

Received 22 July 2023, accepted 9 August 2023, date of publication 21 August 2023, date of current version 29 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3306951

## RESEARCH ARTICLE

# Detection of Commodities Based on Multi-Feature Fusion and Attention Screening by Entropy Function Guidance

AN XIE<sup>1</sup>, KAI XIE<sup>1,2</sup>, HAO-NAN DONG<sup>1</sup>, AND JIAN-BIAO HE<sup>1,3</sup>

<sup>1</sup>School of Electronic and Information, Yangtze University, Jingzhou 434023, China

<sup>2</sup>Western Research Institute, Yangtze University, Karamay 834000, China

<sup>3</sup>School of Computer Science and Engineering, Central South University, Changsha 410083, China

Corresponding author: Kai Xie (500646@yangtzeu.edu.cn)


This work was supported in part by the National Natural Science Foundation of China under Grant 62272485, in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2020DO1A131, in part by the Teaching and Research Fund of Yangtze University under Grant JY2020101, and in part by the Undergraduate Training Programs for Innovation and Entrepreneurship of Yangtze University under Grant Yz2022056.

**ABSTRACT** Although traditional convolutional neural networks (CNN) have been significantly improved for target detection, they cannot be completely applied to objects with occlusions in commodity detection. Therefore, we propose a target detection method based on an improved YOLOv5 model and an improved attention mechanism algorithm is proposed to solve the commodity occlusion problem. This method improves the traditional YOLO deep convolution network, features a more detailed BiFPN layer, and performs lightweight two-way feature fusion, where the multidimensional features of the commodities are convolved and fused, thus improving the overall detection speed and accuracy of the YOLO-R algorithm. Feature entropy is introduced to the attention channel to restrict the threshold value and obtain the global information of the occlusion target. The global information obtained is fused with a bidirectional feature pyramid layer to enhance the robustness of the features. This method could accurately and quickly detect the occluded commodities and the detection accuracy has been greatly improved. Experiments show that the improved YOLO-R model can improve the accuracy and speed of commodity detection, and can achieve good results in objective evaluation. The average accuracy of commodity detection on the self-made product dataset is up to 97.80%, and the detection rate is 22.72F/s. Therefore, the method in this paper has high detection accuracy and fast detection speed.

**INDEX TERMS** Lightweight feature fusion, attention channel, feature entropy, bidirectional feature.

## I. INTRODUCTION

Target detection is an important research subject in the field of computer vision [1]. It detects whether a video or image has a target object that needs to be detected and determines its regional coordinates and category information. Over the years, image processing and recognition have become mainstream target detection methods and the accuracy and speed of target detection have improved significantly. Such as the two-stage detector Faster-RNN [2],

The associate editor coordinating the review of this manuscript and approving it for publication was Wei-Yen Hsu .

the single-stage detector RetinaNet [3], YOLOv5 [4], YOLOv7 [6], [7], [8], [9], [10], [11] and YOLO8 [12], which improve network detection accuracy by increasing network depth. At present, intelligent visual retail commodity detection technology requires both high precision and high speed. As for the single-stage detector, YOLO5 has the advantages of high accuracy and speed.

Currently, the methods for object occlusion detection and processing can be divided into methods based on constrained visible parts and components, methods based on the optimization loss function, and combined innovation based on multiple methods. Different parts are used to distinguish the

location of the occlusion for effective feature fusion or to define the information of certain areas. The loss function is optimized by constraining the distance loss between the prediction and rear frames to make it closer to the real detection frame. Another method is to start with the occlusion dataset, train the detection model by generating a large number of occlusion models, and use the attention mechanism to enhance the reliability of the network and improve the detection effect and robustness.

In the process of commodity testing, we first obtain the movement information and angle information of commodities. Then, the feature of the proportion of unobstructed parts is extracted through the ResNet feature extraction network [13], [14] and the features of some location areas of commodities are combined with feature entropy  $E_i$  assigning different weights to the occluded and visible parts of the commodities. Subsequently, a spatial attention feature map is generated from the first-layer structure hole convolution of the squeeze-and-excitation network [15] and the attention module of the second layer. This feature map under the original feature entropy under the activation of spatial attention is the input and the effective features of the commodities are the output. Thus, the basic features of layer commodities or foreign objects are obtained. Feature entropy is introduced into the attention channel to limit the threshold and obtain global information on the occlusion target. Furthermore, the global information [16], [17] is obtained and convoluted with the adaptive hole under global embedding, and the effective feature points of the target commodity are extracted. Based on the effective feature points, a commodity detection frame is constructed in the channel. Moreover, the influence of the attention mechanisms is examined. In addition to the characteristic attributes of the commodity itself, the generated commodity location-related information is extracted in the case of commodity occlusion and is imported into the scale-adaptive module. The detection frame is generated adaptively through K-means++ clustering, following which the distance measurement loss regression of the intersection over union (IOU) [18], [19], [20] is used to reduce the impact of the redundant features. The commodity detection frame is constructed using the channel correlation of the attention mechanism [21]. In addition to the characteristic attributes of the commodity itself, the information relate to the position of the commodity generated by the optical flow difference extraction in the case of commodity occlusion is imported into the scale-adaptive module. According to the occlusion ratio, the attenuation weight method is introduced to screen the high-quality positive samples for inclusion in the model training, effectively improving the detection frame size under occlusion. Based on the reconciliation of the region proposal network (RPN) network [22], [23], the detection confidence is obtained to compare the confidence weights and to generate the appropriate and accurate detection frames.

#### (1) Bi-directional feature pyramid networkfeature fusion

The first step in feature fusion is the extraction of high-level features from the backbone for direct prediction. However, this structure does not include feature fusion, which results in low accuracy. Subsequently, a feature pyramid network (FPN) [24] which is based on the concept of feature fusion is proposed. First, a top-down path is established for feature fusion. Then, the fused feature maps are then used for higher-level features to obtain the semantic information, which is used to improve the commodity accuracy. However, this top-down FPN is easily limited by the one-way information flow. Therefore, PANet has been proposed in recent years, and it builds a top-down path based on the FPN and uses the strong location information of the underlying feature map to fuse features, thus improving detection accuracy. However, PANet extracts feature from lower to higher levels which can easily lead to missing feature extractions. The BiFPN used in this study is superimposed by a simple feature map, following which different resolutions of the feature map are entered into the same node, thereby fusing more features without increasing the parameters.

#### (2) Attention mechanism

In deep learning, the attention mechanism is inspired by the human visual processing mechanism, whereby people focus on areas of interest. Similar algorithms, such as those for eliminating the background information of interference detection when detecting images to quickly locate the region of interest, are an important research field in target detection. Therefore, the purpose of introducing an attention mechanism is to eliminate redundant information and extract effective information about commodities, thereby improving the network performance.

#### (3) Non-maximum suppression (NMS) based on the soft-IOU algorithm

The traditional NMS algorithm [25] uses greedy clustering at a fixed distance. This is achieved by selecting a large number of high-score detection results and deleting the neighboring results beyond the threshold, thus balancing the accuracy and recall rate. However, if the prediction box is not aligned with the real box, the IOU cannot accurately reflect its coincidence and therefore cannot filter the detection box effectively. Therefore, a Dsoft-IOU algorithm is used to improve the NMS in this study. The algorithm uses a new intersection-to-ratio formula and optimizes the penalty function to increase the consideration of the center distance of the prediction box, thereby reducing the confidence of the prediction box, preventing the deletion of the preselection box, improving the recall rate, and enhancing the prediction ability of the model.

The structure of this article is as follows: Section II introduces the overall algorithm and related theory. Section III introduces the details of the experiment, including the experimental platform, experimental dataset, comparative experiment, ablation experiment, visualization of experimental data and results, experimental results, and analysis. Section IV summarizes the proposed algorithm.

## II. RELATED WORK

In traditional smart retail containers, the most common technologies are gravity induction, RFID wireless radio-frequency identification [26], face recognition, and two-dimensional code. With the heightened pursuit of improved consumer experience, RFID and gravity-sensing technologies [27] are not in line with the current trend. Simultaneously, the process of real-time target detection for blocking commodities [28], [29] which is affected by objective factors such as illumination, motion blurring [30], environmental impact, and morphological changes, reduces the accuracy and stability of target detection. In the field of visual detection, occlusions have a significant impact on scene reconstruction, object recognition [31], behavior recognition [32], [33] target tracking, stereo matching, visual measurement, and other visual tasks. Thus, the occlusion issue has become separated from related visual tasks over time and has been widely studied extensively by both domestic and foreign researchers. Therefore, reducing detection and misdetection and increasing the detection accuracy of occlusion targets in the process of occlusion detection has become a research hotspot in the field of commodity target detection.

In this study, a new feature extraction algorithm that combines global information and an adaptive matching algorithm with a feature entropy mechanism is presented for detecting occlusions and foreign bodies in smart retail containers. The algorithm applies a feature entropy mechanism and adds an attention module [34], [35] to the uncovered part to extract as many effective commodities features as possible. To solve the problems of illumination change, the color similarity between the background and target, and the difficulty in feature extraction under occlusion during detection, C. Xiu proposed a method combining the Camshift algorithm [36], [37] and Kalman filter to improve the robustness of the algorithm and ensure its real-time performance. Liang et al. [38] propose a DetectPreer category auxiliary transformer object detector based on Transform, which can use data augmentation technology to improve the backbone network and use the attention mechanism to extract channel spatial characteristics and direction information. Liang et al. [39] propose an improved sparse R-CNN, which integrates the attention module with ResNeSt to construct a feature pyramid and modify the backbone network to extract more important effective features.

The influence of different occlusion levels on the detection performance is quantified based on an analysis of the commodity detection performance under different occlusion ratios. Based on the analysis, high-quality positive samples are selected for inclusion in the model training by introducing attenuation weights according to the occlusion ratio, which effectively improved the detection performance under occlusion and located the uncovered areas. This method uses VGG to convolute the commodity characteristics. To obtain information on the commodity characteristics, the threshold value of the feature entropy is introduced to determine and convolute the holes in SENet [15], set the weight of the feature entropy threshold [40], import into the attention channel, and

finally entered into the pooling layer to obtain effective features. By calculating the position offset of the detection target over time, the IOU loss regression of the occlusion target prediction box and the real box is improved, the redundancy of the detection box is reduced, and the appropriate size of the detection box is determined under non-extreme suppression (RPN) [41]. This significantly improves the accuracy of the feature extraction for occlusion detection, enhances the detection robustness, and constructed an appropriate detection framework.

