) < N_t \\ 0, iou(M, b_i) \geq N_t \end{cases} \quad (5)$$

In Formula (5), s_i represents the score or confidence between the predicted bounding box, M , and the real bounding box, b_i ; iou represents the intersection over union between the predicted bounding box, M , and the real bounding box, b_i ; M represents the predicted bounding box; b_i represents the real bounding box; and N_t represents the threshold (usually less than 0.5).

Soft-NMS is an improved algorithm of NMS, which retains some bounding boxes with larger overlap to some extent by reducing the score of the bounding box when calculating the overlap. Soft-NMS aims to solve the bounding-box missing problem, while traditional NMS deals with highly overlapping objects. Soft-NMS usually shows better results in scenes with dense targets or highly overlapping targets and can improve recall rates of test results. The rescore function of the Soft-NMS algorithm is shown in Formula (6).

$$s_i = \begin{cases} s_i & , iou(M, b_i) < N_t \\ s_i(1 - iou(M, b_i)) & , iou(M, b_i) \geq N_t \end{cases} \quad (6)$$

$$s_i = s_i e^{-\frac{iou(M, b_i)^2}{\sigma}} \quad (7)$$

In Formula (7), σ represents the variance of the Gaussian function.

3.6. YOLOv5-OCDS

The YOLOv5-OCDS model in this paper uses ODConv to replace a part of ordinary convolution in the neck and introduces adaptive adjustment to perceptual field and shape in convolution operation, which makes the convolution operation more flexible and adaptable and increases adaptability to small target objects and objects with large shape changes. The deformable convolution and C3 are fused to form a new structure, the C3DCN structure, which replaces a part of the C3 structure in the neck. Because the garbage classification dataset usually contains objects of various shapes and sizes, this structure can help the model better capture and identify these objects. Compared with the traditional convolution operation, the structure introduces a deformable convolution kernel, which can adaptively adjust the sampling points according to the object's shape to capture the features of the target better. In addition, C3DCN can also reduce the error of the target bounding box and improve the accuracy of detection and positioning. In the face of this garbage data containing 38 kinds, the original coupled head is replaced by the decoupled head, allowing for the classification and regression branches to learn the category and location features of the target freely, making the model more flexible and more suitable for complex targets. Soft-NMS can reduce redundant detection boxes, enable the model to locate and identify garbage more accurately, and reduce the false detection rate. Due to the different sizes and types of daily household waste, there may be a complex identification environment. ODConv can deal with small-target garbage more effectively, C3DCN can better adapt to the change of the object shape, decoupled head is better for identifying complex objects, and Soft-NMS can reduce redundant detection boxes when dealing with a large number of overlapping targets. Figure 4 shows the YOLOv5-OCDS model's structure.

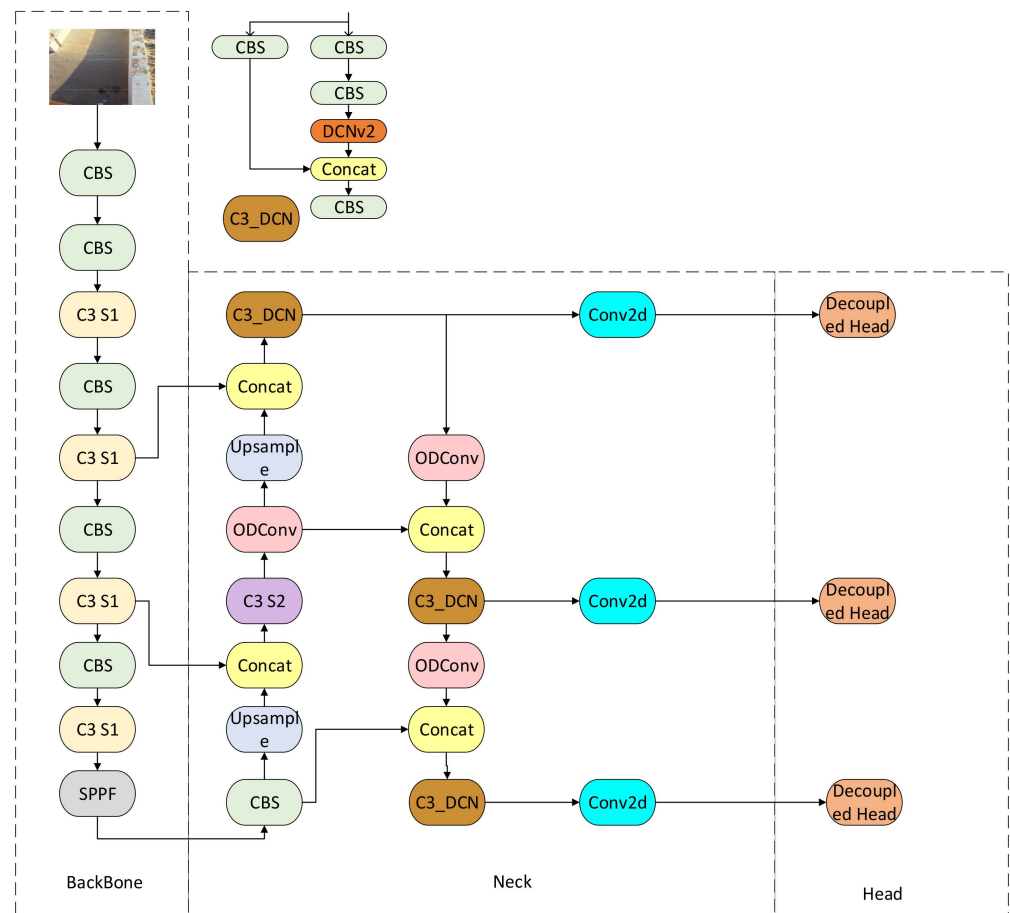


Figure 4. YOLOv5-OCDS structure diagram.

