ITC 2/54
Information Technology and Control
Vol. 54 / No. 2/ 2025
pp. 504-519
DOI 10.5755/j01.itc.54.2.40728

Yolov5-based Intelligent Detection Method for Retail Goods

Received 2025/05/23 | Accepted after revision 2025/05/01

HOW TO CITE: Jiang, Z. (2025). Yolov5-based Intelligent Detection Method for Retail Goods. *Information Technology and Control*, 54(2), 504-519. https://doi.org/10.5755/j01.itc.54.2.40728

Yolov5-based Intelligent Detection Method for Retail Goods

Zixin Jiang

School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China; e-mail: 13677873594@163.com

Corresponding author: 13677873594@163.com

In the current context, intelligent unmanned retail checkout systems offer the prospect of efficient and innovative development. This study proposes an enhanced lightweight YOLOv5 merchandise detection and recognition method. The method introduces SELayer and a multi-headed self-attentive module of Transformer in YOLOv5 to enable the network to focus more on essential factors such as commodities when performing retail merchandise detection, and improve the recognition performance of the model. Also, the Ghost module is introduced to reduce network parameters and computation, increase computation speed and reduce latency. We validated the performance of the approach on a public dataset. Compared with the existing YOLOv5 model, the model achieves a 0.9% improvement in detection accuracy and a 27.7% reduction in GFLOPs. With this study, we optimise the problem of small batch identification of retail goods, providing a basis for automated processing of intelligent retail supply and marketing systems with practical implications.

KEYWORDS: Commodity detection, Multi-target detection, YOLOv5, Attention mechanism, SELayer, Multi-headed self-attention, Deep learning, Ghost, Small-target detection, Intelligent retail

1. Introduction

According to the "2020 China Convenience Store Development Report" published by the China Chain-Store & Franchise Association (CCFA), China's convenience store industry experienced significant growth in 2019, with total sales reaching 255.6 billion yuan and a growth rate of 13%. Among them, 15% of the sample enterprises carried out the pilot project of unmanned stores. Affected by the epidemic and driven by the demand for "contactless" shopping, unmanned retail has garnered renewed market attention. With the popularization of Internet technology and the unmanned retail concept, the users' scale and transaction volume of unmanned retail stores will usher in a blowout outbreak [27, 12]. However, intel-

ligent technology in unmanned stores is still not fully mature, often encountering failures such as user identification issues, account settlement problems, and door-opening malfunctions. At the same time, the slow speed of identification and opening of doors also dramatically affects the experience of consumers [4]. Intelligent technology in unmanned shops is mainly an intelligent retail settlement system, which aims to use image recognition and target detection technology to make accurate, intelligent, and automatic price settlement of the goods purchased by customers. When the customer places his selected goods in the designated area, an ideal intelligent retail settlement system should be able to accurately identify each item and show a correct list and the total price of the item the customer should pay.

At present, there are three leading technologies for commodity identification in domestic unmanned Msupermarkets: bar code identification [13], RFID (Radio Frequency Identification) technology [19], and image recognition technology based on deep learning [29]. Bar code offers several advantages, including a large amount of information, easy readability, reliable coding rules, convenient printing, and high cost-performance. Therefore, it is widely used in supermarkets, libraries, warehouse management, and logistics tracking. At present, many unmanned supermarkets also use the form of customer self-service scanning bar codes for settlement. However, this method takes a long time, there are problems such as broken bar codes that cannot be identified, and at the same time, the scanning process still requires the consumer or salesperson to participate in the process. Radiofrequency identification (RFID) offers advantages such as fast reading speed and durability compared to barcodes. However, its overall application cost in retail remains high. Supermarkets must manually attach RFID tags to each item, incurring significant labor and management costs. Additionally, RFID readers are expensive and require professional maintenance. In some cases, such as near metal or liquid items, RFID may experience reading errors, increasing operational complexity. These factors limit RFID's widespread adoption in retail. In addition, based on the commodity identification technology of deep learning, such as Amazon Go [2], the unmanned convenience store of Amazon, consumers can randomly select commodities in the convenience store. They can leave the convenience store without the need for settlement. Companies such as Jingdong Unmanned Supermarket [11] and Hema Fresh [9] use artificial intelligence to realize the identification and settlement system of commodity types through image recognition technology. The integration of deep learning technologies into retail supply chain management represents a revolutionary transformation in the way global retail organizations operate, optimize, and innovate. A comprehensive review [14] examines the profound impact of deep learning across all aspects of retail supply chain operations, from demand forecasting and inventory optimization to logistics planning and customer experience enhancement.

For the use scenario of unmanned supermarkets, the objects placed by customers on the settlement table are uncertain and diverse, so the identification of goods on shelves and in shopping carts requires accurate positioning, identification, and classification. At present, there are mainly one - stage and two - stage target recognition and detection technologies. The latter method is based on a regional candidate frame, which can usually obtain better detection accuracy. The landmark algorithms are RCNN [6], SPPNet [8], Fast RCNN [24], etc. These methods obtain candidate regions first and then classify them. Although the accuracy is outstanding, the calculation is large, and the speed needs to be improved, so it is unsuitable for real-time processing. The one-stage method is a regression-based method with no candidate region. The location and category of predicted targets can be obtained in one step by directly inputting images, so it is also called end-to-end detection. Typical algorithms include SSD series [18, 5, 30, 3], YOLO series [21, 22, 23, 1], Retina-Net [16], etc. These methods are faster than two-stage methods. Moreover, YOLOv 3 [23] is three and four times faster than SSD and Retina-Net, which achieves almost the same accuracy as Faster RCNN and SSD on the PASCAL VOC dataset.

YOLOv5, proposed by Ultralytics in 2020, is a lightweight target detection network that integrates several advanced network design methods based on YOLOv4. For example, for the backbone network, YOLOv 5 still retains the CSPDarknet 53 used in YOLOv 4, but YOLOv 5 adopts the structure of Feature Pyramid Network (FPN) [15] and Pixel Aggre-

gation Network (PAN) [17] to improve the detection performance of the neck network. YOLOv5 has multiple versions, including YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, each designed for different trade-offs between accuracy and computational efficiency. YOLOv5s is the smallest and fastest version, making it suitable for real-time applications with limited computational resources. YOLOv5m offers a balance between speed and accuracy, while YOLOv5l and YOLOv5x provide higher accuracy at the cost of increased computational load. In addition to YOLOv5, other lightweight recognition models such as MobileNetV2 [26] and ShuffleNetV2 [19] have been widely used in real-time applications due to their balance between accuracy and computational efficiency.

As a new and efficient identification method of the YOLO series in recent years, YOLOv 5 is characterized by high speed, high detection accuracy, and a small number of parameters and therefore has been highly studied and applied in target detection tasks in different fields [31, 20]. For example, Zhu et al. [32] integrated the CBAM attention module and the prediction head in Transformer based on YOLOv 5, thus improving the ability to capture and enrich information to obtain higher performance in target detection tasks in UAV capture scenarios. However, its detection speed is less than ideal. Good detection accuracy is obtained for most of the current target detection tasks using YOLOv 5, but some problems still exist: (1) The majority of the methods improve the detection effect by adding various modules, but their detection speed cannot reach the standard of real-time detection. (2) Although quite a few lightweight network models have advantages in model size and can improve the detection speed, they cannot meet the requirements of high-precision detection.