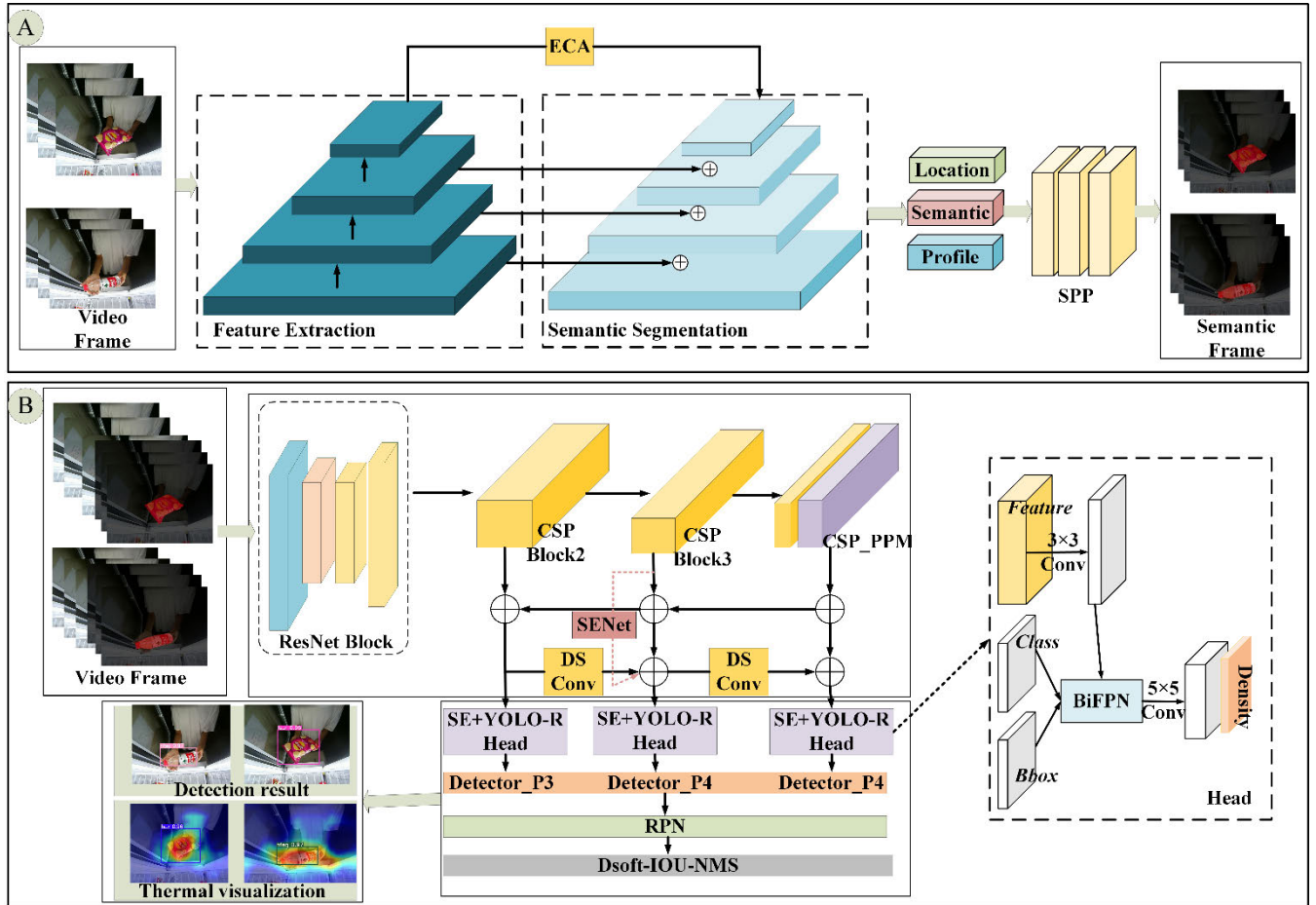
In this study, the structure of the target detection algorithm is modified. (1) An attention enhancement algorithm based on semantic segmentation is proposed to solve the missing high-frequency information in occlusion detection and to achieve the feature extraction of commodities in the unobstructed section of semantic segmentation, as well as also provide location information for video frame localization. (2) In the detection module, the designed YOLO-R algorithm, which combines a residual module to enhance the residual characteristics of the convolution network is deployed. The algorithm improves the NMS algorithm and sets the BiFPN feature pyramid. In feature extraction, it enhances the detailed information, and thus, fuses the features of the different scales of the network, enhances the robustness of the features, and improves the model accuracy. In YOLO-R, the attention mechanism of the SENet channel is added, which make the network model pay more attention to the unobstructed areas of commodities and extract effective information on commodities. Finally, in the processing stage, the NMS model is adjusted automatically using the Dsoft-IOU with the location information of the threshold set by the eigenvalue function, which prevents the true frame from being filtered out and enhances the generalization of the model.

## III. METHOD

In a related study, a new YOLO-R algorithm is designed. The YOLO-R algorithm uses the BiFPN feature pyramid for feature fusion. In the upsampling model, CARAFE lightweight sampling operators are used to increase the sampling characteristics. In this study, the SENet attention mechanism is added to YOLO-R and the attention mechanism threshold is increased by feature entropy. YOLO-R consists of two CSP structures designed to reduce inference and computing power, namely CSP blocks for backbone networks and CSP\_PPM mainly for neck network structures. In post-processing, the NMS algorithm of DSoft-IOU is redesigned. The penalty function threshold is increased, and samples are filtered out by DSoft-IOU range loss regression to reduce the impact of redundant features. The algorithm used in this study is shown in Fig. 1.

### A. COMBINING ECA SEMANTIC FEATURE EXTRACTION

The improved FCN used in this study uses U-Net improvement in the FCN network extractor to increase the ECA attention mechanism [33] in the U-Net network sampling



**FIGURE 1.** Overview of the proposed architecture. (a) We first perform initial feature extraction convolution based on U-Net to get the basic features and fuse the basic features to get the semantic segmentation results. (b) The semantic segmentation features and also the commodity unmasked features are convolutionally pooled, in which the features are sampled to obtain the commodity effective features, the detection framework is built according to the effective features, and the suitable detection frame is filtered using non-maximal suppression.

and to speed up model estimation. In this study, improvements are made to the U-Net semantics segmentation module, and a channel interaction strategy module (ECA module) is proposed. The module uses a one-dimensional convolution module, which effectively reduces the number of parameters and computing power, thus improving performance. The module includes only a few additional parameters to avoid the effect of dimension reduction convolution on the channel attention mechanism.

The selected ECA module optimizes the computational performance and model complexity. Firstly, the features are aggregated by global average pooling to obtain channel global information, and the global average pooling operation formula is as follows:

$$y = \frac{1}{H \times W} \sum_a^H \sum_b^W x_i(a, b) \quad (1)$$

In the formula,  $x_i(a, b)$  represents the  $i$ -th feature map with input size  $H \times W$ , which represents the global average pooling of feature  $x$ .

Secondly, using the channel dimension  $C$  adaptive to calculate the number of channels sharing weight  $k$ . The adaptive function formula is as follows:

$$k = \varphi(C) = \left\lfloor \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right\rfloor \quad (2)$$

In the formula:  $C$  is the channel dimension;  $b$  and  $r$  are constants, where  $b=1$  and  $r=2$ . The ratio of vector  $C$  to  $r$  is used to obtain the channel sharing weight  $k$ .

The extraction performance is improved through multi-channel shared weight information interaction as follows:

$$w = \sigma(C1D_k(y)) \quad (3)$$

In this study,  $C1D$  represents a one-dimensional convolution, and  $\sigma$  represents a sigmoid function. The amount of information in this study is relatively small, which is conducive to reducing the model's complexity of the model. Therefore, this method of channel attention information interaction ensures effectiveness and model efficiency. This enables efficient and fast extraction of commodity features.

## B. FEATURE EXTRACTION

### 1) CHARACTERISTIC ENTROPY WEIGHT DISTRIBUTION

Local feature points are extracted using the YOLO-R residual module, a convolution self-coding neural channel is added for the low-dimensional pixel points, and global features are exploited to enhance the detection characteristics. The adaptive weight redistribution algorithm originated from the similarity principle of the gestalt grouping. The degree of color similarity is used as an influencing factor for weighting, and the influencing factor is determined by the Euclidean distance. In this study, a fully connected layer  $F \in R^{C_F \times C_D}$  is used to integrated it into the model architecture, and the bias it learned is  $b_F \in R^{C_F}$ . These two parts are combined into a global feature that summarize the discriminatory content of the entire image,  $g \in R^{C_F}$ :

$$g = F \times \left( \frac{1}{H_D W_D} \sum_{h,w} d_{h,w}^p \right)^{1/p} + b_F \quad (4)$$

$p$  is expressed as a hyperparameter of the average pooling.  $d_{h,w}^p$  is the mapping features of the signature map.

Assuming the color features extracted by the keyframe are  $F_0, F_1, \dots, F_n$ , the features fused are

$$F_n = H_{Concat}(B_n^1, B_n^2, \dots, B_n^i) \quad (5)$$

Therefore, the fused feature is  $\{F_1, F_2, \dots, F_n\}$ , and there is

$$p(F_i) = \frac{F_i}{\sum_{i=1}^n F_i} \quad (6)$$

Its corresponding eigenvalue is

$$E_i = -\frac{2}{n+1} \sum_{i=1}^n p(F_i) \ln p(F_i) \quad (7)$$

where  $H_{Concat}(B_n^1, B_n^2, B_n^3 \dots B_n^i) \cdot p(F_i)$  is the characteristic ratio and  $E_i$  is the characteristic entropy used in this study. The  $E_i$  value is directly proportional to the effective characteristic quantity of commodities.

First, according to the feature weight set in this study  $\{\omega_1, \omega_2, \dots, \omega_n\}$ , the following formula is obtained:

$$\omega_i = \begin{cases} \frac{E_i}{\sum_{i=1}^n E_i}, & E_i > \tau \\ 0, & else \end{cases} \quad (8)$$

According to the weight, other IOUs can be filtered below the set threshold, and the obtained attention mechanism weight can be introduced into SENet, and the characteristics can be filtered to obtain the effective characteristics of the commodity.

### 2) YOLO-R MULTI-CASCADE CONVOLUTION FEATURE EXTRACTION

Commodity detection is the result of detection based on the target position and semantic information of the video frame. In this study, YOLO5 is used with a cascade for feature extraction according to the characteristics of sheltered commodities [43]. ResNet residual features are added based on

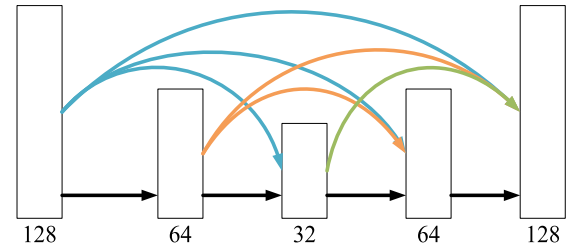


FIGURE 2. YOLO-R network structure.

YOLO5, which yielded the YOLO-R algorithm to learn the feature relationship between the enhancement layer and layer, enhance the network perception field, complete the feature convolution fusion of high and low resolution, compensate for the loss of high-resolution semantic information, enrich edge information, and improve the detection accuracy of the occluded targets.

Numerous residual module area features are used in the YOLO-R algorithm employed in this study. Therefore, a simple feature module is used instead of the representation algorithm module structure. In the residual extraction module, the backward-propagating module propagates from forward to backward, and finally selects the last layer in the feature block, as shown in Fig. 2.