4. Experiment

4.1. Experimental Environment

All experiments were conducted on a computer equipped with AMD EPYC 7601, 32.0 GB RAM, NVIDIA GeForce GTX 3080TI (GPU) with 12GB video memory, and an Ubuntu 20.04 operating system. The code written for this experiment used Python 3.8 as the programming language and Jupyter Notebook as the development tool, with Pytorch version 1.12.0, torchvision version 0.13.0, and Cuda version 11.3.

4.2. Experimental Dataset

This experiment used part of the “Huawei Garbage Classification Challenge Cup” garbage image dataset and some self-made data to compose the experimental dataset. The dataset contains 17,057 pictures, including 38 categories, such as book, bag, basin, metal, daily paper, cardboard box, pot, ceramic utensil, shoe, fish bone, bubble, etc. The dataset is divided into the training dataset, validation dataset, and test dataset according to the ratio of 6:2:2. The training dataset has 10,233 images, the test dataset has 3412 images, and the validation dataset has 3412 images.

This experiment was trained, validated, and tested under the same hyperparameters. Where epochs are set to 500, the batch size is 32, momentum is 0.937, and the learning rate is 0.001. Moreover, mAP@50, mAP@50:95, precision, recall, and FPS were used as the evaluation indices of model performance.

4.3. YOLOv5-OCDS Model Training

Figure 5 shows the training process diagram of each model. The Soft-NMS method is a data postprocessing method, so it cannot be shown in the training diagram. When

using the YOLOv5-OC and C3DCN methods, the training automatically stops at 384 and 466 rounds because the mAP does not grow for more than 50 consecutive epochs. When other parameters, such as epoch, learning rate, etc., are set the same, they can intuitively provide important information about the entire training process. The figure shows that when the IoU is 0.5, the model has a small improvement but a higher value. In contrast, when the IoU increases, the model significantly improves but has a relatively low value. recall and precision were also improved to varying degrees.

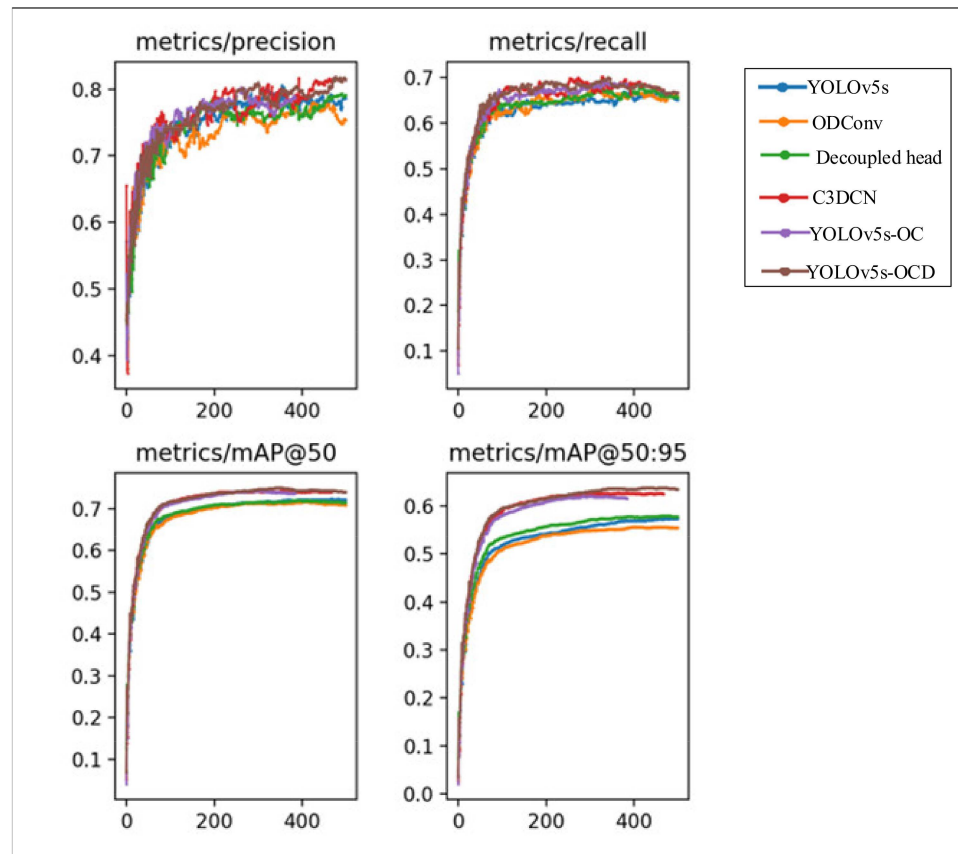


Figure 5. The training curves of the model.