In order to better meet the requirements of recognition accuracy and speed in intelligent retail commodity detection, we have developed a lightweight retail merchandise detection network model, YOLOv5-AG, with the following main improvements:

1 The Multi-head Self-attention module C3TR adapted from Transformer is introduced in the Backbone network, and SELayer is introduced in the Neck layer to make the model pay more attention to the key factors important to commodity detection, reduce the influence of interference factors, and make the model have better detection effect.

2 The Ghost module is added to the YOLOv5 backbone network and neck network to reduce the convolution operation and delay while maintaining a similar recognition performance so that the model can better meet the demand of real-time performance.

This paper will focus on how to obtain the improved YOLOv5-AG. Section 2 explains the working principle and the base model of YOLOv5, as well as the YOLOv5-AG model that obtains significant improvement. Section 3 discusses and analyzes the experimental data results. Section 4 summarizes the proposed work and improvements.

2. Material & Methods

2.1 YOLOv5 Principle

2.1.1 Network Overview

Based on the different widths, depths, and weights of the model, UItralytics has designed four versions. The residual components of the four target detection versions increase in turn, and the number of residual components and the number of convolution kernels for different network structures of the YOLOv5 algorithm are shown in Table 1. With the deepening of the YOLOv5 algorithm network, the ability of the model for feature extraction and multi-feature fusion of objects to be detected is continuously enhanced. The depth and width of each model can be easily adjusted through code modifications. In the experiment, from the perspective of saving memory costs and computing costs, YOLOv5s can make the network more lightweight. Therefore, YOLOv5s, which have a small depth and width of the network model, are selected for model training in this experiment. The YOLOv5 network structure consists of four main components: Input, Backbone (backbone network), Neck (multi-scale feature fusion network), and Head (prediction classifier). Figure 1 shows the structure of YOLOv5s, and the structure of each part of the network is described as follows.

The YOLOv5 input incorporates three improved technologies. By combining four images, scaling, cutting, splicing, and randomly arranging the input images, Mosaic data enhancement increases the number of samples of small targets, enhances the recognition

Table 1
Number of convolution kernels and residual components for different network structures of YOLOv5.

Structural network	Number of residual components	Total number of convolution kernels
YOLOv5s	12	1 001
YOLOv5m	24	1 488
YOLOv5l	36	1 984
YOLOv5x	48	2 480

ability. Adaptive image scaling can automatically add different minimum black edges to the original image during processing analysis, uniformly zoom to the standard size, reducing information redundancy and speeding up the network. The adaptive anchor frame calculation method is also adopted, and the reduced genetic algorithm and K-means algorithm are embedded in the training network. Based on the continuous iteration of the training data set, the appropriate anchor frame size is automatically learned.

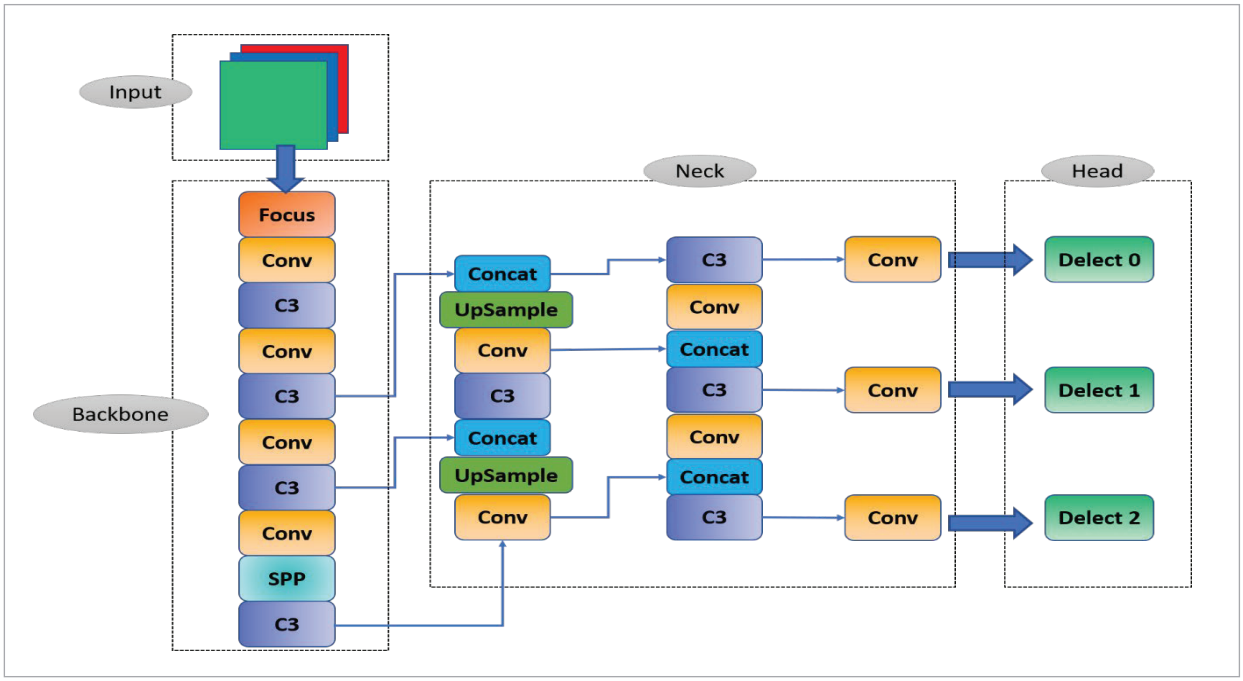
The backbone of YOLOv5 adopts the structure of CSPDarknet53, which integrates the idea of partial connection in the crossover stage with low computational complexity and high identification accuracy. It is mainly composed of Focus, C3, Conv, and Spatial pyramid pooling (SPP), which is used to extract recognition features of images, including edge features, texture features, and location information.

The Focus module extracts pixels from each of the three RGB channels, resulting in each channel generating four channels. This process halves the height and width while quadrupling the number of channels.

The Conv module consists of three parts, extracting local spatial information through a convolution operation, normalising the distribution of feature values through the BN layer and finally introducing a non-linear transformation capability through the activation function, thus enabling the transformation and extraction of the input features.

The bottleneck serves as the basic residual block, combining the two parts of information through summation and then passing the result downstream. The shortcut parameter controls whether to make a residual connection using ResNet. The backbone of YOLOv5 sets Shortcut to True by default, and Bottleneck in Head does not use the shortcut. Correspond-

Figure 1
Diagram of YOLOv5s network structure.



ing to ResNet, ADD is used instead of CONCAT for feature fusion so that the number of fused features remains unchanged.

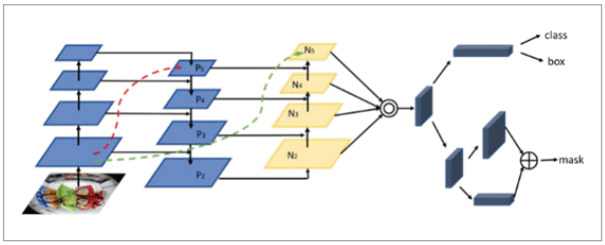
The C3 module is an improved version of the BottleneckCSP module. While its structural function is similar to the CSP architecture, it differs in the selection of correction units and includes three standard convolutional layers and multiple Bottleneck modules for input feature maps.

