Jestr

Journal of Engineering Science and Technology Review 17 (6) (2024) 41 - 48

Research Article

JOURNAL OF
**Engineering Science and Technology Review**

www.jestr.org

# Detection Method for Transport State of C80B Coal Open Wagons Based on Improved YOLOX Algorithm

**Wenge Song[1], Jun Yang[1], Haisheng Zhang[1], Adiya Yadamsuren[2,3] and Youqing Ma[2,4,*]**

[1]*Shenhua Shendong Coal Group Corporation, Ltd. Shanxi 719315, China*
[2]*Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China*
[3]*Wild Camel Protection Foundation Mongolia, Ulaanbaatar 17010, Mongolia*
[4]*International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China*

Received 11 September 2024; Accepted 9 December 2024

***Abstract***

When transporting coal under the effect of various factors, open wagons often encounter issues, such as red ore, coal residues, foreign objects, and frozen and snowy bottoms, which increase consumption and costs. The inspection of open-wagon transport status mostly relies on manual monitoring, which has low automation and is labour intensive and inefficient. This study proposed a You only look once X (YOLOX) model based on a parameter-free attention residual structure and lightweight adaptive feature fusion pyramid to improve the accuracy of the detection of the state of open-vehicle transport. A parameter-free attention residual structure (SARM) and self-adaptive spatial feature fusion module (SASFF) modules were introduced into the neck network of YOLOX. The improved YOLOX model was built by using a highly perceptive detection structure (DCBS) module proposed in this study in the detection process, and the accuracy of the model was verified by experiments. Subsequently, a high-quality open-truck-transportation image dataset was constructed for the open-truck transportation target detection task and accurately labelled. Results demonstrate that, in experiments on the Common Objects in Context 2017 dataset, the improved YOLOX algorithm shows improvements of 11.9%, 4.4%, and 1.9% over the YOLOX, YOLO-V5, and YOLO-V8 algorithms, respectively, and the proposed algorithm exhibits considerable improvements in missed and wrong detections. The proposed method provides the rapid detection of the transport status of open wagons for subsequent automated reporting and logging.

*Keywords:* Object detection, C80B open wagon, Transport state, Attention mechanism

## 1. Introduction

With the booming development of information technology, deep learning algorithms have received increasing attention in numerous disciplines. These algorithms are based on massive data and can perform diverse tasks quickly and accurately, especially in the fields of digital city construction and image processing, and thus show promise for a wide range of applications in the field of foreign object and state detection. Computer image detection technology has become the core science and technology in the fields of manufacturing, biology, transport, and military and has made important contributions to the progress of society and development of the national economy.

Traditional technical means of foreign object and vehicle load detection, such as real-time on-site video monitoring, plug-in sensors [1], and millimeter-wave radar [2], have become the direction of research in domestic railway scenarios to meet the demand for high performance and precision in specific situations. These techniques either rely on manual discrimination or make use of instruments for rapid detection. However, these sensors cannot easily meet the actual needs of monitoring open-vehicle transport conditions due to the limitations imposed by their strict requirements for the detection environment, high maintenance costs, and susceptibility to external interference. With the continuous advancement of information technology, computer image-based defect detection technology has become ideal for the monitoring of open-vehicle transport conditions due to its advantages of high automation, efficient execution, and non contact detection. In contrast to other inspection techniques, computer image inspection technology effectively overcomes the numerous problems of manual inspection and machine-mediated detection.

The above analysis shows that although scholars have conducted numerous studies on foreign object and vehicle load detection [3-6], the current research still has some shortcomings. First, detection accuracy and real-time performance still need to be improved. In particular, how to ensure the accuracy and real-time performance of test results in complex and changing traffic environments is an urgent problem to be solved. Second, existing detection methods are often affected by environmental factors, such as lighting and weather conditions, thus limiting their wide promotion in practical applications.

In consideration of the above-mentioned real-world problems and theoretical background, this study aims to explore new methods and techniques for foreign object and vehicle load detection at the theoretical level. Through an in-depth study of existing detection technologies in combination with the latest research results and technological trends, we propose an efficient and accurate detection method to solve the problem of detecting foreign objects and vehicle loads in real traffic environments. Our proposed method not only has important theoretical value, it will also provide strong technical support for the

*E-mail address: mayq@aircas.ac.cn
ISSN: 1791-2377 
doi:10.25103/jestr.176.06

development of coal mine traffic safety and logistics industry.