4.4. Ablation Experiment

To verify the model’s effectiveness after combining ODConv, C3DCN, decoupled head, and Soft-NMS, a set of ablation experiments were designed. The YOLOv5s was selected as the benchmark model in the ablation experiment. The experimental results are shown in Table 1. The PR curves of different models on the test dataset are shown in Figure 6, the PR curves of different models on the validation dataset are shown in Figure 7, the comparison of mAP@50 and mAP@50:95 values of different models are shown in Figure 8, and the comparison of the precision and recall of different models is shown in Figure 9.

Table 1. Ablation experiment.

Methods	mAP@50	mAP@50:95	P	R	FPS
YOLOv5s	72%	56.7%	76.7%	65.6%	80 f/s
YOLOv5s+ODConv	72.3%	56.7%	77.5%	66.2%	71.4 f/s
YOLOv5s+Decoupled Head	73.1%	59%	77.9%	65.9%	82.6 f/s
YOLOv5s+C3DCN	74.9%	63.4%	79.7%	68.8%	78.7 f/s
YOLOv5s+Soft-NMS	74.7%	61.9%	77%	65.6%	83.3 f/s
YOLOv5s-OC	75.2%	63.1%	77.1%	68.8%	72.4 f/s
YOLOv5s-OCD	75.3%	64.6%	78.3%	68.9%	68 f/s
YOLOv5s-OCDS	77.3%	69%	78%	69.4%	61.7 f/s

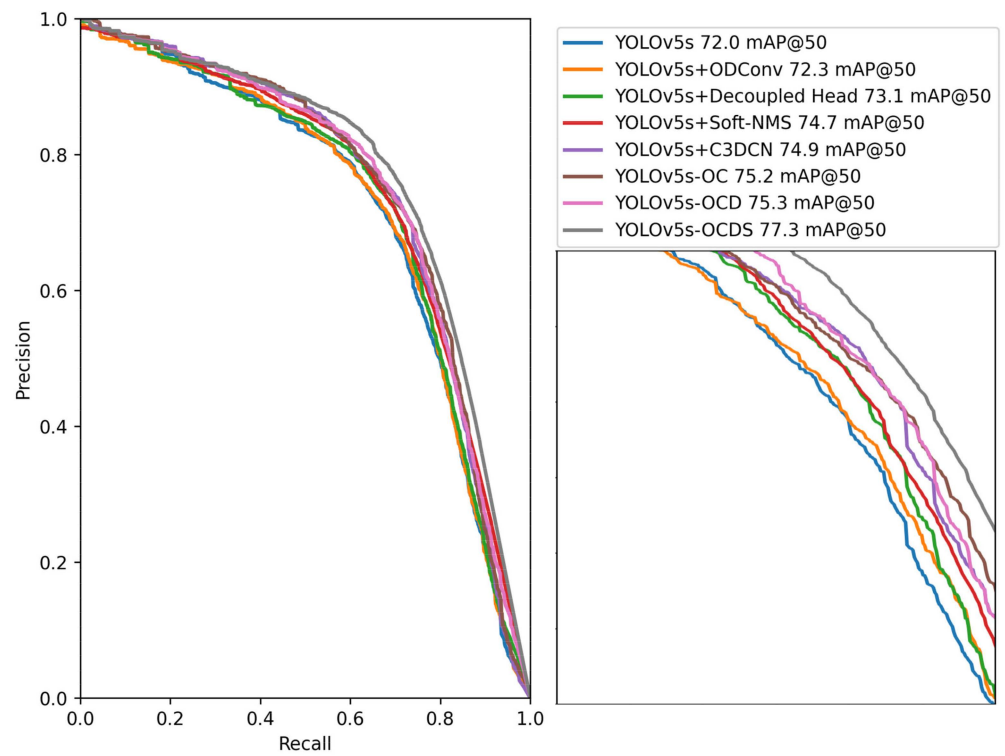


Figure 6. PR curve for the YOLOv5s ablation experiment on the test dataset.