The YOLO-R residual convolution is a ResNet deep residual network that can effectively improve the performance of feature extraction detection. This module is a framework that stacks blocks with the same connection shape. The blocks used in this study are also known as residual units. The residual element calculation process is as follows:

$$y_n = h(x_n) + \mathcal{F}(x_n, M_n) \quad (9)$$

$$x_{n+1} = f(y_n) \quad (10)$$

In this formula,  $F$  is a residual convolution function and  $5 \times 5$  convolution stacks are commonly used.  $x_n$  represents the feature layer entering for the  $n$  residual unit modules. Function  $f$  is the operation of adding feature weights, which is expressed as the ReLU activation function. Function  $h$  is an identical mapping,  $h(x_n) = x_n$ . The residual units are shown in Fig. 3.

Because  $h$  is an identical map:  $x_{n+1} = y_n$ , the following can be obtained:

$$x_{n+1} = x_n \mathcal{F} + (x_n, W_n) \quad (11)$$

By making recursive calls, it can be observed that

$$x_{n+2} = x_{n+1} + \mathcal{F}(x_{n+1}, W_{n+1}) = x_n + \mathcal{F}(x_n, W_n) + \mathcal{F}(x_{n+1}, W_{n+1}) \quad (12)$$

For cell  $L$  of any depth and cell  $l$  of any shallow layer, the following is obtained:

$$x_{n+1} = x_n + \mathcal{F}(x_n, W_n) \quad (13)$$

Therefore, the low-resolution image  $I_{LR}$  is extracted through feature extraction, and the shallow feature extracted

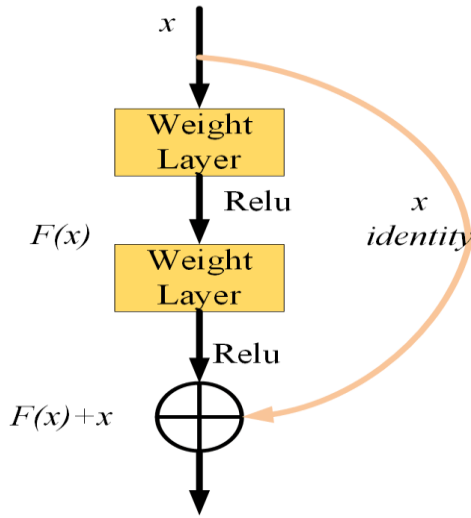


FIGURE 3. Residual element structure diagram.

by the YOLO-R residual network is  $F_0$ , which includes

$$F_0 = H_{Cov}(I_{LR}) \quad (14)$$

where  $H_{Cov}(I_{LR})$ , representing the YOLO-R residual convolution feature extraction, extracted shallow semantic features and transfers them to a deeper feature extraction module to extract deeper occluded commodity features.

The image features are then added pixel-by-pixel before being fused with the BiFPN to obtain an effective feature map of the commodity.

$$B_n^i = B_n^{i-1} + H_{Concat}(F_n^{i,1}, F_n^{i,1} + F_n^{i,2}, F_n^{i,1} + F_n^{i,2} + F_n^{i,3}) \quad (15)$$

where  $B_n^i$  is the deep-seated feature map output of the  $i$  mixed YOLO-R residual convolution block in the  $n$  multi-scale feature extraction module,  $F_n^{i,1}, F_n^{i,2}, F_n^{i,3}$  represent the distance feature outputs of the three different scale modules.

### 3) MULTISCALE FEATURE FUSION

With the deepening of the YOLO-R network level of the algorithm in this study, the commodity features are convoluted from low to high dimensions. However, with the deepening of the feature extraction in each layer of the YOLO-R extraction network, a few features are missing. Therefore, it is necessary to fuse features at different levels to enhance the feature semantics. A lightweight BiFPN is used for feature fusion, and the YOLO-R backbone network is used for bidirectional feature fusion at different scales.

A new BiFPN, where  $P_3 \sim P_7$  denote five input features, is adopted in this study. The algorithm adopted operation methods, such as lightweight (CARAFE) up-sampling (the up-sampling core prediction module and the feature reorganization module are used to predict the sampling core using the up-sampling core module, and then the feature

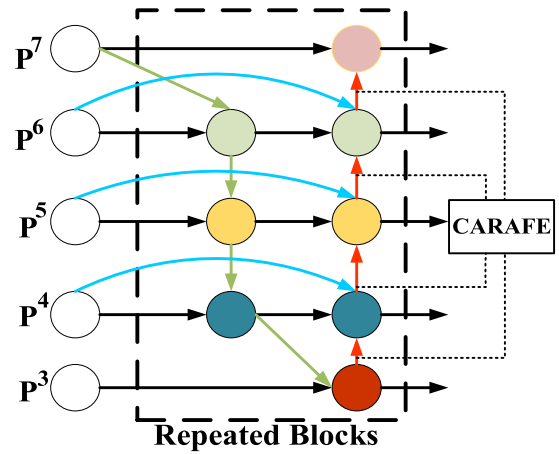


FIGURE 4. Structure diagram of BiFPN.

reorganization module is used to complete the up-sampling), down-sampling, and superposition in BiFPN to output the five extracted features of a single channel. According to the requirements of this study, this algorithm regards a BiFPN as a bidirectional stackable network structure for feature superposition to effectively enhance the feature fusion information of the occluded commodities. The structure is shown in Fig. 4.

The size of the feature map output by YOLO-R in this study is  $480 \times 640$ , and the five feature layers input in the BiFPN network are  $P_3^{IN} = (160, 160, 128)$ ,  $P_4^{IN} = (80, 80, 256)$ ,  $P_5^{IN} = (40, 40, 142)$ ,  $P_6^{IN} = (20, 20, 1024)$ ,  $P_7^{IN} = (10, 10, 2048)$ . The BiFPN assigns new weights  $w_i$  to the input features, and then quickly normalizes the weights linearly from  $P_3$  to  $P_5$ . The input formula for the fusion node is given by Equations (16)-(19):

$$P_4^{TD} = Conv(\frac{w_1 \cdot P_4^{IN} + w_2 \cdot Resize(P_5^{IN})}{w_1 + w_2 + \epsilon}) \quad (16)$$

$$P_3^{OUT} = Conv(\frac{w_3 \cdot P_3^{IN} + w_4 \cdot Resize(P_4^{TD})}{w_3 + w_4 + \epsilon}) \quad (17)$$

$$P_4^{OUT} = Conv(\frac{w_5 \cdot P_4^{IN} + w_6 \cdot P_4^{TD} + w_7 \cdot Resize(P_3^{OUT})}{w_5 + w_6 + w_7 + \epsilon}) \quad (18)$$

$$P_5^{OUT} = Conv(\frac{w_8 \cdot P_5^{IN} + w_9 \cdot Resize(P_4^{OUT})}{w_8 + w_9 + \epsilon}) \quad (19)$$

where  $Conv$  refers to the convolution operation,  $Resize$  refers to the upsampling (CARAFE) or downsampling operation on the input features. A lightweight operator (CARAFE) is introduced for upsampling to replace the nearest neighbor difference. The lightweight operator has low redundancy, strong feature fusion ability, and fast speed.  $w_i$  is the characteristic entropy weight,  $\epsilon = 0.0001$  is the stability coefficient.

In the fusion network, different feature entropy weights  $w_i$  are assigned to the input feature map such that the network can constantly adjust and determine each output feature.

A fast normalization method is adopted.

$$\text{Out} = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} \times \text{In}_i \quad (20)$$

where,  $\text{Out}$ ,  $\text{In}_i$  is the output and input characteristics.

According to the formula, each normalized feature entropy weight is  $w_i \in (0, 1)$ . Owing to the absence of the softmax operation in BiFPN, the efficiency is improved to a certain extent.

#### 4) SENET ATTENTION ALLOCATION COMBINED WITH CHARACTERISTIC ENTROPY

In SENet, the commodity feature information is entered into the channel attention mechanism through stacked clustering feature layers. The feature map under the restriction of information entropy is beneficial for filtering out the effective features and learning the weights of each layer automatically. In the SENet module, a  $5 \times 5$  Gaussian convolution kernel is used to increase the field of sensation, convolute to a lower dimension through a  $1 \times 1$  convolution layer, and finally enhance the nonlinear characteristics of commodities through a sigmoid activation function to obtain the characteristic graph  $M_s \in R^{H \times W \times 1}$ , as follows:

$$M_s = \sigma \left\{ f_{Conv}^{1 \times 1} \left[ f_{Conv}^{5 \times 5} \left[ f_{Conv}^{7 \times 7} (F'_i) \right] \right] \right\} \quad (21)$$

where  $\sigma$  is a sigmoid function,  $f_{Conv}^{1 \times 1}$ ,  $f_{Conv}^{5 \times 5}$ ,  $f_{Conv}^{7 \times 7}$  represent the convolution layers, and  $F'_i$  is the attention output feature map of the SENet module.

The squeeze operation in SENet enlarged the global receptive field of a commodity, obtained the spatial feature information through maximum pooling calculation, and used a convolution kernel to ascend the dimensions to obtain a spatial attention feature map  $M_s$ , whose  $M_s \in R^{H \times W \times 1}$ . The formula used is as follows:

$$M_s = \sigma \left\{ f_{Conv}^{3 \times 3} \left[ \text{MaxPooling} (F') \right] \right\} \quad (22)$$

where MaxPooling represents maximum pooling and  $f_{Conv}^{3 \times 3}$  is  $3 \times 3$  convolution layer.

Finally, according to the obtained spatial attention feature map  $M_s$ , feature map  $F'_i$  under input the feature entropy is activated. By multiplying the valid feature map by the original feature map, the parameter quantity is reduced, and the valid feature map  $F''$  of the commodity is filtered through the linear normalization weight:

$$F'' = M_s \odot F'_i \quad (23)$$

where  $\odot$  is the commodity of the feature graphs by the elements.

The attention model is introduced into the residual network ResNet, which is first squeezed to compress the feature dimension, and then the ReLu activation function is added to the fully connected layer to complete the construction of the attention channel, and finally the feature weight of the channel is obtained through the Sigmoid function, and then the original channel dimension is weighted with the new

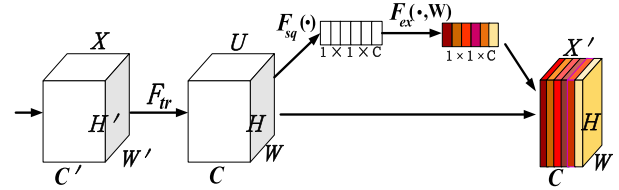


FIGURE 5. Squeeze-and-Excitation Module.