The Spatial Pyramid Pooling (SPP) module fuses different receptive fields, local features, and global features, thereby improving the network's recognition performance for targets of varying sizes.

The neck network is designed to mix and combine image features to generate feature pyramids. YOLOv5 utilizes the PANet structure, which consists of the FPN + PAN structure. FPN uses a top-down approach to transfer solid semantic features from the full feature map to the feature map below. The PAN structure uses a bottom-up approach to transfer powerful positioning features from a lower feature map to a higher feature map. The PANet structure is shown in Figure 2. These two structures effectively alleviate the problem of deep networks losing the characteristic information of shallow network.

Figure 2

Schematic Diagram of PANet Structure.



The prediction network is usually used to take the feature map obtained from the backbone or neck network to predict the location and category of the object, apply anchor boxes, and output a vector with the category probability, score, and bounding box location of the target object. Each of the three detection layers in the Head network has a different feature map size and outputs a corresponding result vector that can be used to detect objects of different sizes. After this layer, the model predicts and labels the bounding box and the class of each object in the input image.

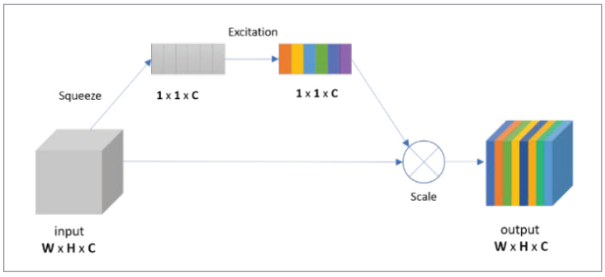
2.2 Improvement of the YOLOv5 Algorithm

2.2.1 Attention Mechanism

Squeeze-and-Excitation (SE) is a network improvement scheme published by Hu et al. of Momenta Autopilot Company [10]. SE module is an attention mechanism module focusing on channel information, which can solve the loss caused by the different channel importance of feature map in the convolution pooling process, and only a few parameters are introduced. It can significantly improve the performance of the original network with a mechanism similar to attention enhancement and low computational power overhead. The SE module comprises three steps: extrusion, excitation, and scaling. Its module structure is shown in Figure 3.

Figure 3

Schematic diagram of SE attention module structure.



F_{sq} stands for Squeeze operation, which is equivalent to a global average pooling. The two-dimensional $H \times W$ feature is compressed into an actual number along the spatial dimension. For example, an ordinary 3×3 convolution operation has a local receptive field, which is limited in scope and cannot use global and contextual information. In contrast, the global average pooling operation compresses the global information into a channel descriptor ($1 \times 1 \times C$), and its calculation formula is as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j), \quad (1)$$

where F_{ex} denotes citation operation. Through a set of parameters W of $1 \times 1 \times C$ that can be learned, a new matrix of $1 \times 1 \times C$ is obtained by combining the statistics of different channels extracted. Moreover, it is regarded as the weight value of each characteristic channel. It can be seen as the use of information between channels to expand the receptive field of subsequent oper-

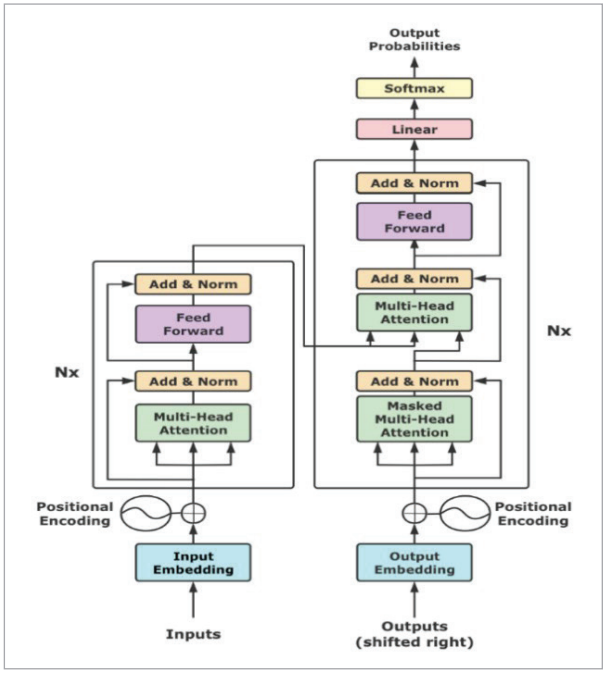
ations in the dimension. The following formula can summarize the operation process:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \quad (2)$$

where F_{Scale} represents the Scale step, which makes adjustments for each channel of the original $W \times H \times C$ diagram based on the weights for each channel we found earlier. In other words, the proportion of values in $H \times W$ on each channel is kept unchanged, and the whole is scaled up and down according to the previous weight value to obtain a new feature graph.

The transformer model was proposed by Vaswani and Shazeer et al. in 2017 [28], who argued that the transformer encoder block is more sensitive and comprehensive to the global and contextual information than the original bottleneck block in CSPDarknet53. The structure of the general Transformer is composed of an encoder and decoder, each containing N=6 layers, where each layer consists of two sub-layers, the multi-headed self-attention and the fully connected network, and uses residual connections between each sub-layer, as shown in Figure 4.

Figure 4
Schematic Diagram of Transformer Structure.



Source: Vaswani, Shazeer et al. (2017) [26]

The self-attentive mechanism introduced by the Transformer allows the output of each neuron layer to take into account all the inputs of that layer, and Each neuron can be computed in parallel. The problem of inefficient attention and distortion during training in RNNs is solved. We found on the VisDrone2021 dataset that the transformer encoder block performs better for obscured objects with high density, which is key to our improvement. The success of the Transformer model is mainly due to the Multi-head Self-attention mechanism. It enables the model to focus on multiple vital areas simultaneously. In the preorder language model, we find that the self-attention model calculates itself as the most noteworthy object. If more than one attention head is added, the model may focus on some objects other than itself. There are multiple weight matrices of Q, K and V in the multi-head self-attention, which are initialized independently and randomly. Then the input vectors are mapped to different subspaces, thus enriching the feature expression of information.

Assuming that the number of attention heads per layer is h, the h-th attention head can be represented by the following formula:

$$Q_h = XW_h^Q, K = XW_h^K, V = XW_h^V \quad (3)$$

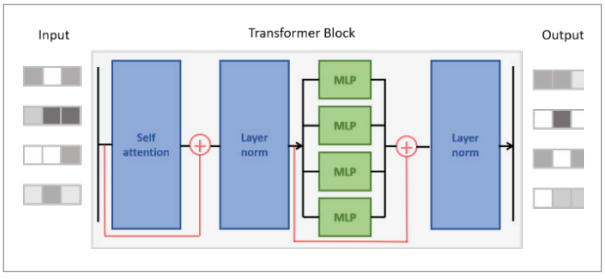
$$A_h = softmax\left(\frac{Q^h K_h^T}{\sqrt{d_k}}\right) \quad (4)$$

$$H_h = AttentionHead(X) = A_h V_h. \quad (5)$$

Then each layer of multi-head attention can be expressed as:

$$MultiH(X) = [H_1, \cdots, H_{|h|}]W^O \quad (6)$$

Figure 5
Schematic Diagram of TransformerBlock Structure.