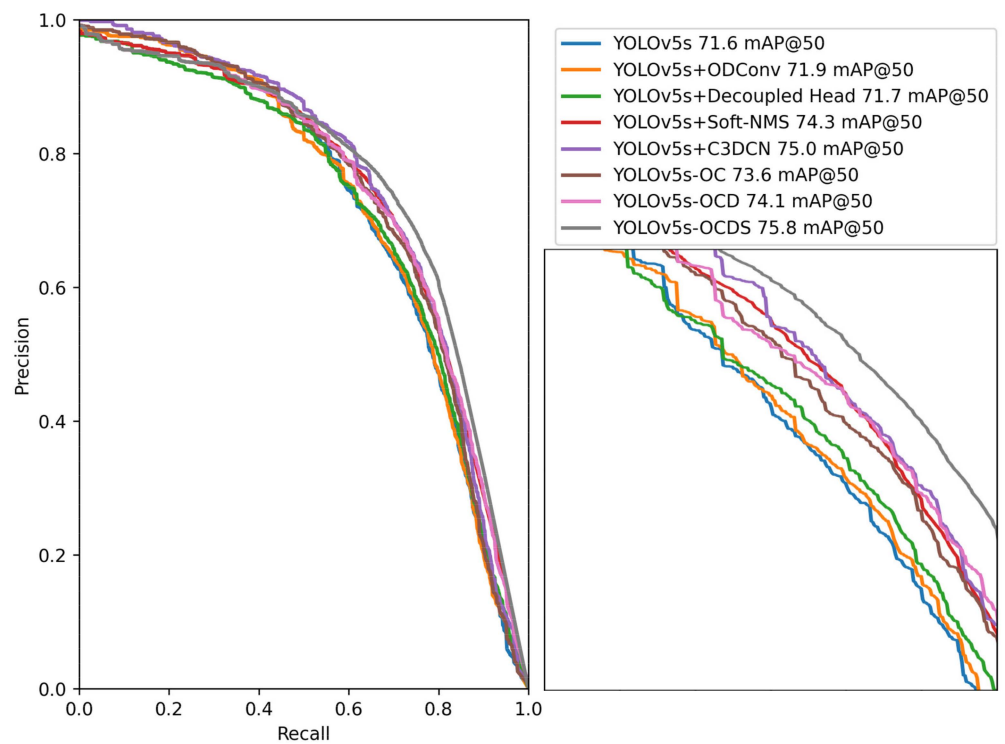


Figure 7. PR curve for the YOLOv5s ablation experiment on the validation dataset.

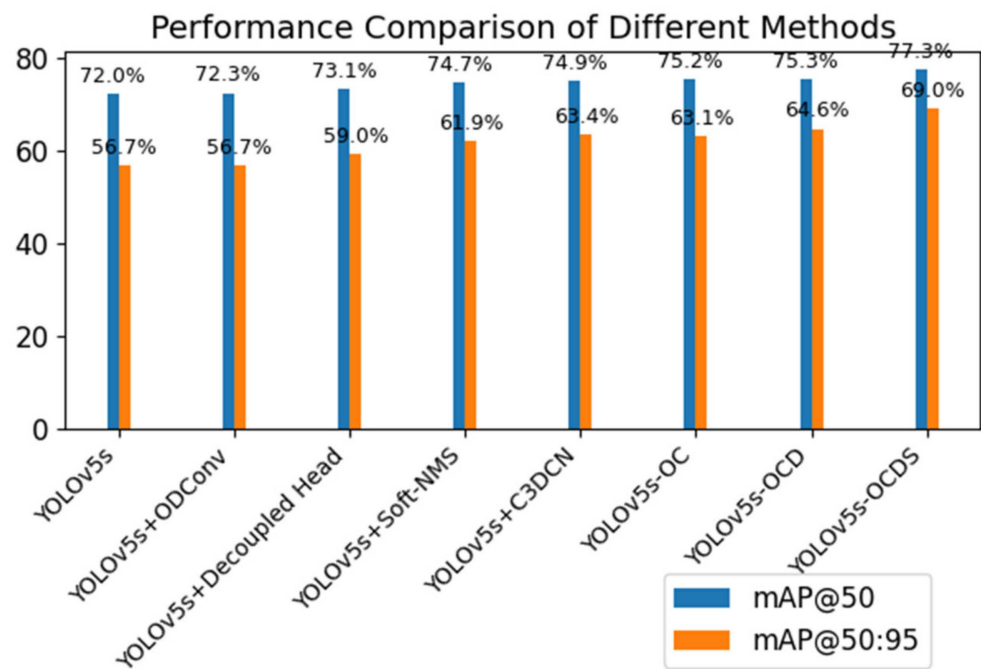


Figure 8. Comparison of the values of mAP@50 and mAP@50:95 for different models.

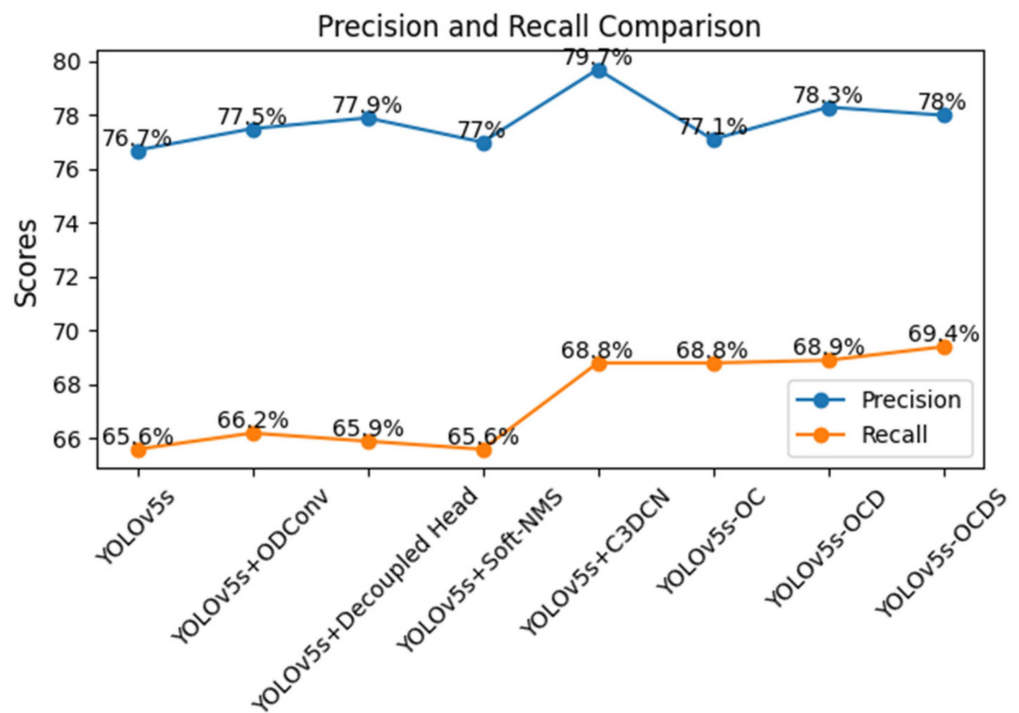


Figure 9. Comparison of precision and recall of different models.