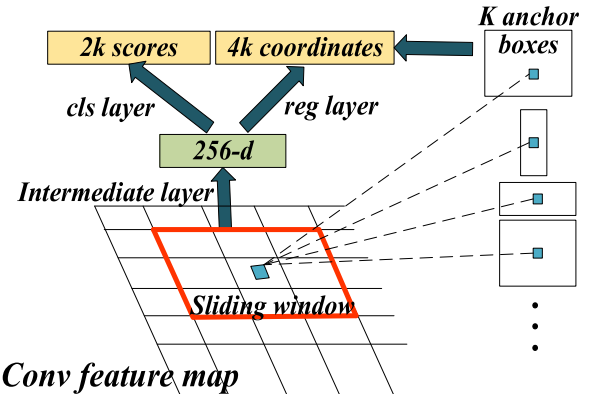


FIGURE 6. Schematic diagram of anchor construction.

channel, and finally the effective commodity features are output. The improved attention mechanism in this study has few parameters and good embedding, which can be quickly embedded in YOLO-R residual networks. The structure is shown in the Fig. 5.

### C. INSPECTION FRAME CONSTRUCTION

#### 1) STRICT DECISION RPN

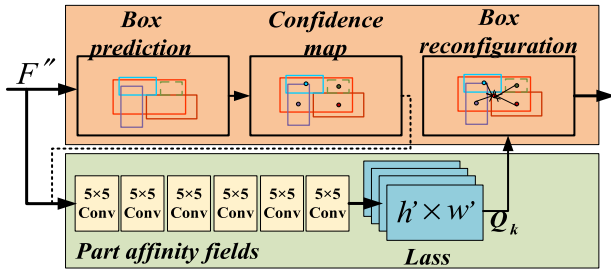
The core of the RPN [22], [23] is an anchor. The detection prediction box is generated using an anchor that could be used to select the location of the following detection box. In this study, the K-means clustering algorithm is used to adaptively generate the anchor parameters to improve the clustering effect.

Commodity detection is primarily used for the overlap between the prediction and real detection boxes and the overlap ratio referenced in this study. IOU obtains the similarity between both boxes according to the degree of overlap between them. In this study, a new measure based on the Dsoft-IOU is used. Its formula is as follows:

$$IOU(a, b) = \frac{|a \cap b|}{|a \cup b|} \quad (24)$$

$$d_i = \beta \sqrt{1 - IOU(b_{bbx}, c_{cluster, i})} \quad (25)$$

where  $a \cup b$  is the union area of frames  $a$  and  $b$ .  $a \cap b$  is the area where both frames intersected;  $d_i$  is the distance between the bounding box and the first cluster center,  $b_{bbx}$  is the bounding box,  $c_{cluster, i}$  is the  $i$  cluster center,  $\beta$  represents a coefficient. The influence of the IOU is amplified using a distance measurement formula, and the influence of the Euclidean distance variation is mitigated to generate more appropriate clustering results, as shown in Fig. 6.



**FIGURE 7.** Structure diagram of decision RPN. According to the pre-designed box, it is revolutionized pooled to obtain the final detection box center area.

Second, a loss function is combined with the IOU in this study to reduce the impact of the redundancy characteristics, and the following conclusions are made:

$$L_{IOU} = 1 - IOU(a, b) + \frac{d_i^2(a, b)}{c^2(a, b)} + \alpha\beta \quad (26)$$

where IOU represents the size of the merge ratio of the prediction and real boxes,  $\alpha$  is the similarity factor for measuring the aspect ratio, and  $c^2(a, b)$  is the square of the minimum diagonal length of the overlay rectangle.

$$\beta = \frac{\alpha}{(1 - IOU) + \alpha} \quad (27)$$

$$\alpha = \frac{4}{\pi^2} \left( \arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \quad (28)$$

where  $w^{gt}$  and  $h^{gt}$  are the true box width and height, respectively,  $w$  and  $h$  are the predicted box width and height, respectively. This process is illustrated in Fig. 7.

## 2) NONMAXIMAL SUPPRESSION BASED ON SCALE ESTIMATION

The NMS algorithm is a processing algorithm that deletes the redundant prediction boxes of a network. The box is scored according to the confidence level. Then, a bubble sort is performed according to the size of the confidence level score, and the IOU threshold is compared with the high-score prediction box. If the threshold is set higher, the prediction box is deleted and the maximum confidence score prediction box is not deleted. This is repeated until all the checkboxes are processed.

In this study, the DSoft-NMS algorithm is used. The algorithm is based on the Euclidean distance between the IOU true box and the prediction box. This is based on the IOU of the preselected box and the real box: the larger the confidence level of the detection box, the smaller the confidence level of the prediction box. The probability formula is derived based on the IOU crossover ratio as follows:

$$S_i = \begin{cases} S_i, & IOU(M, b_i) - \frac{d^2(b_i, M)}{c^2(b_i, M)} < w_i \\ S_i \left[ 1 - IOU(M, b_i) + \frac{d^2(b_i, M)}{c^2(b_i, M)} \right], & \\ IOU(M, b_i) - \frac{d^2(b_i, M)}{c^2(b_i, M)} \geq w_i \end{cases}, \quad (29)$$

$d(b_i, M)$  expressed as  $b_i$ ,  $M$  is the Euclidean distance, where  $b_i$  and  $M$  represent the prediction box and the center point of the optimal detection box, respectively;  $c(b_i, M)$  is the diagonal length between the center of the two boxes.  $w_i$  is the threshold.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the algorithm system, experimental parameters, and procedures, are described in detail. First, the overall process is introduced step-by-step. Subsequently, the key experimental parameters are determined during the training process. Finally, these indices are used to evaluate the occlusion detection results.

### A. COMPARISON WITH STATE-OF-THE-ARTS

The experiments are analyzed and compared using self-made smart retail container datasets. The experimental device used is the Lenovo Xiaoxin Air-12IIL 2020 with an Intel Core i5-1035G1 CPU and a discrete graphics card with NVIDIA GeForce MX3502GB. The system and software running the algorithm are Win10 and PyCharm, respectively. HF899 with a 2.7-mm (135 °distortion-free) camera is used.

A self-made dataset with common commodities in daily retail containers is used and 4,270 pieces of commodity data are collected. The target datasets are labeled according to the VOC2007 format. Among the 4,270 datasets, 3,843 pictures are used for training, and 427 for validation. The test commodity datasets are constructed separately.

### B. EXPERIMENTAL EVALUATION CRITERIA

Based on a review of various test studies, the average accuracy (AP) and the average AP (mAP) are used in this study to obtain the average value of the detection accuracy of commodity categories. Additionally, the F1 index is used as the evaluation standard for model stability [43]. The detected samples are considered positive when the confidence level of the commodity detection box is equal to the threshold value and negative when the confidence level of the commodity detection box is equal to or lower than the threshold value. The recall rate R is the percentage of the total sample of positive samples correctly tested, defined as

$$R = \frac{TP}{TP + FN} \quad (30)$$

The TP is the number of samples correctly classified as positive. FN is the total number of positive samples incorrectly identified as negative samples. T represents the maximum spacing between video frames.

Precision P represents the proportion of positive samples detected by the algorithm to the total number of positive samples detected by the detection result, which is defined as

$$P = \frac{TP}{TP + FP} \quad (31)$$

Set the total number of samples to  $n$ , and detect  $k$  samples. The completion rate is expressed as  $r_k$ , and  $p_k$  is the



maximum accuracy rate that is greater than  $r_k$ . The average accuracy is defined as:

$$AP = \sum_{k=1}^n p_k(r_{k+1} - r_k) \quad (32)$$

mAP is the average accuracy for all categories, defined as

$$mAP = \frac{1}{L} \sum_{L=1}^L AP_q \quad (33)$$

where  $L$  represents the total number of categories, and  $AP_q$  represents the average accuracy of category  $q$ .

The model stability is detected using the F1 value (H-mean value), which is obtained by dividing the arithmetic mean by the geometric mean. The F1 value is inversely proportional to model stability. The formula used is as follows:

$$F_1 = \frac{2PR}{P+R} = \frac{2TP}{2TP+FP+FN} \quad (34)$$

The formula shows that F1 is the weighted summation of the precision and recall, expressed as a harmonic mean.

### C. ANALYSIS OF EXPERIMENTAL RESULTS

In this study, a semantic segmentation algorithm is used to enhance the semantic information of the commodities and improve the accuracy of the data. Subsequently, the commodities are detected using YOLO-R, an algorithm that semantically split the network association, which constitutes the overall process. The test results for this process are as follows.

The results of the algorithm detection are shown in Fig. 8. The first, second, and third rows represent the original image, semantically segmented prediction results, and detection results, respectively. It can be observed from Fig. 8 that in all semantic segmentation maps of commodities, the commodities are separated from the background, and the algorithm accurately detected obscured commodities.

At the beginning of this study, the ideal number of iterations is determined to be 100 based on training sessions conducted for each comparison model. Furthermore, the stability of the proposed model is compared to that of the traditional models, as shown in Fig. 9.

#### 1) PUBLIC DATASET DETECTION AND COMPARISON

This study first verifies the performance of the algorithm. The algorithm uses in this study and the traditional algorithm is trained and verified using the VOC2007 public dataset. The data contains 21,503 pictures, which are divided into a training set and a verification set through cross-validation at 9:1.

Table 1 uses the mAP and F1 data obtained by VOC2007, and according to the literature, YOLOX [44] and DETR [45] models are introduced for comparative analysis, and the comprehensive analysis shows that the detection accuracy of mAP is similar, and the higher the F1 index of the model, the stronger the stability of the model. Therefore, this paper uses a more mature YOLO5 model to improve the model in this study, and the performance of YOLO5 is more stable than other networks.

**TABLE 1. Comparison of public dataset model data. The best score is highlighted in bold.**

Model	Backbone	mAP	F1
SSD	Resnet50	80.58%	74.50%
Centernet	Resnet50	84.22%	72.60%
Retinanet	VGG	84.83%	80.45%
Faster-RCNN	VGG	87.83%	72.95%
YOLO4	CSPDarknet53	84.89%	78.90%
YOLO5	CSPDarknet53	87.27%	<b>83.50%</b>
YOLO7	CSPDarknet53	81.00%	76.15%
YOLOX [44]	CSPDarknet53	86.42%	79.60%
DETR [45]	Resnet50	<b>90.94%</b>	67.05%