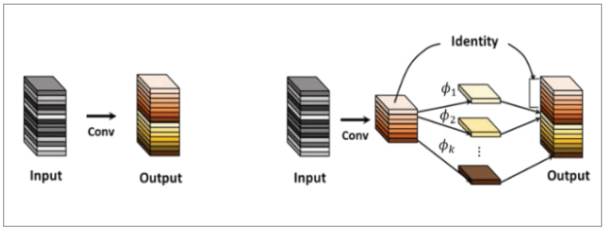


We used the encoding part of the Transformer module and made some changes. The C3TR module is an autonomic force module based on the C3 module that replaces the Bottleneck with TransformerBlock. The structure of TransformerBlock is shown in Figure 5.

2.2.2 Ghost Module

The CNN, composed of a large number of convolutional modules, is computed to obtain a highly redundant intermediate feature map, which will lead to increased computational costs. The Ghost module [7] is a lightweight network that effectively addresses hardware resource limitations and computational constraints. Ghost, as a plug-and-play module, enables a more compact model of YOLOv5 while maintaining high performance.

Figure 6
Schematic diagram of ordinary convolution and Ghost module structure.



Source: Han, Wang et al. (2020) [14]

Figure 6 shows the comparison between ordinary convolution and Ghost module structure. In contrast to the widely used 1×1 pointwise convolution of cells, the main convolution in the Ghost module allows for custom kernel sizes. In the first stage, Ghost uses multiple convolution kernels to computationally generate the original feature mappings, and in the second stage, more Ghost feature mappings are generated by an inexpensive shift operation to increase the number of channels. Whereas in previous efficient architectures, processing of each feature mapping was limited to a depthwise convolution or shift operation, linear operations in the Ghost module allow for a great deal of diversity. In the third stage, the identity is parallelised with the linear transformation in the Ghost module to maintain an intrinsic feature map consistent with the output of normal convolution.

2.2.3 Loss Function

Most of the current target detection algorithms still choose the IoU (intersection and merge ratio) method of the original target frame and the predicted frame shown in Equation (11)(12) as the measure and use 1-IoU as their loss regression method.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

$$L_{IoU} = 1 - IoU\,. \tag{8}$$

Although the original IoU loss solves the problem of the relative fit overlap between two mutually independent variables between the target frame and the prediction frame, it still has some limitations: firstly, for the case where the two frames do not intersect, according to the definition, it does not adequately reflect the degree of overlap between them. Meanwhile, when Loss =0, there is no gradient return, so it is impossible to learn and train. Secondly, the IoU does not accurately reflect how the target and prediction frames intersect.

Generalized Intersection over Union (GIoU), proposed by Rezatofighi et al. [25], addresses the limitations of IoU by considering both intersection and disjunction cases. The YOLOv5 model incorporates GIoU to improve bounding box regression. GIOU incorporates a minimum rectangle C, including the target box A and the predicted box B on top of the original IOU. The improved formulation is as follows.

$$GIoU = IoU - \frac{|C(A \cup B)|}{|C|} \tag{9}$$

$$GIoUloss = 1 - GIoU\,. \tag{10}$$

The areas of the minimum closure regions of the two frames are calculated, and the areas of the minimum frames of the predicted frame and the actual frame are included at the same time; Calculate the proportion of areas in the closed area that do not belong to these two frames. Moreover finally, the proportion is subtracted from IoU to obtain a GIoU. GIoU considers the fact that there is no overlap between the detection frame and the actual frame and has a faster convergence rate. However, GIoU degenerates to IoU when the two target frames intersect, and does not achieve

better optimization of the target and prediction frames, so the convergence of the actual and prediction frames in the horizontal and vertical directions is not improved.

2.2.4 Improved YOLOv5 Structure

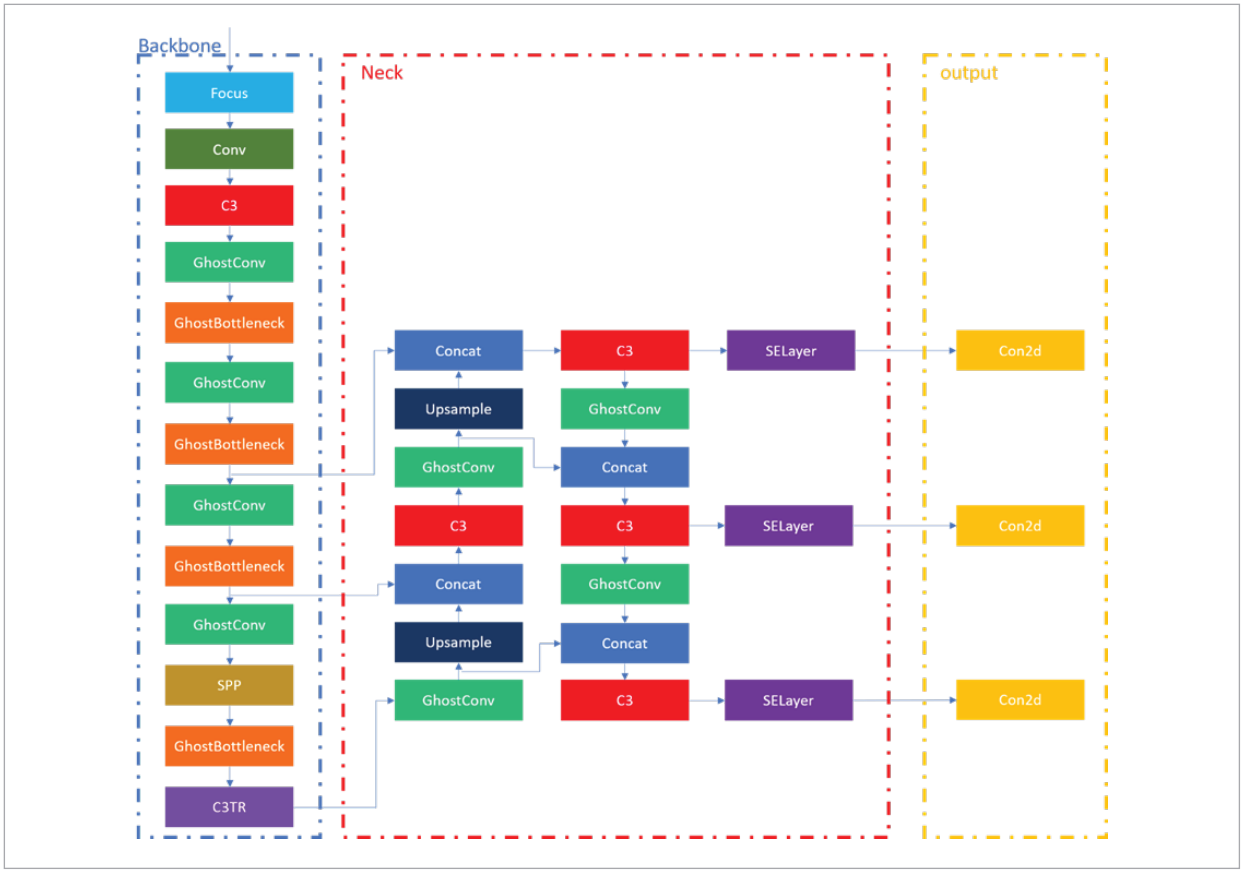
After the experiment, we updated YOLOv5 to YOLOv5-AG, and its network structure is shown in Figure 7. First, we changed the activation function in the C3 module of the original YOLOv5 network from SiLU to hard swish. In the Backbone layer, the first C3 module is retained, the remaining C3 modules are replaced by GhostBottleneck modules (orange), all Conv modules are replaced by GhostConv modules (grass green), and a C3TR module (bright purple) is added at the end of the Backbone. In the Neck layer, all Conv modules have been replaced with GhostConv modules, and a SElayer module (dark purple) is attached after each upsample + concat.