## 2. State of the art

The task of foreign object and vehicle load detection mainly involves a fusion of detection algorithms and practical application scenarios. However, complex environments and effective feature extraction have always been a problem for detection algorithms. Existing deep learning foreign object tracking and detection algorithms are susceptible to the influence of complex environments and target occlusion, leading to problems, such as leakage and low detection accuracy. Chen Yong et al. [3] proposed a spatial localization and feature generalization enhancement of the GhostNet feature network for a railway foreign object tracking and detection algorithm. However, the stability and robustness of the algorithm in the face of extreme bad weather or highly similar foreign object interference still need further validation and optimization. Ye Tao et al. [4] simplified feature extraction through the linear transformation of feature maps, employed adaptive multicale feature fusion to optimize feature expression capability, and subsequently used lightweight attention to improve accuracy in combination with the Jetson embedded platform. They thus proposed an autonomous detection system for foreign objects in track intrusion boundaries based on LAM-Net. However, the system's balance between real-time performance and accuracy in addressing the interior state of fast-moving carriages, especially in complex and dynamic environments, still requires further optimization.

With the wide application of the YOLO model, research on optimization algorithms for YOLO for foreign object detection is also being conducted for the first time. For example, Hao Shuai et al. [5] used the adaptive histogram equalization algorithm to enhance the contrast of conveyor belt images in underground coal mines and reduce interference from coal dust. Subsequently, network detection speed was improved by introducing depth-separable convolution into the framework of the YOLOv5 algorithm, and the detection accuracy of the whole network was improved by optimizing the loss function of the detection network. Second, the CBAM-YOLOV5 algorithm for downhole foreign object detection was constructed by introducing a convolutional block attention model into the YOLOv5 detection network to enhance the saliency of foreign object targets in images. However, the detection performance and stability of this algorithm may be affected to some extent when dealing with extremely complex or poorly illuminated underground coal mine images, and further optimization is still required. Guan Ling et al. [6] proposed a CSPShuffleNet structure using a fusion cross-stage structure and channel shuffling strategy, introduced a multihead attention mechanism into the neck network, applied separated convolution to replace the traditional convolution in the head network, and synthesized the improved YOLOV4-Tiny algorithm to solve the problem of intrusion detection in a rail line environment. However, further improvement in the detection accuracy and robustness of the algorithm when facing complex and changing track line environments, such as extreme weather and obstructions, is still necessary. He Zifen et al. [7] designed a CSPTNet algorithm for the nighttime detection of airport runway foreign objects incorporating a self-attentive feature embedded by replacing the bottleneck module with the transformer bottleneck module, introducing a multihead self-attention mechanism, and changing the IOU loss function to the CIoU loss function. However, the algorithm may exhibit reduced detection sensitivity and accuracy when dealing with faint reflections or hidden foreign objects on airport runways at night, and additional in-depth research and optimization are still required.

In addition to the aforementioned foreign object detection algorithms based on target detection, numerous researchers and scholars have taken advantage of the continuity of video information to propose interframe computation and tracking for detection. For example, Wang Linfeng et al. [8] introduced the interframe differential optimization algorithm for weighted evaluation to achieve multiframe sequential recognition and proposed an interframe differential optimization method for detecting foreign objects on foggy tracks. However, their method may experience difficulty in accurately distinguishing foreign objects from the background under foggy conditions when fog concentrations are extremely high or when the foreign objects are similar in color to the background, resulting in limited detection. Weiming Liu et al. [9] collected background and detection images when a train stops at a platform, extracted the feature information of the images to obtain the feature pyramid through the encoding part of the network, and connected the feature maps of the two images. They then computed the feature differences to acquire the foreground heat maps of the images to be detected by the decoding part and finally obtained the detection results through threshold segmentation and contour screening. They proposed combining semantic segmentation with the background reference foreground detection method. However, their method may face the problem of decreasing detection accuracy or increasing false detection rates when dealing with complex and changing station environments, such as dense crowds and changing light. Wang Guoyi et al. [10] employed the velocity-assisted linear interpolation of positional information to obtain the most relevant frames between a background template image library and an image sequence to be detected. The image was then divided into different subblocks, and adaptive weights were set by using the complexity of the image texture within each subblock in combination with the ORB algorithm for feature point extraction. Furthermore, the feature points of the subblocks were normalized to the original image and aligned with the most relevant frames of the background template library for difference. The foreign object was then obtained. Finally, a kernel correlation filter was introduced to track multiple foreign object targets. However, this method may lead to inaccurate feature point extraction when dealing with high-speed moving train images with train vibration and image blurring. This effect, in turn, affects the accuracy and stability of foreign object detection.