From Table 1 and Figures 6 and 8, it can be found that adding ODConv, C3DCN, decoupled head, and Soft-NMS to YOLOv5s can improve mAP@50 and mAP@50: 95 to various degrees. When ODConv is added, the detection effect of the model for small targets is improved, and the sampling position and sampling weight of the convolution kernel are adjusted, which increases mAP@50 by 0.3% and improves the precision and recall. By adding OD-Conv and C3DCN, the convolution kernel can be dynamically adjusted to adapt to the shapes of different targets. The mAP@50 and mAP@50:95 increase by 3.2% and 6.4%, respectively, while the number of parameters in the model is reduced. When

ODConv, C3DCN, and the decouple head are added, the model's training efficiency and generalization ability are improved, and the mAP@50 and mAP@50:95 improve by 3.3% and 7.9%, respectively; still, the parameter amount of the model is increased. When ODConv, the decouple head, C3DCN, and Soft-NMS are added, repeated detections are effectively reduced, making mAP@50 and mAP@50:95 increase by 5.3% and 12.3%, respectively. However, for severely stacked targets, Soft-NMS may lead to some missing detections because it will reduce the confidence score of targets with more overlap, and the model may not achieve good results when dealing with such pictures.

4.5. Comparison Experiment

This study conducted a comparison experiment with various popular object detection models to verify the superiority of the improved model compared with other models. The results are shown in Table 2. First, it was compared with the classic two-stage Faster R-CNN target model; then with the classic single-stage RetinaNet model; then with the end-to-end DINO model; and finally with the YOLOv3, YOLOv4-tiny, YOLO-WASTE, YOLOX YOLOv7-tiny, and YOLOv8 models.

Table 2. Comparison experiment of different models.

Models	Backbone	mAP@50	Parameters	GFLOPs
Faster R-CNN	ResNet50	64.9%	41.5 M	193.78
RetinaNet	ResNet50	67.8%	36.9 M	220.73
DINO	ResNet50	75.4%	46.6 M	279
YOLOv3	Darknet53	72.9%	61.7 M	155.2
YOLOv4-tiny	CSPDarknet53	68.3%	6.0 M	16.3
YOLO-WASTE	CSPDarknet53	72.5%	32.5 M	140.3
YOLOv5s	CSPDarknet53	72%	7.1 M	16.1
YOLOX	Darknet53	64.4%	8.95 M	26.84
YOLOv7-tiny	CSPDarknet53	71.4%	6.1 M	13.3
YOLOv8s	CSPDarknet53	76%	11.1 M	28.5
YOLOv5s-OCDS	CSPDarknet53	77.3%	13.8 M	25.9

According to the data in Table 2 and Figures 10 and 11 in the comparison experiment, among the models of the YOLO series mentioned in this paper, the mAP@50 of YOLOv5s-OCDS is the highest, which is 4.8% higher than that of YOLO-WASTE, 12.9% higher than YOLOX, 5.9% higher than YOLOv7-tiny, and 1.3% higher than YOLOv8, reaching 77.3%. Compared with the classic Faster R-CNN and RetinaNet models, mAP@50 improves by 12.4% and 9.5%, respectively, while the parameters and GFLOPs are significantly reduced. Compared with end-to-end DINO, mAP is also enhanced, while the parameter quantity and GFLOPs are also lower. Although YOLOv5s-OCDS has slightly higher parameters and GFLOPs than YOLOv5s, it has a considerable improvement on mAP@50. YOLOv5s-OCDS replaces part of the convolution in the neck with ODConv to better capture the features and boundary information of the target, and it replaces part of the C3 module in the neck with C3DCN, which improves the accuracy of the model and can better identify small garbage objects or objects with complex shapes. While the decoupled head usually shows a better performance when dealing with complex targets such as garbage sorting, Soft-NMS can effectively suppress the bounding boxes with larger overlap. Therefore, this model achieves better results in garbage detection tasks.