2.2.5 Evaluation Index

In order to verify the effectiveness of the proposed model, both qualitative and quantitative evaluations are conducted. For qualitative evaluation, the performance of the model is evaluated by comparing the detected images of YOLOv5-AG with the comparison method, including comparing the positioning accuracy of the target frame, whether it is missed or false. The main indexes of quantitative evaluation are the number of parameters, GFLOPs, mAP@0.5, and mAP@0.5: 0.95.

The time and space complexity of the model is represented using model parameters and GFLOPs. mAP@0.5 is used to evaluate the recognition ability of the target detection model, while mAP@0.5:0.95 is mainly used to reflect the positioning effect of the model and the overstep rollback ability due to the need for higher total value of the IOU. The values of

Figure 7
Diagram of YOLOv5-AG network structure.



these four evaluation indexes are positively correlated with the detection effect.

The AP calculation formula in mAP calculation is as follows:

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{11}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{12}$$

$$AP = \int_0^1 P(r)dr\,. \tag{13}$$

Taking the detected "hongniu" categories as an example, is the number of items correctly detected by the model labeled as "hongniu", refers to the number of items that identify the " hongniu " category as other items or that do not detect the presence of " hongniu "., and F_N is the number of items detected as "hongniu" targets. AP denotes the integral of accuracy over recall, and two metrics, P and, R are usually used to measure how good the model is. Average all categories of AP as the average accuracy average, mAP which can measure the performance of the whole model.

3. Results & Discussion

3.1 Data Set

This experiment uses the public data set, which has 113 categories, and the data set is the product images taken from the top view, JPG format. There are 5,422 photos, all of which are marked in Pascal VOC format. There are 4238 training sets and 1184 test sets. Figure 8 is a sample image of the dataset.

Figure 8
Sample Dataset Picture.

3.2 Experimental Platform

The experiment uses Ubuntu 20.04 system with Ubuntu 20.04 system, and the software versions are Python 3.8 and Pytorch 1.8. The hardware configurations and model parameters related to the experiment are shown in Table 2. YOLOv5 can adaptively scale images, select 640×640 images as input, and obtain feature images of equal scale as detection scale. The default learning rate is 0.01, the batch size is 16, and the training period is 1000 times.

Table 2
Lab configuration information and Experimental model parameters.

Name	Configuration	Name	Value
GPU	RTX2080 Ti	Size of images/ pixel	640 × 640
CUDA	11.2	Learning	0.01
CuDNN	8.2.0	Batch size	16

3.3 Ablation Study

This study uses the classical YOLOv5 network model, whose overall recognition accuracy is satisfactory. MAP@0.5 reaches 97.5%. However, we also found that the original YOLOv5 network model is not accurate in the identification of some commodities, such as "hongniu," "bingqilinniunai," and "yezhi", and the average identification accuracy is less than 0.90, only 0.851,0.580 and 0.884.

Table 3 lists the categories of commodities with identification accuracy lower than 0.90. The physical objects and their identification renderings are shown in Figures 9-10. As shown in Figure 9(b), the original YOLOv5 model mistakenly detected items outside the detection area as items within the detection area. In contrast, Figure 9(c) demonstrates

Table 3
Category and accuracy of commodity. with low identification accuracy

Category	mAP@0.5	Category	mAP@0.5
bingqilinniunai	58.0%	daofandian1	88.5%
daofandian4	85.0%	daofandian2	88.9%
hongniu	83.5%	dongpeng	86.8%
kafei	88.6%	yezhi	88.4%

Figure 9
Detection effect case one, (a) original image, (b) figure of Yolov5S test results, (c) figure of Yolov5-AG test results.

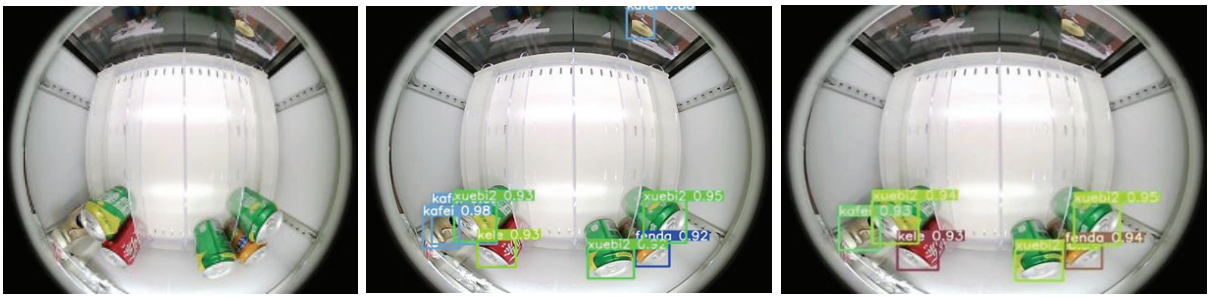
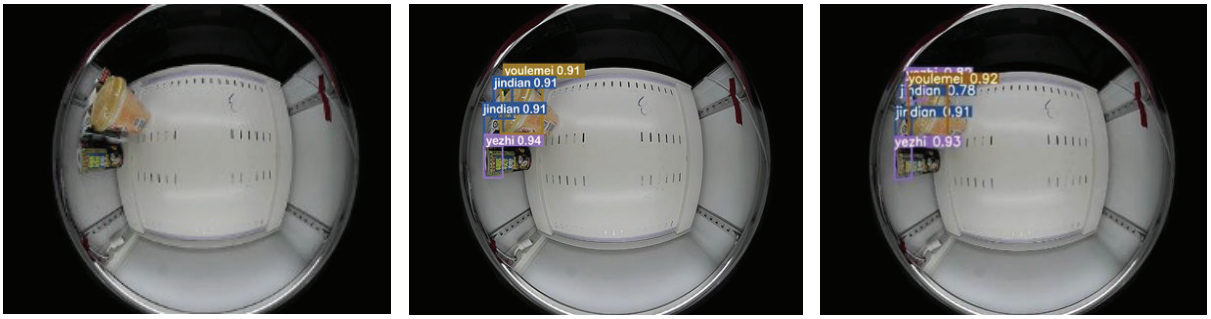


Figure 10
Detection effect case two, (a) original image, (b) figure of Yolov5S test results, (c) figure of Yolov5-AG test results.



that our proposed model accurately identified all the items within the detection area without any false detections. Similarly, in Figure 10(b), the original model failed to detect one of the five items due to stacking and blocking. However, our proposed model, as shown in Figure 10 (c), successfully recognized the "yezhi" item that was missed by the original model due to occlusion.