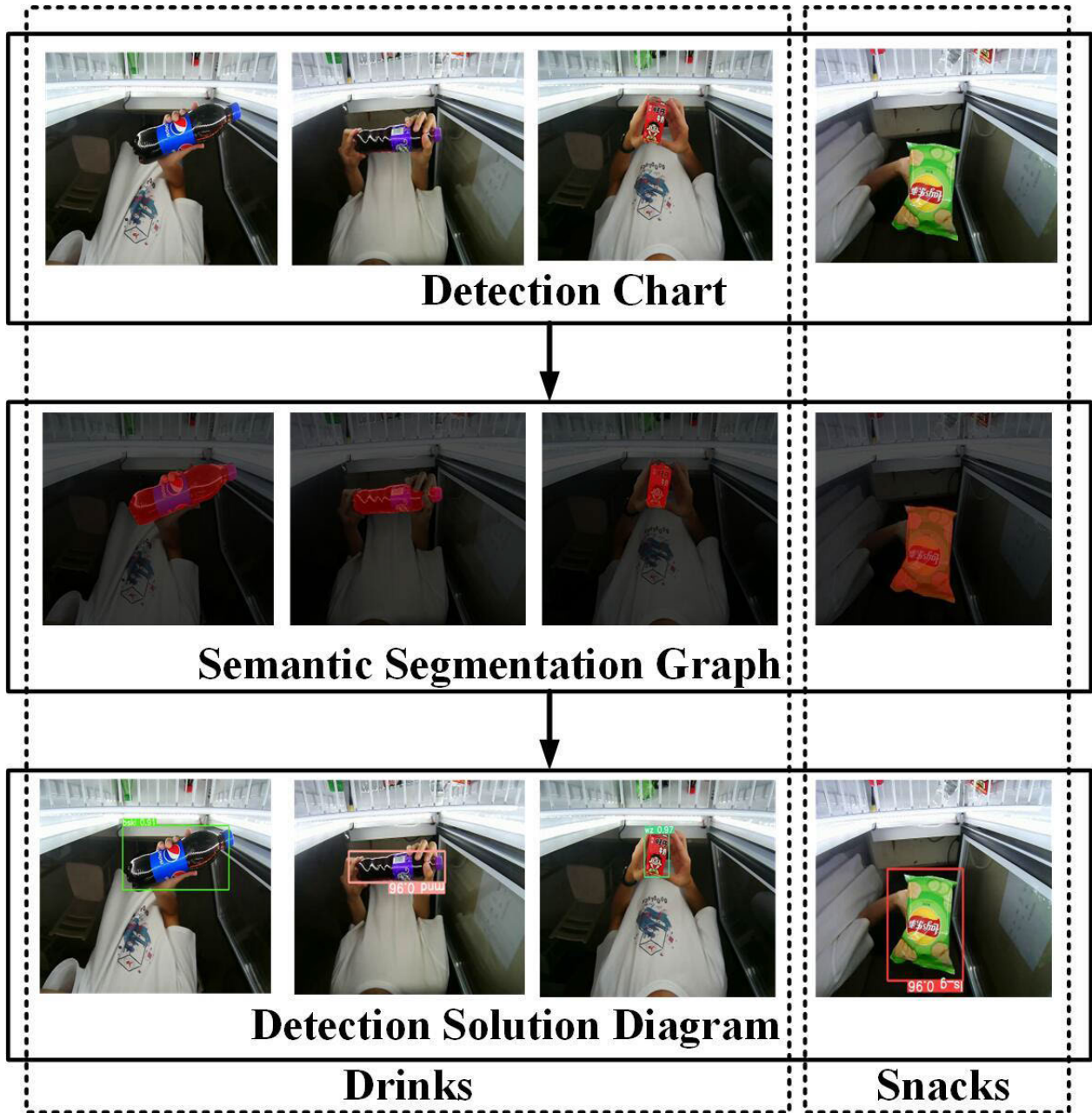
**TABLE 2. Comparison of average accuracy, map value and recall rate on self-made commodity datasets. The best score is highlighted in bold.**

Method	mAP	Precision	Recall	Speed (F/s)
YOLO4	95.43%	97.50%	95.76%	3.36
YOLO5	97.40%	95.70%	99.03%	18.89
YOLO7	96.50%	96.40%	95.00%	20.24
SSD	85.63%	85.35%	90.12%	4.07
Retinanet	97.13%	97.01%	98.06%	2.63
Centernet	94.12%	97.09%	90.17%	3.77
Faster-RCNN	97.74%	96.60%	97.80%	1.67
YOLOX [44]	97.79%	<b>98.52%</b>	98.31%	3.03
DETR [45]	97.62%	97.60%	97.80%	1.82
YOLO-R(ours)	<b>97.80%</b>	97.80%	<b>99.93%</b>	<b>22.72</b>

Note: Speed =  $\frac{\text{Detect the number of video frames(Frames)}}{\text{time(s)}}$ .

#### 2) COMPARISON OF SELF-MADE DATASET DETECTION MODELS

In Table 2, The results are based on 100 training iterations. This study refers to the current mainstream comparative experimental models, and finds that the proposed model is superior to other network models in AP and Recall, among which the YOLO-R model is 0.01% higher than YOLOX in AP, and the YOLO-R model is better than DETR, but the accuracy is slightly lower than YOLOX. In terms of commodity detection speed, the detection speed of the original model YOLO5 is lower than that of YOLO7, but the YOLO-R improved by the algorithm in this paper increases the speed by 3.83 F/s, and the improved YOLO-R detection speed is significantly better than YOLO7, and 2.48F/s faster than YOLO7. YOLO-R also improves the detection accuracy by 0.9% compared to the original model and the speed by 3.83F/s. This comparative experiment verifies the feasibility and superiority of the proposed algorithm. In the self-made dataset, the improved model accuracy in this paper is more accurate, and the comprehensive performance has also been improved to a certain extent.



**FIGURE 8.** The results are output during the experiment. First line: Input masking Commodities. Second line: Output semantic segmentation result graph. Third line: Output test result graph.

Compared with other algorithms, YOLO is an end-to-end target-detection neural network. YOLO predicted multiple candidate boxes at one time, regressed the object location area and the category of objects in the area at the output layer, and is faster. The faster-RCNN and other algorithms must generate several candidate frames in the picture and they have several parameters and a long training time.

### 3) COMPARISON AND ANALYSIS OF DIFFERENT COMMODITY TESTS ON SELF-MADE DATASETS

The occlusion detection performance of the proposed model is compared with those of other networks based on AP using

test sets of different commodities. By comparing the detection effect of each network on other datasets, the stability of the proposed model is verified. Simultaneously, the universality and stability of the algorithm model are demonstrated, proving its applicability to different commodity detections, as shown in Table 3.

As shown in Table 3, The inspection progress of YOLO-R on different commodities is higher than that of other models, but the detection accuracy of YOLO7 models on Scream commodity is **0.02%** lower, and the overall detection accuracy is **2.69%** higher than that of YOLO7. The detection accuracy of the improved model on different commodities is **0.05%**–

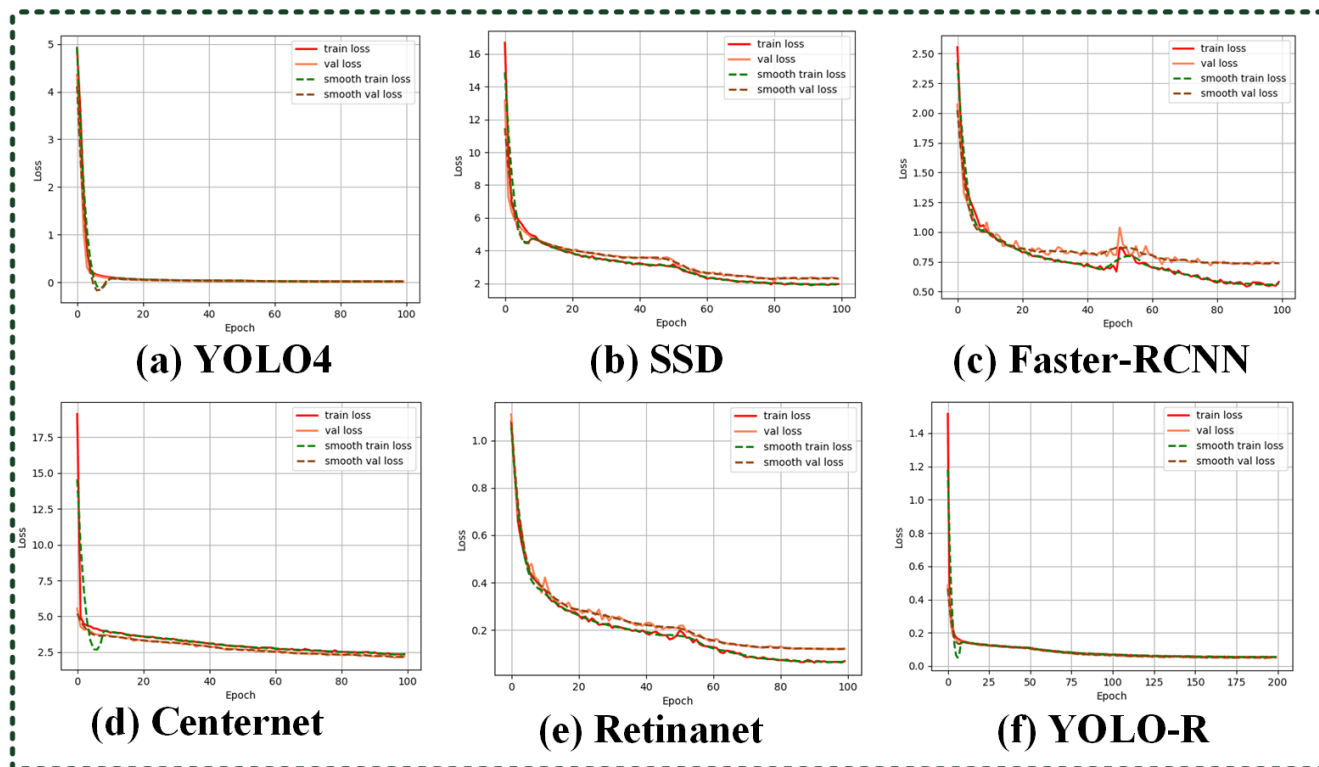


FIGURE 9. Iterative stability of loss training. In (a)(b)(c)(d)(e)(f), the network models are stable after 100 trainings.

TABLE 3. Comparison of various network models in different commodity detection. The best score is highlighted in bold.

Categories	AP					
	YOLO4	Retinanet	SSD	YOLO7	YOLO5	YOLO-R
Pepsi Cola	83.42%	80.23%	63.75%	82.81%	80.67%	<b>83.60%</b>
Cestbon	80.11%	87.23%	80.18%	81.21%	84.38%	<b>88.78%</b>
Lays	82.56%	76.60%	86.22%	88.16%	88.56%	<b>89.67%</b>
Sprite	65.47%	73.55%	66.30%	73.78%	74.60%	<b>74.65%</b>
Scream	67.36%	69.73%	69.21%	<b>69.91%</b>	67.69%	69.89%
Spring	86.62%	90.81%	90.81%	89.80%	92.83%	<b>95.25%</b>
Average Precision	77.59%	79.69%	79.69%	80.95%	81.46%	<b>83.64%</b>

2.42% higher than that of the second-best model. In the selected comparison model, the overall performance of the improved YOLO-R commodity detection method improved by 2.18% compared to the original model. In conclusion, the model outperforms the general approach.

#### 4) COMPARATIVE ANALYSIS OF NETWORK MODEL STABILITY ON THE HOMEMADE DATASETS

A series of comparisons are also made for commodities with different shielding to verify the stability of the model. Two representative commodities (packed potato chips and bottled mineral water) are shielded to different degrees and their average accuracies are obtained. The comparison analysis is based on the average accuracy.