Given that 3D LiDAR point cloud data are also often used for foreign object detection, Zhu and Hyyppa [11] employed airborne and mobile laser scanning techniques to model the railway environment in 3D. The method constructs a detailed 3D model of the railway environment from high-precision laser scanning data. The model provides intuitive and accurate data support for the maintenance, management, and planning of railway infrastructure and helps improve the efficiency and safety of railway operations. However, this method may face the problems of large computational volume and long time-consumption in data processing and model construction, which requires high hardware resources. Zhangyu et al. [12] proposed a camera

and LiDAR data fusion method for railway object detection. This method improves the accuracy and robustness of detection by fusing camera and LiDAR data. In the railway environment, the system can identify and track railway objects in real time, providing strong technical support for railway traffic safety. The method provides new ideas and solutions for railway safety monitoring. However, the performance of the method may be affected by environmental factors, such as extreme weather and changing light conditions, which may have an effect on the stability and accuracy of the system.

Meanwhile, some researchers have solved the foreign object detection problem for specific railway monitoring scenarios by using different neural network structures. Tao et al. [13] presented a feature fusion refinement neural network for the detection of railway traffic objects in shunting mode. This network improves the accuracy and robustness of detection by fusing multiple features. In a complex and changing shunting environment, the system can effectively identify and track railway traffic objects, providing a strong guarantee for railway traffic safety. However, the method has high computational complexity and may need to run on high-performance hardware to meet real-time requirements. Zhang et al. [14] proposed a real-time detection method based on deep learning for the problem of foreign object detection in complex open railway environments. This method utilizes the powerful feature extraction capability of the deep learning model to achieve the efficient and accurate detection of railway foreign objects. In the complex and changing railway environment, the system can identify and warn against potential safety hazards in real time, improving the safety of railway operations. However, the detection performance of the method may suffer under extreme weather or highly variable lighting conditions. Chen et al. [15] discussed a semisupervised learning method for the problem of foreign object detection in railway ballasted track beds. The method uses a small amount of labeled data and a large amount of unlabeled data to improve the generalization ability of the detection model by semisupervised learning. The system can accurately identify and locate foreign objects in the special environment of ballasted track beds, providing strong support for the maintenance and management of railway tracks. However, this method has certain requirements regarding the quality and quantity of unlabeled data that may affect the training effect of the model. Guo et al. [16] used generative adversarial networks to generate images of high-speed railway intrusions. This method provides rich data support for railway safety monitoring by simulating foreign object images in a high-speed railway environment. The generated images have a high degree of realism and diversity, helping improve the training effect and practical application of the railway safety monitoring system. However, the method may be affected by the training data when generating images, resulting in some differences between the generated images and real environment. Feng et al. [17] established an efficient foreign object recognition model for rail traffic. The model achieves the fast and accurate recognition of rail traffic foreign objects through real-time railway region extraction and object detection. In the complex rail environment, the system can efficiently identify and deal with potential safety hazards, providing a strong guarantee for the safe operation of rail transport. However, this method may require long training times and high amounts of computational resources when dealing with large-scale data. Wang et al. [18] investigated a detection method for neural network foreign objects that maintain differential privacy on distributed devices. This method protects the privacy security of data by adding noise on the basis of heterogeneity while achieving the accurate detection of railway foreign objects. In a network composed of distributed devices, the system can ensure the safe transmission and processing of data, providing a reliable guarantee for railway safety monitoring. However, this method may have some effect on the detection accuracy while increasing noise.