To address the issue of low recognition accuracy caused by the stacking and blocking of some goods, we incorporated channel attention technology by adding the SELayer module (YOLOv5-V1) to the original network. Table 4 shows the test results of the model under the same training parameters. Compared with the original YOLOv5s network, YOLOv5-V1 improves precision by 0.9%, recall by 1%, F1 score by 0.6%, mAP@0.5 by 0.8%, and mAP@0.5:0.95 by 2.6%. This result shows that, the inclusion of the SE attention mechanism allows the neural network to focus on certain feature channels by automatically learning the importance of each channel of the feature map and then using this importance to assign a weight value to each feature. Boosting the channels of the feature map that are useful for the current task and suppressing the channels of features that are less useful for the current task.

At the same time, we observed that the recognition accuracy rate of some products was lower than 90%. In this regard, we tried to use the C3TR module (YOLOv5-V2) adapted from the combination of the Transformer and C3 module to optimize the low recognition accuracy of some products. C3TR is adapted from the multi-head self-attention module in Transformer, which is a variant of the attention mechanism and can reduce dependence on external information and can ameliorate the ability of the model to deal with the internal correlation of data. We tested the YOLOv5-V2 network with the same training parame-

ters as the original YOLOv5s, which can be compared and analyzed in Table 4. We found that the effect of using C3TR is not ideal. Not only is the accuracy hardly improved, but also the F1 is reduced by 0.4%. It may be because the self-attention mechanism model is added, and the convergence rate of the model will be slow under the same number of iterations.

To further enhance the detection performance, we developed the YOLOv5-V3 model, which integrates both the SELayer and C3TR modules into the YOLOv5s framework, and the detection effects were also compared in Table 4. It shows that the mAP@0.5 is increased by 0.8% and 0.2%, respectively, after adding SELayer and C3TR attention mechanisms into the original network of YOLOv5s. When both SE and C3TR were added, precision improved by 0.9%, recall by 1.4%, F1 by 12.6%, mAP@0.5 by 0.8%, and mAP@0.5:0.95 by 3.5%. Where precision and mAP@0.5 are the same as adding only the SE module, although the effect is not ideal when adding C3TR alone when working with SE, it gives a good boost to recall, F1 and mAP@0.5:0.95. The combination of SELayer and C3TR enables the network to automatically assign different weights to different features, constantly capture information between features, and effectively integrate learning, so as to effectively improve the network's ability to identify features and improve the recognition effect of the model.

Testing in YOLOv5-V3, we found that all categories have reached the recognition accuracy rate of 0.9 or above. In order to show it better, we only draw the product categories with improved accuracy before and after improvement. In contrast, the product categories with the same accuracy and above 0.95 accuracies before and after improvement are not shown in Figure 11. From Figure 11, we can find that the improved version's recognition rate (YOLOv5-V3) is greatly improved.

Some commodities with poor recognition accuracy in the YOLOv5s network have been improved in the proposed model YOLOV5-V3. The effect is shown in Table 7, where the detection accuracy of commodities with the category "bingqilinniunai" improved by 32%.

Table 4

Comparison of results between YOLOv5s and three improved methods.

model	SE	C3TR	Precision	Recall	F1	mAP@0.5	mAP@0.5:0.95
YOLOv5s	-	-	97.1%	96.2%	63.3%	97.5%	87.9%
YOLOv5-V1	√	-	98.0%	97.2%	72.7%	98.3%	90.5%
YOLOv5-V2	-	√	97.2%	97.0%	62.9%	97.7%	88.2%
YOLOv5-V3	√	√	98.0%	97.6%	75.9%	98.3%	91.4%

Figure 11

Comparison of YOLOv5 and YOLOv5-V3 results.

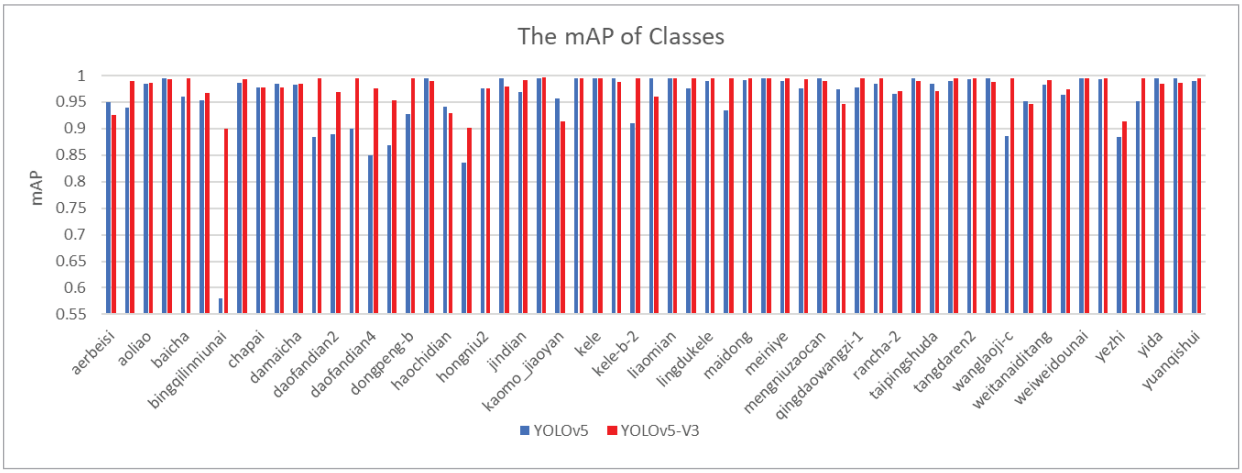


Table 5
Accuracy improvement results for goods with poor identification accuracy

Category	YOLOv5s	YOLOv5-AG
bingqilinniunai	58.0%	90.0%
daofandian4	85.0%	97.6%
hongniu	83.5%	90.1%
wanglaoji-c	88.6%	99.5%
daofandian1	88.5%	99.5%
daofandian2	88.9%	96.9%
dongpeng	86.8%	95.4%
yezhi	88.4%	91.3%

It can be found from Figure 11 and Table 5 that in the YOLOv5-AG network based on the double attention mechanism, the SE attention mechanism pays more attention to the channel information, which can optimize the loss problem caused by the unbalanced weight of each channel of the feature map in the convolution pooling. C3TR, on the other hand, is an adaptation of the multi-headed self-attention mechanism introduced in Transformer, which can ameliorate the model's ability to deal with the internal correlation of data and reduce the need for additional external information. The combination of the two can make the model automatically assign the weight of each channel feature and capture the relationship and information between the two learning features. Thus, it effectively improves the ability of the network to identify features, thereby improving the detection accuracy of the network. The effect of YOLOv5-AG network detection is shown in Figure 9(c) and Figure 10(c). In the original network, the situation of false and missed detection has improved.

As shown in Table 6, YOLOv5-V3 is an enhanced version of YOLOv5s that integrates both the SELayer and C3TR modules to improve detection accuracy.