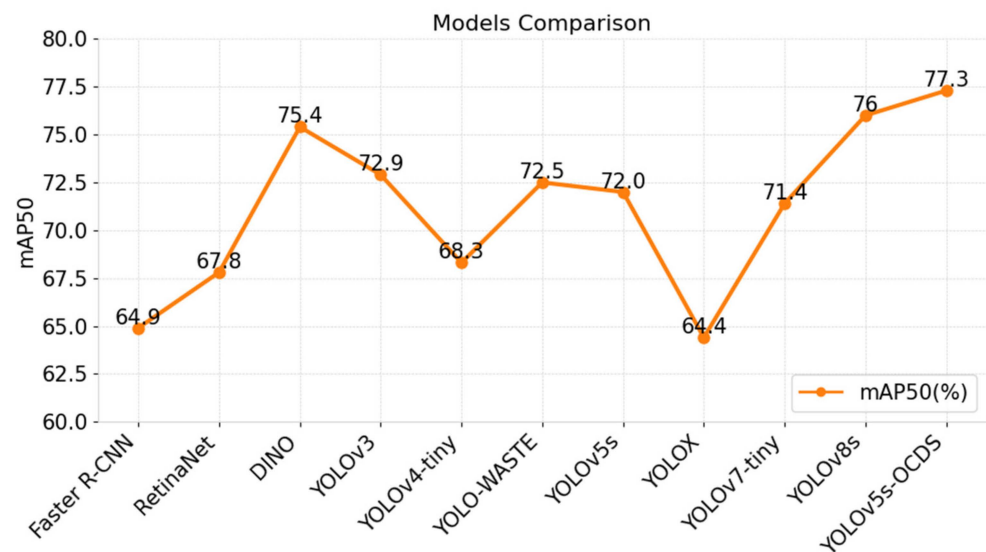


Figure 10. Comparison of mAP@50 between classical model and improved model.

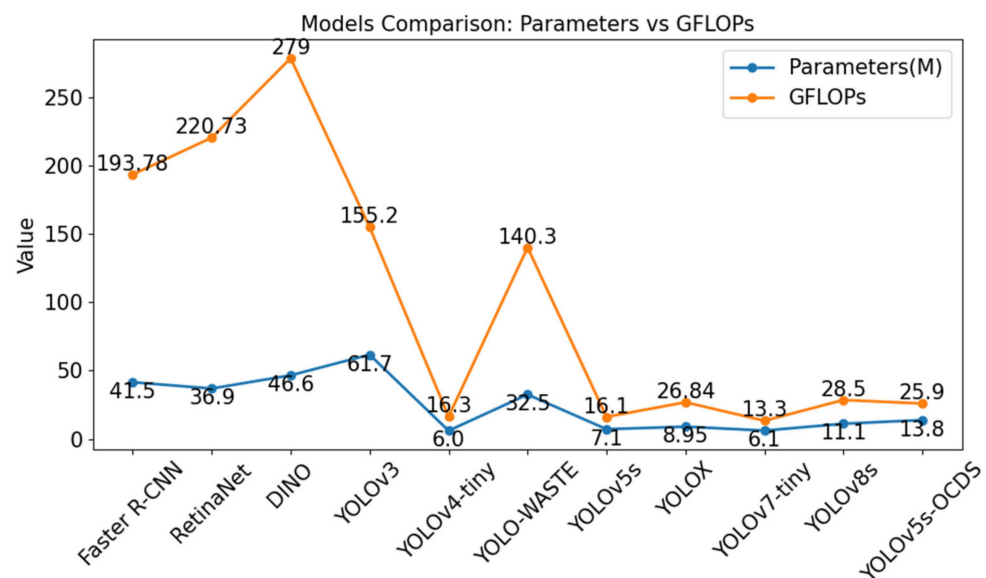


Figure 11. Comparison of parameters and GFLOPs between classical model and improved model.

4.6. Visualization

In some complex scenarios, YOLOv5s may miss and error detect some targets, as shown in Figure 12. In the figure, (a) is the ground truth, that is, our manual annotation; (b) is the detection result of YOLOv5s; and (c) is the detection result of YOLOv5s-OCDS. It can be seen that, compared with the ground truth, YOLOv5s mistakenly detected a metal at the bottom, mistaking the branch for a piece of metal. In contrast, YOLOv5s-OCDS detected the defiled plastic, which is in the ground truth, so the detection effect of YOLOv5s-OCDS is better.

If some photos of the object are incomplete, there will be a case of missing detection. As shown in Figure 13, compared with the ground truth, the detection result of YOLOv5s is not the same as that of the ground truth, but the YOLOv5s-OCDS can detect the paper at the lower right corner, although its accuracy is not high.

YOLOv5s is also prone to missing detection when facing images with smaller targets. As shown in Figure 14, YOLOv5s does not recognize the defiled plastic on the right side of the picture when facing small-target pieces of garbage, while YOLOv5s-OCDS can recognize them as well as ground truth.



(a) Ground truth.



(b) YOLOv5s detection results.



(c) YOLOv5s-OCDS detection results.

Figure 12. Effect of complex scene detection.



(a) Ground truth.

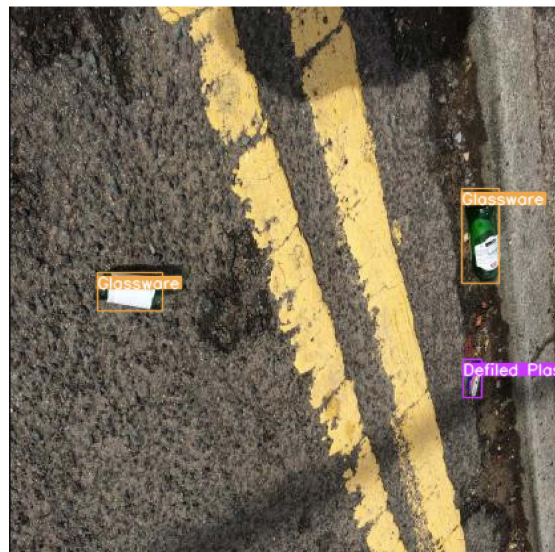


(b) YOLOv5s detection results.