In Table 4, to compare the stability of the network models, we select the packed potato chips and bottled water are selected as the experimental objects and shield the two commodities are shielded to varying degrees to detect the accuracy of each network model. From the analysis of the experimental results, it can be concluded that the detection accuracy of the proposed algorithm is not different from that of the other network models when there is little or no occlusion. However, the detection performance of the model decreased with an increase in the occlusion. Nevertheless, with an increase in the occlusion ratio, the detection performance of the YOLO-R model remained high, indicating greater stability on the self-made occlusion dataset. The detection accuracy for potato chips reached an astonishing 97.88% at

**TABLE 4.** The stability between network models. The best score is highlighted in bold. This data indicates the accuracy of commodity detection.

Occlusion degree		No occlusion	Slight occlusion	Moderate occlusion	Severe occlusion
YOLO5	Lays	90.89%	83.94%	74.06%	72.06%
	Spring	94.17%	92.83%	75.89%	52.75%
SSD	Lays	97.89%	87.50%	79.44%	65.42%
	Spring	93.44%	82.21%	78.14%	66.25%
YOLO4	Lays	93.17%	93.00%	62.05%	67.41%
	Spring	97.83%	97.72%	74.62%	67.79%
Retinanet	Lays	99.17%	99.01%	78.20%	74.92%
	Spring	<b>99.88%</b>	98.94%	77.50%	66.24%
Faster-RCNN +attention	Lays	99.11%	99.06%	98.61%	96.53%
	Spring	99.44%	97.06%	<b>89.56%</b>	67.69%
YOLO-R (ours)	Lays	<b>99.94%</b>	<b>99.87%</b>	<b>99.08%</b>	<b>97.88%</b>
	Spring	99.50%	<b>99.06%</b>	73.08%	<b>68.88%</b>

**TABLE 5.** Comparing models of attention mechanisms. The best score is highlighted in bold.

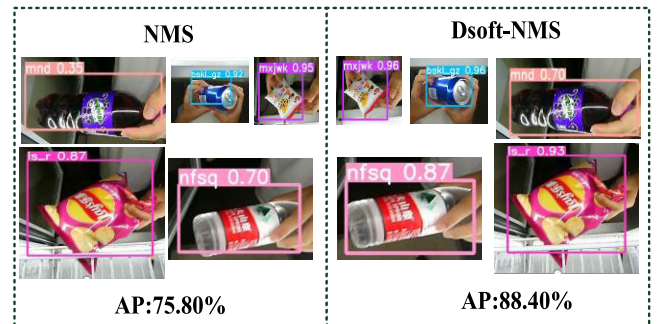
Model	F1	mAP	Predict	speed (F/S)
YOLO	97.71%	97.40%	95.70%	19.94
YOLO+CA	97.69%	97.30%	95.60%	5.16
YOLO+CBAM	97.67%	97.30%	95.72%	19.53
YOLO+ECA	97.72%	97.60%	95.10%	20.12
YOLO+SE	97.77%	97.60%	96.50%	20.09
YOLO-R	<b>97.79%</b>	<b>97.80%</b>	<b>96.70%</b>	<b>22.32</b>

approximately 60% of the serious occlusion degree, which is 1.35% higher than that of the second-best network model. For the detection of Farmer Spring, although the faster R-CNN is not better in the presence of moderate occlusion, the overall performance is better. For severe occlusions, the detection accuracy of YOLO-R reached 68.87% and 1.08% higher than that of the second-best network, respectively.

5) COMPARISON OF ATTENTION MECHANISMS ABLATION EXPERIMENTS

To identify the attentional mechanism that best matches the proposed model, numerous studies are consulted and three representative attentional models are selected for ablative comparison with the attentional model in this study.

As shown in Table 5, the YOLO algorithm of the improved SE attention mechanism is 0.2% better than that of the original model. The ECA attention mechanism network appears to be similar to the improved SE model during the analysis process. The focus of this study is the comparison and validation of the actual commodity detection of each network model. Through a comparative analysis, YOLO-R is found to



**FIGURE 10.** Comparing the improved NMS algorithm with the original one. We use different types of commodities, compare the results of testing, and solve the average detection accuracy of these commodities.

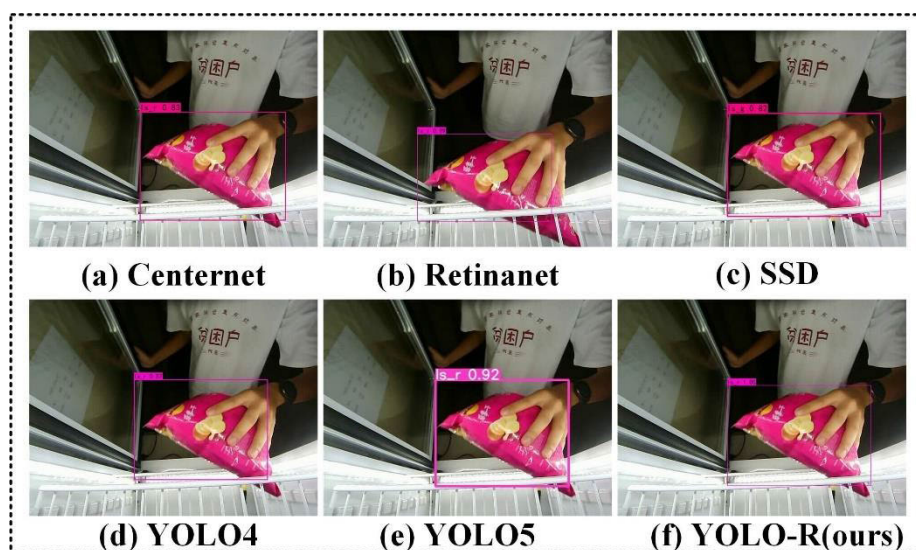
improve both accuracy and speed, with 0.40% accuracy and 2.38 F/s.

Table 6 presents an extension of the ablation experiments described previously. Several representative commodities are screened in extensive experiments and tested to verify the superiority of the improved attention mechanism.

The four model attention mechanisms are compared, and it is found that the improved SE attention mechanism in the study is more stable, and the detection accuracy is approximately 0.2% higher in terms of the mAP, compared to when no attention mechanism is used. In Table 6, six commodities (the test datasets in Tables 6 and 3 are not the same) are selected to verify the detection accuracy of the commodities randomly sampled from the self-made dataset used in this study. The average accuracy is the average value of the detection accuracy for the commodities mentioned above. As shown in Table 5, the average detection accuracy improved when the attention mechanism is incorporated. The greatest improvement is observed with YOLO with the SE

**TABLE 6.** Commodity detection accuracy of YOLO model in the self-made dataset. The best score is highlighted in bold.

Categories	AP					
	YOLO	YOLO +CA	YOLO +CBAM	YOLO +ECA	YOLO +SE	YOLO-R
Panpan Snack	93.60%	95.67%	95.73%	95.00%	<b>95.79%</b>	95.33%
Lays	88.56%	93.38%	92.20%	91.72%	92.06%	<b>93.39%</b>
Coca Cola (Small)	85.72%	90.50%	90.06%	91.17%	91.86%	<b>92.75%</b>
NongFu Spring	91.77%	<b>93.15%</b>	92.15%	87.85%	93.07%	<b>93.15%</b>
Meida	88.65%	85.33%	87.20%	88.89%	88.90%	<b>92.20%</b>
Pepsi Cola (Canned)	88.00%	93.25%	91.80%	91.71%	93.69%	<b>94.29%</b>
Average Precision	89.38%	91.88%	91.52%	91.06%	92.56%	<b>93.52%</b>



**FIGURE 11.** Detection results of common network models and this paper's network model on Ritz-Carlton potato chips.

attention mechanism model, whose detection accuracy is 3.18% higher than that of YOLO. On this basis, the YOLO-R accuracy of the improved algorithm is 4.14% higher than that of the original model. Tables 5 and 6 show that the algorithm is faster, more accurate, and has better detection stability.

6) INNOVATIVE COMPARATIVE ABLATION EXPERIMENT

The BiFPN feature fusion model is used to replace the original PANet fusion structure and to reduce the model parameters in the BiFPN layer. Second, the detection box algorithm is improved and the threshold limit of the eigenvalue is increased so that a suitable detection box could be generated. Therefore, an innovation point verification is conducted in this study.

As indicated in Table 7, the lightweight BiFPN in this study has the highest stability detection accuracy, and the stability and detection accuracy are higher. The lightweight network model is 0.01-0.02 higher than other networks on the F1 index and **0.3%-0.4%** higher than other networks on mAP.

To verify that the proposed NMS algorithm for the Dsoft-IOU is better in terms of the accuracy of the con-

**TABLE 7.** Comparison of YOLO-R's multi-feature fusion module network. The best score is highlighted in bold.

Model	F1	mAP
YOLO (PANET)	97.71%	97.40%
YOLO (BiFPN)	97.73%	97.50%
YOLO (Lightweight BiFN)	<b>97.74%</b>	<b>97.80%</b>

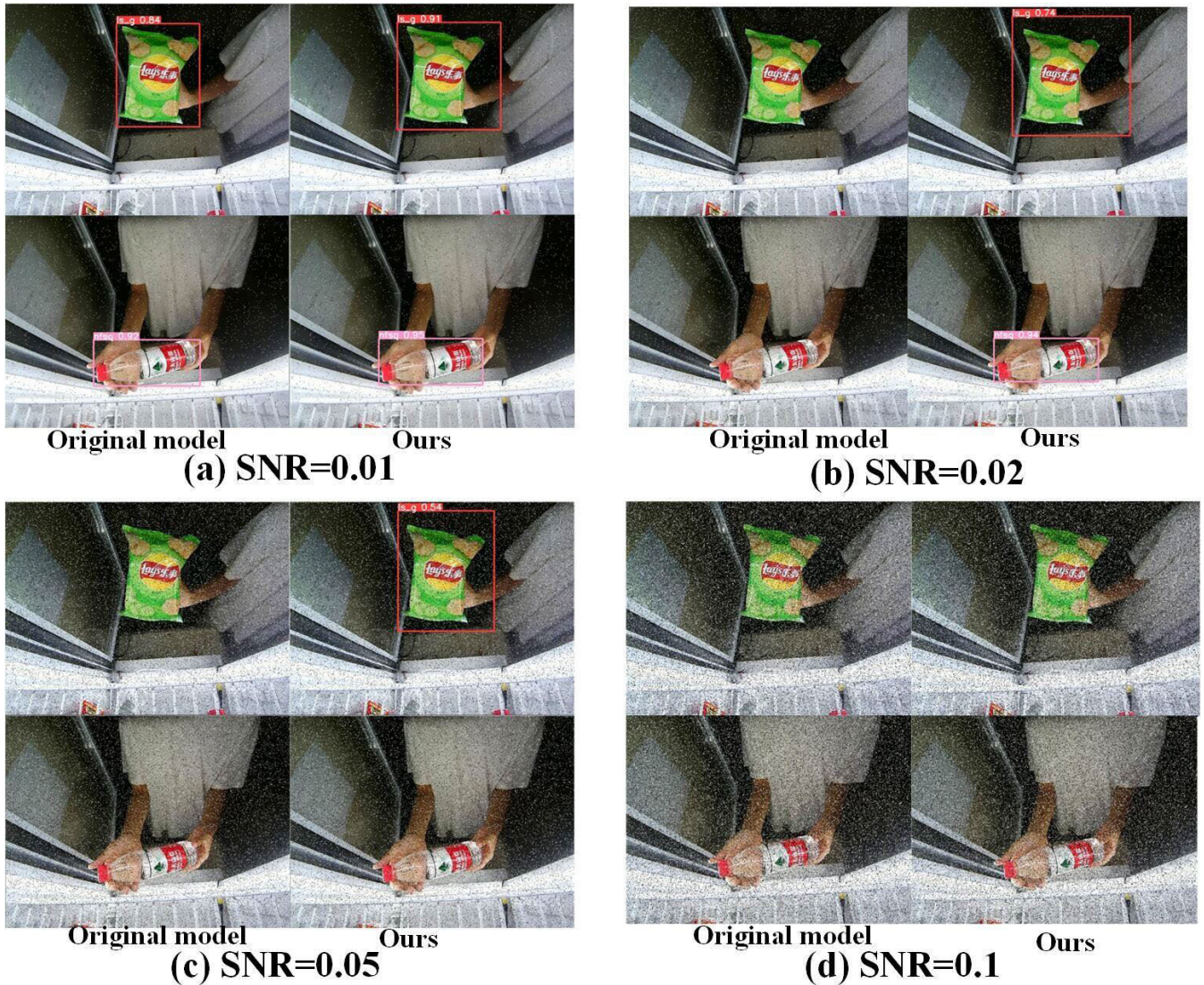
struction of the detection box and its size, a similar model structure is guaranteed in this study, and the NMS algorithm is modified for comparison experiments, as shown in Fig. 10.