The above results are mainly for foreign body intrusion and other conditions of track state detection. No algorithm specifically for the transport of coal in open wagons exist, indicating that some shortcomings persist. Datasets dedicated to the detection of the operating state of open wagons do not exist and cannot be directly applied to the algorithm of the task . Moreover, the detection capability and adaptability of current algorithms for the complex and changing internal states of carriages, especially those involving multiple dynamic changes, still need to be strengthened. In particular, for multitarget detection with similar features inside a carriage, the structural redundancy, poor detection accuracy, and other problems of the current detection algorithm network will lead to missed and wrong detections. Therefore, the algorithmic model still requires further in-depth research and exploration. In this study, a high-precision open-wagon transport detection algorithm is proposed for the internal state detection of C80B carriages on the Shendong Dachen line. The main contributions of this work include constructing and accurately labeling an open-wagon transport image dataset; introducing a parameter-free attention residual structure (SARM) into the You only look once X (YOLOX) neck network to improve the ability of acquiring the location information of feature maps; adding a self-adaptive spatial feature fusion module (SASFF) to improve multiscale feature utilization rates; adopting a highly perceptive detection structure (DCBS); and enhancing target feature information through the expansion of perceptual field utilization. Finally, the accuracy of the algorithm proposed in this study is verified on a publicly available dataset and the open-vehicle transport state dataset.

The remainder of this study is organized as follows: The structure of the foreign body detection algorithm and construction of the network model of the detection algorithm are described in Section 3. The comparison and ablation experiments on the algorithm using a publicly available dataset; the verification of the feasibility, excellent performance, and detection accuracy of the algorithm in real application scenarios; and qualitative and quantitative analyses on the datasets of the application scenarios are reported in Section 4. The last section summarizes the study and provides relevant conclusions.

## 3. Methodology

This study, which aims to address the poor detection of foreign objects and redundancy of the network structure of most algorithms, proposes an efficient detection algorithm for open-truck transport images with the network structure shown in Fig. 1. The overall backbone network of the algorithm adopts the CSPdarnet53 structure that incorporates the attention of SimAM. The neck network architecture adopts the path aggregation feature pyramid network (PAFPN) structure, and the head network adopts the decoupled detection head structure. The image to be detected

passes through the backbone network, and the SARM structure fused with SimAM attention is designed to enhance the feature extraction of the image to be detected and provide effective feature information for subsequent learning. The neck structure is designed with a lightweight PAFPN–SASFF structure for the fusion of the target features to increase the reliability of target feature information. The convolution operation (Conv) in the detection network is replaced with the introduction of dilated convolution into the detection network to construct the DCBS structure in combination with the above SASFF and thus achieve the optimal fitting of the model and obtain the final detection results. The overall structure is shown in Fig. 1.