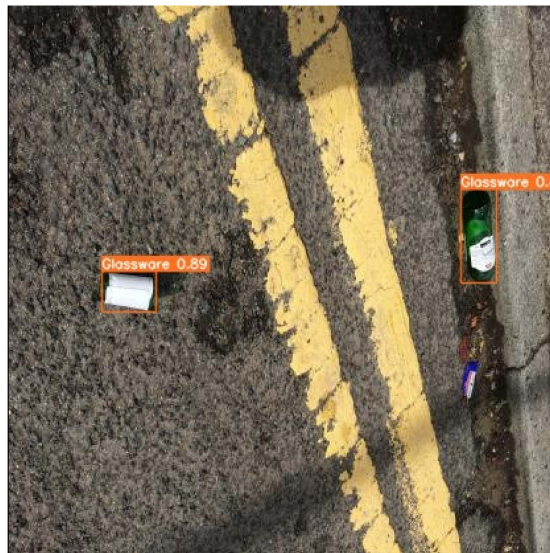


(c) YOLOv5s-OCDS detection results.

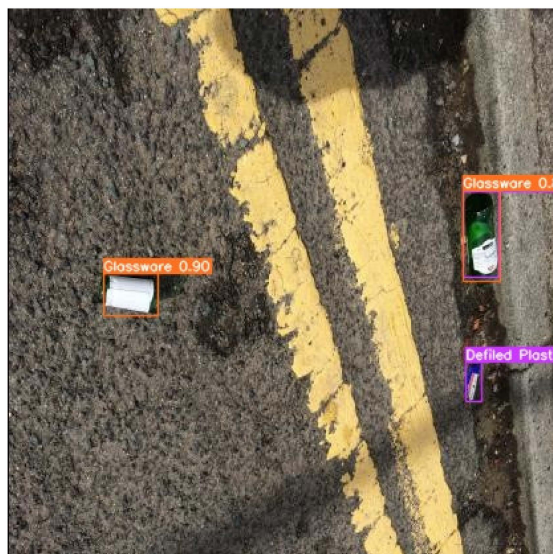
Figure 13. The detection effect when the object display is incomplete.



(a) Ground truth.



(b) YOLOv5s detection results.



(c) YOLOv5s-OCDS detection results.

Figure 14. The detection effect of a small target.

5. Conclusions

This paper proposes a garbage detection model, YOLOv5-OCDS, based on improved YOLOv5s to improve the accuracy in complex environments, objects displayed incompletely, and small detected objects.

In the YOLOv5-OCDS model, ODConv is first used to replace partial convolution in the neck, and the interaction and importance between channels can be adjusted adaptively by learning weights to extract more distinguishing features; the C3DCN module is proposed and used to replace part of the C3 structure in the neck so that the model can sample the target more accurately and avoid information loss; the decoupled head is used to replace the original coupled head so that the model can learn the fine-grained characteristics of the target better, improve the recognition ability of different categories, help the model judge the category of garbage more accurately, and reduce the misclassification; using Soft-NMS instead of ordinary NMS reduces the confidence in the elimination process to retain the detection boxes of smaller targets, which can help to keep the detection results of these small targets and improve the accuracy of the model.

Compared with the original YOLOv5s, YOLOv5-OCDS improves mAP@50 by 5.3% and mAP@50:95 by 12.3%. Compared with the Faster R-CNN, YOLOv5-OCDS increases mAP@50 by 12.4%. YOLOv5-OCDS can locate and identify targets more accurately and has a better detection ability and robustness. In the higher IoU threshold range, the model can better adapt to changes in different target shapes and sizes. While the model improves mAP, it also increases the number of parameters and GFLOPs. The introduction of Soft-NMS reduces the confidence score of objects with a lot of overlap, there may be a case of missing detection when detecting heavily stacked objects. In the following work, the lightweight degree of the model should be improved as far as possible, without affecting the model's accuracy.

Author Contributions: Conceptualization, Q.S. and X.Z.; methodology, Q.S., X.Z. and J.W.; software, X.Z.; validation, Y.L.; investigation, Y.L. and X.Z.; writing—original draft preparation, X.Z. and Y.L.; writing—review and editing, Q.S. and J.W.; visualization, X.Z. and Y.L.; supervision, J.W. and Q.S.; project administration, X.Z. and Y.L.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Innovation Foundation of Hebei Intelligent Internet of Things Technology Innovation Center under Grant AIOT2203, the Defense Industrial Technology Development Program under Grant JCKYS2022DC10, and China Scholarship Council No.202208130054.