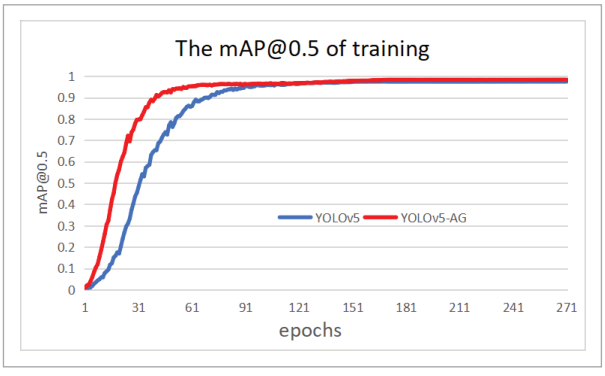
Table 6
Comparison of calculation amount after adding Ghost.

model	Numbers of layers	Parameters	GFLOPS
YOLOv5s	283	7365606	17.3
YOLOv5-V3	349	7805606	18.4
YOLOv5-AG	435	4968006	12.5

YOLOv5-AG, on the other hand, further incorporates the Ghost module to reduce computational load and improve real-time performance. Although the (YOLOv5-V 3) network based on the attention mechanism greatly improves the accuracy of commodity category recognition, it also increases the number of parameters and calculations of the network. The original YOLOv5s network has fewer layers and more extensive parameters. However, with the addition of C3TR and SELayer, the number of layers and parameters is increased. To meet the demand for real-time detection of intelligent retail goods, we are considering adding a Ghost module (YOLOv5-AG) based on YOLOv5-V 3, hoping to reduce the network burden and significantly reduce the amount of network computation. We used GhostConv to replace some

Conv modules and GhostBottleneck to replace some C3 modules and compared the modified YOLOv5-AG with the original YOLOv5s network and the network with a dual attention mechanism (YOLOv5-V 3). Although in the YOLOv5-AG network, the network layers increased by 53.7%. The parameters decreased by 32.6%, and the GFLOPS decreased by 27.7%. Compared with YOLOv5-V 3, when the network layer is increased by 24.6%, the number of parameters decreases by 36.4%, and GFLOPS decreases by 32.1%. Moreover, as shown in Table 7, we find that YOLOv5-AG can effectively reduce the model parameters, and in the meantime, it can handle the balance of accuracy and speed excellently and still maintain the high accuracy of 98.3% mAP@0.5 compared with YOLOv5-V 3. We also compared the convergence rates of YOLOv5s and YOLOv5-AG.

Figure 12
Comparison of convergence rate between YOLOv5s and YOLOv5-AG.



As shown in Figure 12, the YOLOv5 model before improvement needs about 75 epoch iterations to achieve 90% mAP, and the maximum mAP is only 97.5%. The improved YOLOv5-AG can not only reach 90% mAP at about 40 epochs but also reach 98.3% maximum mAP. Our model has higher accuracy and faster convergence speed than the two models. This model is compared with other network models of the same series and different series in the same data set and shown in Table 8 to further prove its performance.

As shown in Table 8, compared with YOLOv3, YOLOv4, and YOLOv5s in the YOLO series, our YOLOv5-AG obtained by us improves the accuracy but also greatly reduces the calculation amount and model parameters. mAP@0.5 of this method improves 5.5%, 3.1%, and 1.2% compared to the three. Meanwhile, YOLOv5-AG still gains advantages in both accuracy and computational effort compared with the two-stage Faster-RCNN, whose mAP@0.5 is 1.3% higher than the Faster – RCNN. To further validate the innovation of our proposed YOLOv5-AG model, we compared it with other lightweight models, including NanoDet, MobileNetV2 and ShuffleNetV2. These models are known for their efficiency in real-time applications. Our experiments show that while these models achieve high frame rates, our YOLOv5-AG model offers superior accuracy with only a moderate increase in computational cost. This demonstrates the effectiveness of our method in balancing accuracy and efficiency.

Table 7
Accuracy comparison after adding Ghost.

model	SE	C3TR	Ghost	Precision	Recall	F1	mAP@0.5	mAP@0.5:0.95
YOLOv5s	-	-	-	97.1%	96.2%	63.3%	97.5%	87.9%
YOLOv5-V1	√	-	-	98.0%	97.2%	72.7%	98.3%	90.5%
YOLOv5-V2	-	√	-	97.2%	97.0%	62.9%	97.7%	88.2%
YOLOv5-V3	√	√	-	98.0%	97.6%	75.9%	98.7%	91.4%
YOLOv5-AG	√	√	√	98.0%	97.6%	75.9%	98.7%	91.4%

Table 8
Compare the performance of different networks.

model	mAP@0.5	mAP@0.5:0.95	Parameters	GFLOPS	FPS
YOLOv5s	97.5%	87.5%	7365606	17.3	30.12
YOLOv3	93.2%	74.8%	8859276	13.5	27.32
YOLOv4	95.6%	80.5%	6028427	17.1	28.64
Faster-RCNN	97.4%	85.9%	3878580	20.0	29.53
NanoDet	93.4%	75.2%	750,000	2.1	89.32
MobileNetv2	98.3%	87.4%	3106214	7.4	39.04
ShuffleNetv2	98.3%	87.1%	4095015	8.9	37.98
YOLOv5-AG (ours)	98.7%	91.4%	4968006	12.5	35.46

4.Discussion

This study aims to optimize the intelligent retail goods detection method by using YOLOv5. In Table 8, the recognition accuracy of the classic YOLOv5s has reached 97.5%. However, it is found in Table 4 and Figure 9 and 10 that some goods are missed due to the stacking and blocking of items during place-

ment. Thus, it makes the accuracy too low for some categories. In order to optimize this problem, we add an attention mechanism to the original network, as shown in Table 2, and adding C3TR and SELayer can achieve better performance than using them alone. It makes the model automatically assign the weight of each channel feature and captures the relationship and information between two features learned. This is because the SE attention mechanism pays more attention to the channel information, which can optimize the loss problem caused by the unbalanced weight of each channel of the feature map in the convolution pooling. Meanwhile, C3TR, which is adapted by introducing the Multi-head Self-attention mechanism in Transformer, ameliorates the ability of the model to deal with the internal correlation of data and reduce the need for additional external information. The combination of the two effectively solves the problem of low recognition accuracy of some product categories in the original network, as shown in Table 5, where the recognition accuracy of the product category "bingqilinniunai" is greatly improved from 58.0% to 90.0%.

Meanwhile, we introduced the Ghost module better to meet detection speed and convenience in intelligent retail detection. It is used GhostConv to replace part of the Conv module and GhostBottleneck to replace part of the C3 module. The improved network can be ported to some mobile devices with relatively weak computational power (e.g., smartphones, etc.) to ensure the algorithm's performance capability.

In order to better verify the performance of this method, we use the same commodity detection data set to conduct tests in different identification networks, and the comparison results are shown in Table 8. The results show that YOLOv5-AG has superior performance in detecting small objects such as commodities, and the recognition accuracy of each commodity in the data set is above 90%. Especially for commodities with low recognition accuracy, the optimized network becomes lighter.