The improved NMS algorithm produced a more accurate detection box, located the detection box more accurately, effectively contained the detected commodities, improved the detection accuracy, and had a better detection effect.

D. ALGORITHM DETECTION RESULT GRAPH

1) DETECTION RESULTS OF DIFFERENT ALGORITHMS

To make this article more convincing, the next section presents a visualization of the test results. The results of the



**FIGURE 12.** Interference immunity comparison chart. (a) When the signal-to-noise ratio is 0.01, our model is compared with the original model. (b) (c) (d)When the signal-to-noise ratio is 0.02, 0.5, and 0.1, our model and the original model are compared for detection.

six comparison network detection models selected in this study are shown in Fig. 11 to further illustrate the detection effect characteristics of the proposed algorithm. The algorithm is effective in detecting occluded images in self-made datasets. This can solve the problem of occlusion when shoppers use commodities in a container. The following is an experimental analysis and detection result diagram of several algorithms. In the diagram, bagged potato chips are used as the detection object to facilitate the comparison and detection of various models.

This section also adds a comparative visualization of commodities detection models under different signal-to-noise ratios. As Fig. 12.

The algorithm results are displayed and analyzed Fig. 11 shows the detection results of the proposed model and the other network models. In this study, a self-made occluded commodity dataset is used for comparative experiments, and

compared with other models, the proposed model achieves a high detection accuracy.

According to the citation of relevant literature [46], the detection and comparative analysis of goods under different noise conditions are carried out. In Fig. 12, our model has better detection results at different signal-to-noise ratios than the original model, but when SNR=0.1, neither our model nor the original model can detect the commodities. This experiment can verify that the algorithm in this paper is more stable through comparison. This signal-to-noise ratio experiment proves that the anti-interference ability of the improved algorithm has been improved, but the anti-interference ability and YOLO-R anti-interference ability need to be improved.

## 2) SELF-MADE DATASETS TESTED IN THIS MODEL

The dataset used in this study is a self-made commodity dataset. The commodity commonly used in intelligent retail



**FIGURE 13.** Display of three categories of self-made datasets. We categorize the datasets into three categories of commodities, and select some of them for visualization and detection.

containers are selected through big data detection and divided into three categories, depending on whether they are bagged, bottled, or canned. A few commodities are selected from these three categories, and the inspection experiment is visualized.

The type of dataset used in this study is similar to the displayed sample data. There are many kinds of commodities, and the degree of occlusion is determined by the proportion of occlusion.

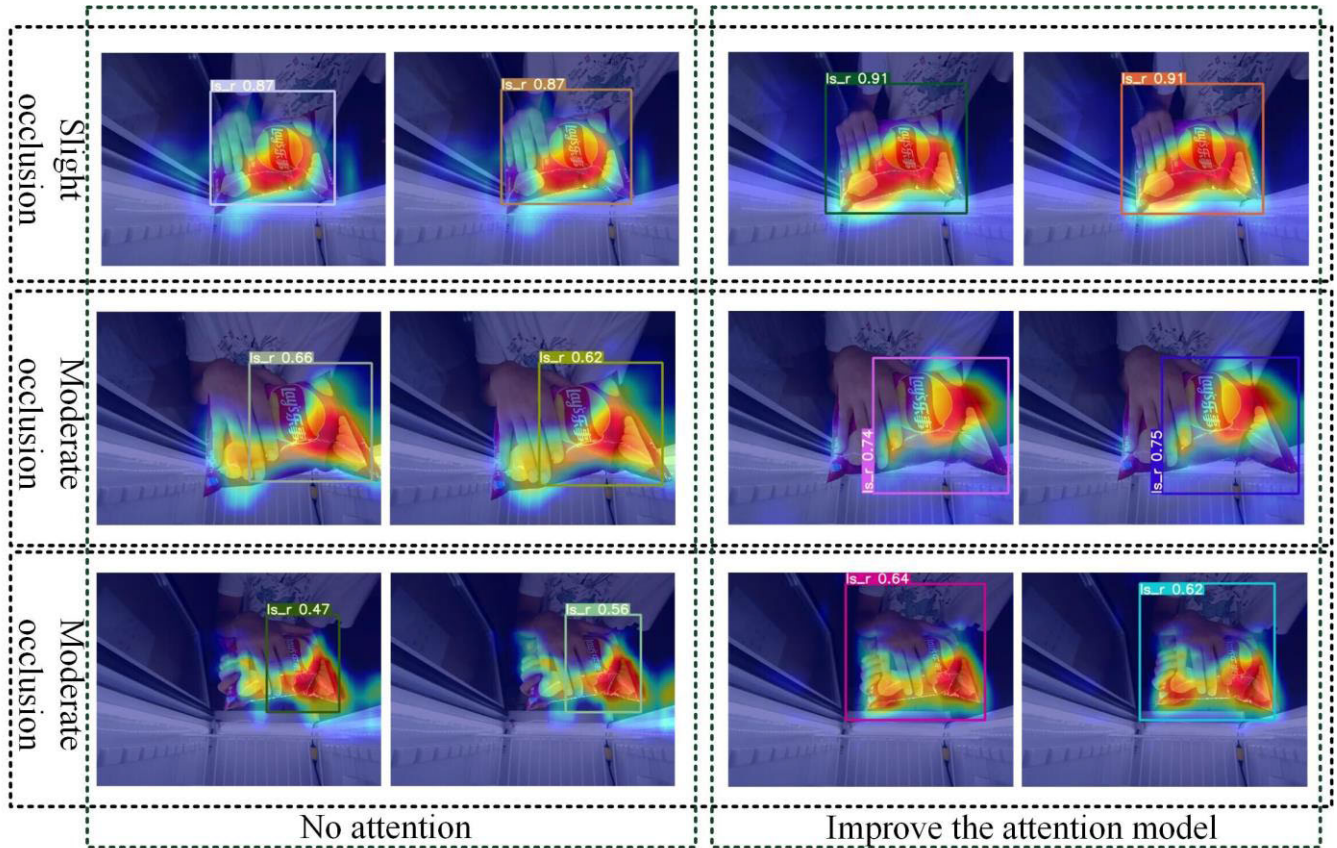
Twenty-one types of commodity datasets are used in this study. Twelve categories of commodities are presented in the homemade datasets and are divided into three categories.

Using a free graph, it is proven that the algorithm is effective on self-made datasets.

### 3) ATTENTION MECHANISM THERMAL VISUALIZATION

In this section, a visualization of the attention mechanism is presented, as shown in Fig. 14. This section further demonstrates that the YOLO model enhances the network’s ability to extract the signs of the model and could focus on the effective feature areas of commodities very well.

The graph shows the thermographic display and detection accuracy of YOLO-R at different occlusion levels and a



**FIGURE 14.** Attention visualization and comparison. First line: Visualization of lightly occlusion heatmaps. Second line: Medium occlusion visualization. Third line: Heavily occlusion heatmap visualization.

comparison of the proposed model with and without an attention mechanism network. As can be observed from Fig. 14, the algorithm focuses on increasing the attention on the unobscured part to make the feature extraction more effective. The detection accuracy decreased with an increase in the occlusion degree, the strengthened attention mechanism network became more effective in the area of interest, and the color is more in-depth. Each test commodity is the same as the training set commodity, and the commodity that obtained the AP from the test is the result of the video-frame excerpt test. Through a comparative analysis, it can be concluded that the attention mechanism model in this study focused more on the unobstructed features of commodities and is more accurate for detection.

## V. CONCLUSION AND FUTURE WORK

To solve the issue of large changes in the target scale, multiple occlusion cases, and target detection accuracy in the process of occlusion commodity detection, a residual network combined with an attention module is proposed to enhance the range of field scale and enhance the multiscale information fusion ability of the model, thus improving the detection accuracy of the model. To address the insufficient feature fusion in YOLO-R and the mix of multilayer features, the BiFPN feature pyramid is used in this approach.

The feature pyramid is sampled as a CARAFE structure, which enlarges the sensing field, fused features at different scales of the network structure, and enhances the robustness of the features. For the processing module of the YOLO-R model, a Dsoft-IoU loss regression module that combines the location information and characteristic entropy threshold is proposed to adaptively adjust the model's Dsoft-IoU, thus, preventing the real detection box from being filtered and improving the prediction accuracy of the model.

The method is tested by masking the homemade datasets. The results show that the improved YOLO-R based on YOLO and the SENet occlusion detection method combined with the attention mechanism uses the eigenvalue to limit the threshold value, thereby increasing the attention model. The mean average accuracy obtained using this method is higher than that obtained using YOLO, and the speed is also improved. The algorithm also achieves good detection results for the commodities with different occlusion ratios. However, the proposed algorithm, similar to the traditional algorithm, has certain limitations. Compared with the original method, the noise immunity of the improved algorithm is significantly enhanced, but with the increase in noise ratio, the commodity detection ability decreases significantly. The algorithm in this study is insufficient in the low anti-interference ability of noise and low detection accuracy. Therefore, in future work,



this paper focuses on optimizing the model network structure, improving the noise resistance and anti-interference ability of the model, enhancing the generalization of the model, and improving the stability of model detection.

## AUTHOR CONTRIBUTIONS

An Xie conceived algorithms of the paper and write the manuscript, Kai Xie reviewed the paper, Hao-Nan Dong and Kai Xie designed experiments, Hao-Nan Dong conducted comparative experiments and collected data, Jian-Biao He checked spelling and grammar and made suggestions.