Data Availability Statement: Data are available upon request due to privacy restrictions. The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ruiz, V.; Sánchez, Á.; Vélez, J.; Raducanu, B. Automatic image-based waste classification. In Proceedings of the 8th International Work-Conference on the Interplay Between Natural and Artificial Computation, Almería, Spain, 3–7 June 2019.
2. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
3. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition, San Diego, CA, USA, 20–25 June 2005.
4. Zhu, M.; Feng, T.; Zhang, Y. Remote Sensing Image Multi-Target Detection Method Based on Fd-Ssd. *Comput. Appl. Softw.* **2019**, *36*, 232–238.
5. Zeng, D.; Zhang, S.; Chen, F.; Wang, Y. Multi-Scale CNN Based Garbage Detection of Airborne Hyperspectral Data. *IEEE Access* **2019**, *7*, 104514–104527. [[CrossRef](#)]
6. Ma, W.; Yu, J.; Wang, X.; Chen, J. Garbage Detection and Classification Method Based on Improved Faster R-CNN. *Comput. Eng.* **2021**, *47*, 294–300.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards realtime object detection with region proposal networks. In Proceedings of the Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
8. Mikami, K.; Chen, Y.; Nakazawa, J.; Iida, Y.; Oya, Y. DeepCounter: Using deep learning to count garbage bags. In Proceedings of the 2018 IEEE 24th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), Hakodate, Japan, 28–31 August 2018.

9. Liu, Y.; Ge, Z.; Lv, G.; Wang, S. Research on automatic garbage detection system based on deep learning and narrowband Internet of things. In Proceedings of the Journal of Physics: Conference Series, Suzhou, China, 22–24 June 2018.
10. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
11. Xu, W.; Xiong, W.; Yao, J.; Shen, Q. Application of Garbage Detection Based on Improved YOLOv3 Algorithm. *J. Optoelectron. Laser.* **2020**, *31*, 928–938.
12. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
13. Zhang, Q.; Yang, Q.; Zhang, X.; Wei, W.; Bao, Q.; Su, J.; Liu, X. A multi-label waste detection model based on transfer learning. *Resour. Conserv. Recycl.* **2022**, *181*, 106235.
14. Jia, K.; Ma, Z.; Zhu, R.; Li, Y. Attention-mechanism-based light single shot multiBox detector modelling improvement for small object detection on the sea surface. *J. Image Graph.* **2022**, *27*, 1161–1175.
15. Salimi, I.; Dewantara, B.S.B.; Wibowo, I.K. Visual-based trash detection and classification system for smart trash bin robot. In Proceedings of the 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), Bali, Indonesia, 29–30 October 2018.
16. Hu, B.; Fu, H.; Wang, W.; Zhang, B.; Tang, F.; Ma, S.; Lu, Q. Research on deep sorting approach based on infrared spectroscopy for HighValue utilization of municipal solid waste. *Spectrosc. Spectr. Anal.* **2022**, *42*, 1353–1360.
17. Ma, W.; Wang, X.; Yu, J. A Lightweight Feature Fusion Single Shot Multibox Detector for Garbage Detection. *IEEE Access* **2020**, *8*, 188577–188586. [[CrossRef](#)]
18. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot MultiBox detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV2016), Amsterdam, The Netherlands, 8–16 October 2016.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
20. Feng, J.; Tang, X.; Jiang, X.; Chen, Q. Garbage disposal of complex background based on deep learning with limited hardware resources. *IEEE Sens. J.* **2021**, *21*, 21050–21058.
21. He, K.; Gkioxari, G.; Piotr, D.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
22. Liu, B.; Wang, X. Garbage detection algorithm based on YOLO v3. In Proceedings of the 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 25–27 February 2022.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
24. Pan, Z. Research on improved Yolo on garbage classification task. In Proceedings of the 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 25–27 February 2022.
25. Li, X.; Tian, M.; Kong, S.; Wu, L.; Yu, J. A modified yolov3 detection method for vision-based water surface garbage capture robot. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420932715.
26. Iqbal, U.; Barthelemy, J.; Perez, P.; Davies, T. Edge-computing video analytics solution for automated plastic-bag contamination detection: A case from remondis. *Sensors* **2022**, *22*, 7821. [[PubMed](#)]
27. Wang, C.; Zhou, Y.; Li, J. Lightweight Yolov4 Target Detection Algorithm Fused with ECA Mechanism. *Processes* **2022**, *10*, 1285. [[CrossRef](#)]
28. Patel, D.; Patel, F.; Patel, S.; Patel, N.; Shah, D.; Patel, V. Garbage detection using advanced object detection techniques. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021.
29. Yan, X.; Yang, Y.; Feng, L.; Wang, L.; Tan, M. A garbage classification method based on improved YOLOv5. In Proceedings of the 2022 International Conference on Networks, Communications and Information Technology (CNCIT), Beijing, China, 17–19 June 2022.
30. Lin, J.; Yang, C.; Lu, Y.; Cai, Y.; Zhan, H.; Zhang, Z. An improved Soft-YOLOX for garbage quantity identification. *Mathematics* **2022**, *10*, 2650. [[CrossRef](#)]
31. Li, C.; Zhou, A.; Yao, A. Omni-dimensional dynamic convolution. *arXiv* **2012**, arXiv:2209.07947.
32. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. *arXiv* **2018**, arXiv:1811.11168.
33. Li, C.; Li, L.; Geng, Y.; Jiang, H.; Cheng, M.; Zhang, B.; Ke, Z.; Xu, X.; Chu, X. YOLOv6 v3. 0: A Full-Scale Reloading. *arXiv* **2023**, arXiv:2301.05586.
34. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving object detection with one line of code. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.