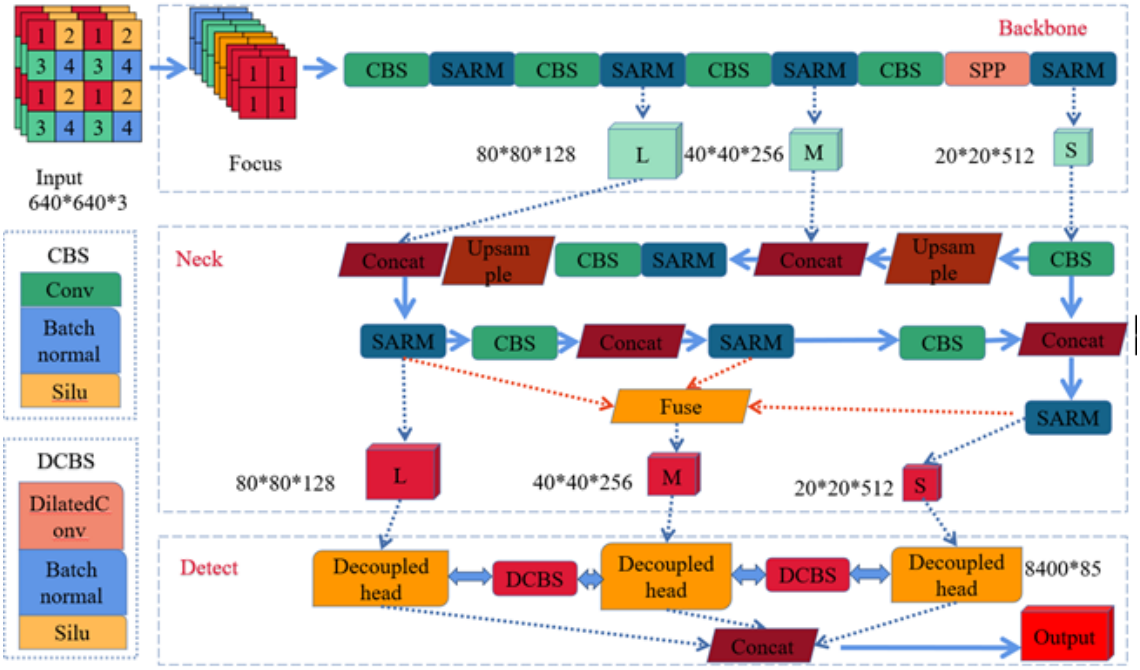


**Fig. 1.** Schematic of the overall structure

### 3.1 Parameter-free attention residual structure

The parameter-free attention residual structure is a network structure inspired by the cross-stage partial network (CSPNet). First, the original input features are divided into two branches: first through Conv then through the batch data with the channel normalization (batch normalization [BN]) layer. Subsequently, the CBS module is constructed by using the commonly used sigmoid linear unit (Silu) activation function. The CBS module halves the number of input feature channels. Subsequently, the residual bottleneck structure (bottleneck) is constructed by using the first branch. Next, the two branches are spliced together, and the residual bottleneck structure is constructed by the parameter-free attention mechanism. After the CBS module, the number of input feature channels is halved. The first branch is then used to build the residual bottleneck structure (bottleneck). Subsequently, the feature vectors of the two branches are spliced, and the three-dimensional attention weights for the feature map are inferred through the parameter-free attention mechanism to pay attention to deep target features. The deep target feature vectors are then derived by fusing the two branches. Finally, the obtained feature vectors are subjected to Conv then passed through the BN layer. After this step, the CBS module is used and the final feature vectors are obtained. The overall module structure is shown in Fig. 2.

The SimAM attention mechanism is implemented primarily by evaluating the importance of each neuron. For each neuron, an energy function that measures the linear divisibility of the target neuron and other neurons is defined:

$$e_t(\omega_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \hat{x}_i)^2 + (y_t - \hat{t})^2 \quad (1)$$

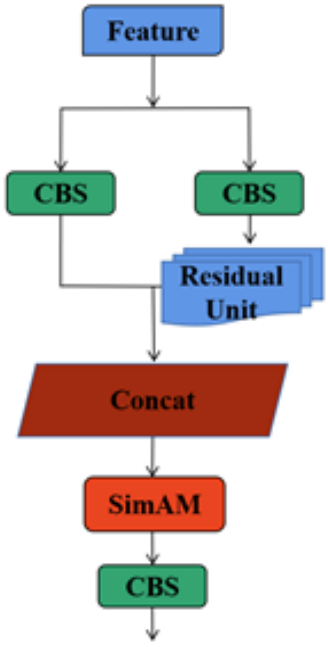


**Fig. 2.** Schematic of the SARM structure

In Equation (1), $t$ represents the target neuron of the input feature $X \in R^{C\times H\times W}$, $x_i$ represents the other neurons, $i$ is the index on the spatial dimension, $\omega_t$ represents the weight of the neuron when it is transformed, and $b_t$ represents the bias of the neuron when it is transformed. $y$ represents the binarized labels, which are also known as $-1$

and 1, and $M = H \times W$ represents the number of all neurons in the channel.

By using $\hat{t} = \omega_t t + b_t$ and $\hat{x}_i = \omega_t x_i + b_t$ and assigning values to $y_0$ and $y_t$ and adding regular terms, we obtain

$$e_t(\omega_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(\omega_t x_i + b_t))^2 + (1-(\omega_t t + b_t))^2 + \lambda\omega_t^2 \tag{2}$$

In Equation (2), $\lambda$ represents the coefficients of the canonical term, and the following relationship can be obtained from the above equation:

$$\begin{aligned} \omega_t &= -\frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \\ b_t &= -\frac{1}{2}(t+\mu_t)\omega_t \end{aligned} \tag{3}$$

In Equation (3), $\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i$, $\sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - u_i)^2$.