## REFERENCES

- [1] C. Bhagya and A. Shyna, "An overview of deep learning based object detection techniques," in *Proc. 1st Int. Conf. Innov. Inf. Commun. Technol. (ICICT)*, Apr. 2019, pp. 1–6.
- [2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
- [3] Q. Wei, X. Hu, X. Wang, and H. Wang, "Improved RetinaNet target detection model," in *Proc. 2nd Int. Conf. Algorithms, High Perform. Comput. Artif. Intell. (AHPICAI)*, Oct. 2022, pp. 470–476.
- [4] J. Wang, T. Xiao, Q. Gu, and Q. Chen, "YOLOv5\_CSL\_F: YOLOv5's loss improvement and attention mechanism application for remote sensing image object detection," in *Proc. Int. Conf. Wireless Commun. Smart Grid (ICWCSG)*, Aug. 2021, pp. 197–203.
- [5] S. Liu, Y. Wang, Q. Yu, H. Liu, and Z. Peng, "CEAM-YOLOv7: Improved YOLOv7 based on channel expansion and attention mechanism for driver distraction behavior detection," *IEEE Access*, vol. 10, pp. 129116–129124, 2022.
- [6] A. Iriani Sapitri, S. Nurmaini, M. N. Rachmatullah, B. Tutuko, A. Darmawahyuni, F. Firdaus, D. P. Rini, and A. Islami, "Deep learning-based real time detection for cardiac objects with fetal ultrasound video," *Informat. Med. Unlocked*, vol. 36, Jan. 2023, Art. no. 101150.
- [7] M. R. Javed and M. Shamsuzzaman, "YOLOBin: Non-decomposable garbage identification and classification based on YOLOv7," *J. Comput. Commun.*, vol. 10, no. 10, pp. 104–121, 2022, doi: [10.4236/jcc.2022.1010008](https://doi.org/10.4236/jcc.2022.1010008).
- [8] D. Wu, S. Jiang, E. Zhao, and Y. Liu, "Detection of *Camellia oleifera* fruit in complex scenes by using YOLOv7 and data augmentation," *Appl. Sci.*, vol. 12, no. 22, 2022, Art. no. 11318.
- [9] Z. Yang, C. Zhao, H. Maeda, and Y. Sekimoto, "Development of a large-scale roadside facility detection model based on the Mapillary dataset," *Sensors*, vol. 22, no. 24, p. 9992, Dec. 2022.
- [10] K. Chen, G. Yan, M. Zhang, Z. Xiao, and Q. Wang, "Safety helmet detection based on YOLOv7," in *Proc. 6th Int. Conf. Comput. Sci. Appl. Eng.*, Oct. 2022, pp. 1–6.
- [11] C. Dewi, A. P. S. Chen, and H. J. Christanto, "Deep learning for highly accurate hand recognition based on YOLOv7 model," *Big Data Cognit. Comput.*, vol. 7, no. 1, p. 53, Mar. 2023.
- [12] T. Sledevic and D. Plonis, "Toward bee behavioral pattern recognition on hive entrance using YOLOv8," in *Proc. IEEE 10th Jubilee Workshop Adv. Inf., Electron. Electr. Eng. (AIEEE)*, Apr. 2023, pp. 1–4.
- [13] G. Zhang, H. Zhang, Y. Yao, and Q. Shen, "Attention-guided feature extraction and multiscale feature fusion 3D ResNet for automated pulmonary nodule detection," *IEEE Access*, vol. 10, pp. 61530–61543, 2022.
- [14] M. F. Haque, H.-Y. Lim, and D.-S. Kang, "Object detection based on VGG with ResNet network," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Jan. 2019, pp. 1–3.
- [15] A. Aditya, L. Zhou, H. Vachhani, D. Chandrasekaran, and V. Mago, "Collision detection: An improved deep learning approach using SENet and ResNext," in *Proc. IEEE Int. Conf. Syst. Man, Cybern. (SMC)*, Oct. 2021, pp. 2075–2082.
- [16] W. Liu and S. Li, "Human motion target recognition using convolutional neural network and global constraint block matching," *IEEE Access*, vol. 8, pp. 69378–69388, 2020.
- [17] H. Fang, M. Xia, G. Zhou, Y. Chang, and L. Yan, "Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [18] M. M. Dasari and R. K. S. S. Gorthi, "IOU—Siamtrack: IOU guided Siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2061–2065.
- [19] H. Peng and S. Yu, "A systematic IoU-related method: Beyond simplified regression for better localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5032–5044, 2021.
- [20] A. Ahmadvadeh, D. J. Kempton, Y. Chen, and R. A. Angryk, "Multi-scale IOU: A metric for evaluation of salient object detection with fine structures," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 684–688.
- [21] Z. Hou, X. Cai, S. Chen, and B. Li, "A model based on dual-layer attention mechanism for semantic matching," in *Proc. IEEE Int. Conf. Intell. Appl. Syst. Eng. (ICIASE)*, Apr. 2019, pp. 105–108.
- [22] G. Liu, C. Wang, and Y. Hu, "RPN with the attention-based multi-scale method and the adaptive non-maximum suppression for billboard detection," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2018, pp. 1541–1545.
- [23] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4012–4021.
- [24] D. Liu and F. Cheng, "SRM-FPN: A small target detection method based on FPN optimized feature," in *Proc. 18th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2021, pp. 506–509.
- [25] A. Kumar, G. Brazil, and X. Liu, "GroomMeD-NMS: Grouped mathematically differentiable NMS for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8969–8979.
- [26] N. Jia, "Design and implementation of smart shopping management system based on RFID," *Comput. Eng.*, vol. 41, no. 9, pp. 25–30 and 38, 2015.
- [27] W. Jun-Fei, Z. Ting-Gang, S. Jing-Wei, D. Jin-Fang, Y. Hua, and S. Chen-Yang, "Research and implementation of an indoor positioning and navigation method based on 2D bar code and geodatabase—Taking supermarket shopping-guide as an example," *J. Southwest Univ., Natural Sci. Ed.*, vol. 36, no. 11, pp. 209–214, 2014.
- [28] L. Liu, J. Cui, Y. Huan, Z. Zou, X. Hu, and L. Zheng, "A design of smart unmanned vending machine for new retail based on binocular camera and machine vision," *IEEE Consum. Electron. Mag.*, vol. 11, no. 4, pp. 21–31, Jul. 2022.
- [29] M. Hu and X. Zhong, "Fast-speed image recognition system on retail commodity image," in *Proc. IEEE Int. Conf. e-Bus. Eng. (ICEBE)*, Nov. 2021, pp. 76–81.
- [30] X. Zhao and Y. Wu, "Automatic motion-blurred hand matting for human soft segmentation in videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1450–1454.
- [31] S. Namiki, K. Yokoyama, S. Yachida, T. Shibata, H. Miyano, and M. Ishikawa, "Online object recognition using CNN-based algorithm on high-speed camera imaging: Framework for fast and robust high-speed camera object recognition based on population data cleansing and data ensemble," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2025–2032.
- [32] Y. Chen, "Human behavior recognition based on deep learning," in *Proc. 2nd Int. Seminar Artif. Intell., Netw. Inf. Technol. (AINIT)*, Oct. 2021, pp. 88–91.
- [33] Y. Shi, B. Zhou, and C. Li, "Parallel optimization of depth learning algorithm based on behavior recognition," in *Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS)*, Jan. 2019, pp. 660–664.
- [34] C. I. Orozco, M. E. Buemi, and J. J. Berlles, "Towards an attention mechanism LSTM framework for human action recognition in videos," in *Proc. IEEE Congreso Bienal de Argentina (ARGENCON)*, Dec. 2020, pp. 1–6.
- [35] Z. Luo, J. Li, and Y. Zhu, "A deep feature fusion network based on multiple attention mechanisms for joint iris-periodic biometric recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 1060–1064, 2021.
- [36] C. Xiu, X. Su, and X. Pan, "Improved target tracking algorithm based on Camshift," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2018, pp. 4449–4454.
- [37] H. Wang, Q. Zhang, L. Yu, and Z. Wang, "Research on CAMshift algorithm based on feature matching and prediction mechanism," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2021, pp. 773–778.

- [38] T. Liang, H. Bao, W. Pan, X. Fan, and H. Li, "DetectFormer: Category-assisted transformer for traffic scene object detection," *Sensors*, vol. 22, no. 13, p. 4833, Jun. 2022.
- [39] T. Liang, H. Bao, W. Pan, and F. Pan, "Traffic sign detection via improved sparse R-CNN for autonomous vehicles," *J. Adv. Transp.*, vol. 2022, pp. 1–16, Mar. 2022.
- [40] Y. Tian, Y. Zhang, and L. Li, "Assessment method of fusion image quality based on region entropy," in *Proc. IEEE 15th Int. Conf. Electron. Meas. Instrum. (ICEMI)*, Oct. 2021, pp. 1–4.
- [41] H. Wang, Y. Li, and S. Wang, "Fast pedestrian detection with attention-enhanced multi-scale RPN and soft-cascaded decision trees," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 12, pp. 5086–5093, Dec. 2020.
- [42] H. Wang and P. Xie, "Research on facial feature extraction method based on semantic segmentation," in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, vol. 6, Mar. 2022, pp. 787–790.
- [43] S.-E. Ryu and K.-Y. Chung, "Detection model of occluded object based on YOLO using hard-example mining and augmentation policy optimization," *Appl. Sci.*, vol. 11, no. 15, p. 7093, Jul. 2021.
- [44] Z. Ge, S. T. Liu, F. Wang, and Z. M. Li, "YOLOx: Exceeding YOLO series," 2021, *arXiv:2107.08430*.
- [45] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," 2022, *arXiv:2203.16527*.
- [46] L. Jia, S. Song, L. Yao, H. Li, Q. Zhang, Y. Bai, and Z. Gui, "Image denoising via sparse representation over grouped dictionaries with adaptive atom size," *IEEE Access*, vol. 5, pp. 22514–22529, 2017.



**KAI XIE** received the M.S. degree in electronic engineering from the National University of Defense Technology, Changsha, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent system from Shanghai Jiao Tong University, Shanghai, China, in 2006. He is currently a Professor with the School of Electronic Information, Yangtze University, Jingzhou, China. He currently works in the field of image processing and signal processing.



**HAO-NAN DONG** was born in Shangrao, China. He participated in the design of the experiment and assisted in the completion of the manuscript. In 2021, he joined the National Demonstration Center for Experimental Electrical and Electronic Education to study image processing and deep learning. He has been conducting research on medical image processing and artificial intelligence.



**AN XIE** was born in Huanggang. He conceived and initiated the research, conceived the algorithms, and designed the experiments. In 2021, he joined the National Demonstration Center for Experimental Electrical and Electronic Education to study image processing and deep learning. He has been conducting research projects on object detection and image processing.



**JIAN-BIAO HE** received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 1986 and 1989, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Central South University. His research interests include artificial intelligence, the Internet of Things, pattern recognition, mobile robots, and cloud computing.

...