5.Conclusion

In order to detect retail goods better and faster and to solve the deficiency of goods recognition effect in intelligent retail, this study combines YOLOv5 with the SELayer attention mechanism, the self-intention module in Transformer (that is, the rewritten C3TR), and the Ghost module to construct an intelligent retail goods detection method. This method aims to improve the network's accuracy in detecting the occlusion of small objects caused by the thick discharge. Furthermore, the target detection network is lightened as much as possible. Experiments show that detection accuracy can be significantly improved by adding the SELayer attention mechanism alone. In contrast, the detection accuracy can hardly be improved by adding the C3TR autonomous intention module alone. However, when the two mechanisms work together, the recognition effect of the network can be further improved, especially for some items with low recognition accuracy in the original YOLOv5 network. Finally, the identification accuracy of 113 kinds of items reached 90.0% or more, and the overall average accuracy improved from 97.5% to 98.7%. With the addition of the Ghost module, the number of layers of the network is increased.

Meanwhile, the loss of parameters is significantly reduced, thus improving the detection speed. It is of great value to the specific application of smart retail products, whether it is to identify each product that customers choose to buy accurately. It is also real-time and fast processing of a particular batch of commodities. With the rapid development of Internet of Things technology, the diversification of consumption patterns, and the continuous stimulation of the e-commerce economy to physical sales, unmanned retail has attracted much attention. In this paper, by improving YOLOV5, higher recognition accuracy is achieved, and the amount of calculation is greatly reduced, which meets the needs of the market today.

Data Availability

The dataset used in this study is publicly available and can be accessed at https://aistudio.baidu.com. The dataset comprises 113 categories of retail goods. All data and materials necessary to replicate the findings of this study are included in the manuscript or available from the corresponding author upon reasonable request.

References

1. Bochkovskiy, A., Wang, C.-Y., Liao, H. Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv Preprint arXiv:2004.10934, 2020. https://doi.org/10.48550/arXiv.2004.10934
2. Calderón-Ochoa, A. F., Coronado-Hernandez, J. R., Portnoy, I. Throughput Analysis of an Amazon Go Retail Under the COVID-19-Related Capacity Constraints. Procedia Computer Science, 2022, 198, 602-607. https://doi.org/10.1016/j.procs.2021.12.293
3. Cui, L., Ma, R., Lv, P., Jiang, X., Gao, Z., Zhou, B., Xu, M. MDSSD: Multi-Scale Deconvolutional Single Shot Detector for Small Objects. arXiv Preprint arXiv:1805.07009, 2018. https://doi.org/10.48550/arXiv.1805.07009
4. Fan, X., Ning, N., Deng, N. The Impact of the Quality of Intelligent Experience on Smart Retail Engagement. Marketing Intelligence & Planning, 2020, 38(7), 877-891. https://doi.org/10.1108/MIP-09-2019-0439
5. Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., Berg, A. C. DSSD: Deconvolutional Single Shot Detector. arXiv Preprint arXiv:1701.06659, 2017. https://arxiv.org/pdf/1701.06659
6. Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. https://doi.org/10.1109/CVPR.2014.81
7. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C. GhostNet: More Features from Cheap Operations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. https://doi.org/10.1109/CVPR42600.2020.00165
8. He, K., Zhang, X., Ren, S., Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9), 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824
9. He, Y., Xie, S. Discussion on New Retail from the Perspective of Artificial Intelligence-Taking Hema Fresh as an Example. Modern Trade Industry, 2020, 41(12), 40-41. https://doi.org/10.19311/j.cnki.1672-21982020.12.020
10. Hu, J., Shen, L., Sun, G. Squeeze-and-Excitation Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745
11. Intel. Intel Helps JD Cope with New Challenges of "Borderless Retail". Intel White Paper, 2019, (3), 43-46.
12. Isharyani, M. E., Sopha, B. M., Wibisono, M. A., Tjahjono, B. Smart Retail Adaptation Framework for Traditional Retailers: A Systematical Review of Literature. 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, 2021, 143-147. https://doi.org/10.1109/IEEM50564.2021.9673056
13. Kjellberg, H., Hagberg, J., Cochoy, F. Thinking Market Infrastructure: Barcode Scanning in the US Grocery Retail Sector, 1967-2010. In: Thinking Infrastructures. Emerald Publishing Limited, 2019, 207-232. https://doi.org/10.1108/S0733-558X20190000062013
14. Lee, S. K., Rah, M.-J. Deep Learning in Retail Supply Chain Management: An Evolution. World Journal of Economics and Business Research, 2024, 2(2), 37-43. https://doi.org/10.61784/wjebr3024
15. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106
16. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P. Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 42(2), 2999-3007. https://doi.org/10.1109/TPAMI.2018.2858826
17. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. Path Aggregation Network for Instance Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913
18. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C. SSD: Single Shot Multibox Detector. In: Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I. Springer International Publishing, 2016, 14. https://doi.org/10.48550/arXiv.1512.02325
19. Mekruksavanich, S. Supermarket Shopping System Using RFID as the IoT Application. 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON). IEEE, 2020, 83-86. https://doi.org/10.1109/ECTIDAMTNCON48261.2020.9090714

20. Qi, J., Liu, X., Liu, K., Xu, F., Guo, H., Tian, X., Li, M., Bao, Z., Li, Y. An Improved YOLOv5 Model Based on Visual Attention Mechanism: Application to Recognition of Tomato Virus Disease. Computers and Electronics in Agriculture, 2022, 194, 106780. https://doi.org/10.1016/j.compag.2022.106780
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Computer Vision and Pattern Recognition, 2016. https://doi.org/10.1109/CVPR.2016.91
22. Redmon, J., Farhadi, A. YOLO9000: Better, Faster, Stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 7263-7271. https://doi.org/10.1109/CVPR.2017.690
23. Redmon, J., Farhadi, A. YOLOv3: An Incremental Improvement. arXiv Preprint arXiv:1804.02767, 2018.
24. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems, 2015, 28. https://doi.org/10.1109/TPAMI.2016.2577031
25. Rezatofighi, H., Tsoi, N., Gwak, J. Y., Sadeghian, A., Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. https://doi.org/10.1109/CVPR.2019.00075
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 4510-4520. https://doi.org/10.1109/CVPR.2018.00474
27. Sujana, A., Fazxal-e-Hasan, S. M., Makam, S. B., Azeem, M. M., Mortimer, G. Examining the Antecedents and Consequences of Perceived Shopping Value Through Smart Retail Technology. Journal of Retailing and Consumer Services, 2020, 52, 101901. https://doi.org/10.1016/j.jretconser.2019.101901
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 30. ttps://dl.acm.org/doi/10.5555/3295222.3295349
29. Wei, X. S., Cui, Q., Yang, L., Wang, P., Liu, L. RPC: A Large-Scale Retail Product Checkout Dataset. arXiv Preprint arXiv:1901.07249, 2019. https://doi.org/10.48550/arXiv.1901.07249
30. Zheng, L., Fu, C., Zhao, Y. Extend the Shallow Part of Single Shot Multibox Detector via Convolutional Neural Network. Tenth International Conference on Digital Image Processing (ICDIP 2018). SPIE, 2018, 10806, 287-293. https://doi.org/10.1117/12.2503001
31. Zhou, J., Zhenbo, B. Ship Target Detection Algorithm Based on Improved YOLOv5. Journal of Marine Science and Engineering, 2021, 9(8), 908. https://doi.org/10.1109/CEI60616.2023.10528110
32. Zhu, L., Geng, X., Li, Z., Liu, C. Improving YOLOv5 With Attention Mechanism for Detecting Boulders from Planetary Images. Remote Sensing, 2021, 13(18), 3776. https://doi.org/10.3390/rs13183776