Therefore, substituting $\omega_t$ and $b_t$ into Equation (1) yields the minimum energy equation:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{4}$$

In Equation (6), a low energy is indicative of the high importance of the neuron t as distinct from peripheral neurons. Therefore, the importance of neurons is inversely proportional to $e_t^*$. Hence, for feature image $X$ and enhanced $\tilde{X}_S$, the following relationship defining the energy function as $E$ exists:

$$\tilde{X}_S = sigmoid(\frac{1}{E}) \odot X \tag{5}$$

### 3.2 Simple adaptive feature fusion pyramid

The simple adaptive feature fusion pyramid structure is optimized for the ASFF structure, which proposes an adaptive feature fusion method that allows a network to learn automatically to filter out useless information from other layers and retain useful information to fuse features efficiently. Specifically, for features in a particular layer, other layer features are first resized to the same size as the current layer before fusing the other layer features. The model is then trained to learn the optimal fusion method (i.e., to learn the weight share of the features in each layer at the time of fusion). In this study, the feature fusion for the second and third layers in ASFF is canceled, and feature fusion and learning information are only performed for the first layer, which is targeted to perform multisize target feature enhancement to construct SASFF and effectively reduce the number of parameters.

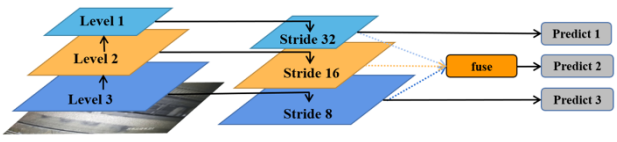


**Fig. 3.** Schematic of the SASFF structure

The feature fusion in Fig. 3 can be represented by the following equation:

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{1\to l} + \beta_{ij}^l \cdot x_{ij}^{2\to l} + \gamma_{ij}^l \cdot x_{ij}^{3\to l} \tag{6}$$

In Equation (6) $x^1, x^2, x^3$ represent the characteristics of the first, second, and third tiers, respectively. $\alpha^1, \beta^2, \gamma^3$ represent the weighting parameter. l stands for the corresponding level, which in this study is set to be $l = 2$, stating that the variables satisfy the following relational equation:

$$\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1, \quad \alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}^l}^l}}{e^{\lambda_{\alpha_{ij}^l}^l} + e^{\lambda_{\beta_{ij}^l}^l} + e^{\lambda_{\gamma_{ij}^l}^l}} \tag{7}$$

In Equation (7), $\lambda$ is the result obtained by 1 × 1 convolution acting on the upper level of the feature, which can be updated by backpropagation for learning.

### 3.3 Detection network

The detection network optimizes the decoupling head design in YOLOX. The structure is designed to reduce the number of feature channels relative to each level of features incoming from the backbone network by first passing through a 1 × 1 convolutional layer then adding two parallel branches, each consisting of two 3 × 3 convolutional layers for classification and regression, i.e., decoupling the commonly used prediction header into classification and regression headers, wherein the regression header is divided into localization and edge regressions. Given that a suitable receptive field for classification will yield suitable features for classification, by using cavity convolution, features are generated by using cavity convolution with a dilation rate of [2,2,2]. This approach effectively expands the receptive field of the layer with cavity convolution. Fig. 4 shows the schematic of cavity convolution.

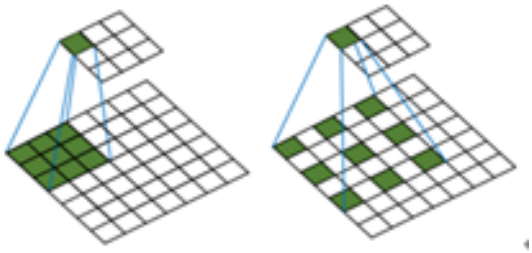

**Fig. 4.** Dilated convolution

Dilated convolution is the insertion of zero values between the neighboring weights of a standard convolution kernel. Therefore, it is also known as dilation or porous convolution. This design can expand the sensory field of the neural network and will not increase weight parameters. The zero value will not be involved in Conv to achieve the use of few parameters and a high amount of background information for the purpose of feature extraction. Dilation rate is an important parameter in null convolution. It determines the number of zero values inserted into the convolution kernel and controls the distance between adjacent nonzero weights when the convolution kernel performs Conv. Its expression is

$$y(m,n)=\sum_q\sum_p x(m+\lambda p+\lambda q)h(p,q) \quad (8)$$

In Equation (8), x represents the input value, y represents the output characteristics, $h(p,q)$ represents the weights of the positions in the convolution kernel, $(m,n)$ represents the coordinates of the pixels in the image, and $\lambda$ represents the dilatancy parameter of the cavity convolution. The resulting detection structure is shown in Fig. 5.

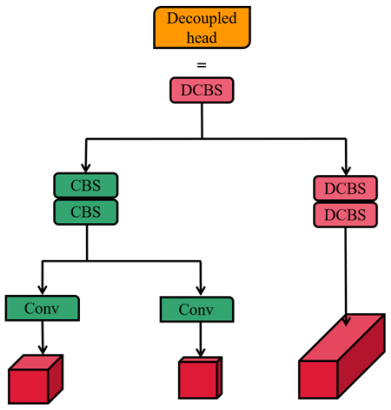


**Fig. 5.** Schematic of the structure of the decoupling head

The structure of the DCBS using the null convolution and the BN layer paired with the Silu activation function in Fig. 5 is indicated in the lower left corner of Fig. 1. Such a structure can increase the sensory field of convolution without increasing the parameters to increase the utilization of feature information.

## 4. Result Analysis and Discussion

### 4.1 Experimental environment and process

The hardware environment in the experiments in this study is an Intel Core i7-10700 CPU with an NVIDIA GeForce RTX 3060ti graphics card using the Python 3.8 programming language, PyTorch 1.8.1 deep learning framework, and Cuda 10.2 environment. The datasets used for the experiments are the Common Objects in Context 2017 (COCO2017) dataset and the open-car transport state dataset. The training parameters are all adopted from Table 1. In the analysis of the experimental results, mean average precision is employed for the quantitative analysis of the algorithm's performance. The graph of the final detection results of the algorithm is shown for qualitative analysis.

**Table 1.** Training parameters

| Indicators | Parameter setting |
|---|---|
| Batch size | 16 |
| Epoch | 300 |
| Weight decay | 0.0001 |
| Hyperparameter | 0.9 |
| Optimizer | Adam |
| Learning rate | 0.0004 |

### 4.2 Analysis of the experimental results for the COCO2017 dataset

The COCO2017 dataset is a dataset provided by the Microsoft team that can be applied for image recognition. It includes 118 287, 5000, and 40 670 training, validation, and test images, respectively, in 80 categories.

**Table 2.** Comparative experiment on the COCO dataset (bold font in the table indicates optimal indices)

| Method | Backbone | Size | FPS (3060ti) | MAP:50-95 |
|---|---|---|---|---|
| YOLOV5-S | CSPResNet50 | 640 | 86.1 | 37.4 |
| YOLOV8-S | ELAN | 640 | 88.2 | 44.9 |
| YOLOX-DarkNet | CSPDarkNet50 | 640 | 87.2 | 47.4 |
| Improved YOLOX-DarkNet | CSPDarkNet50 | 640 | **85.4** | **49.3** |

The relationship between the accuracies of the YOLOV5-S, YOLOV8-S, and YOLOX-DarkNet models and that of the improved YOLOX model on the COCO2017 dataset is validated in Table 2. The accuracy of the improved YOLOX model has improved by 11.9%, 4.4%, and 1.9% compared with those of the other models.

This study selects a railway scene, a classroom scene with repetitive feature targets, and a confusing crowded scene in the COCO2017 dataset as the validation images to demonstrate the effect of the proposed algorithm intuitively. Moreover, it takes the YOLOV8 and YOLOX algorithms with high accuracy, as shown in Table 2, to conduct comparative experiments in the environment presented in Table 1. The detection results are then visualized and analyzed.

Fig. 6 shows that the YOLOV8 and YOLOX algorithms cannot complete the detection of people at the edge of the red dotted line in the railway scenario and that of chairs in the red dotted line in the classroom scenario with repetitive feature targets. Moreover, they exhibit numerous omissions of crowded pedestrians in the dashed lines in the confusing crowded scenario. By contrast, the improved YOLOX can be used in the above scenarios. The improved YOLOX can detect the results in all the above scenarios.

### 4.3 Ablation experiment

The accuracy results in Table 3 are obtained by using the structure proposed in the study for ablation experiments, the COCO2017 dataset, and YOLOX-DarkNet as the baseline network trained with the training parameters in Table 1.

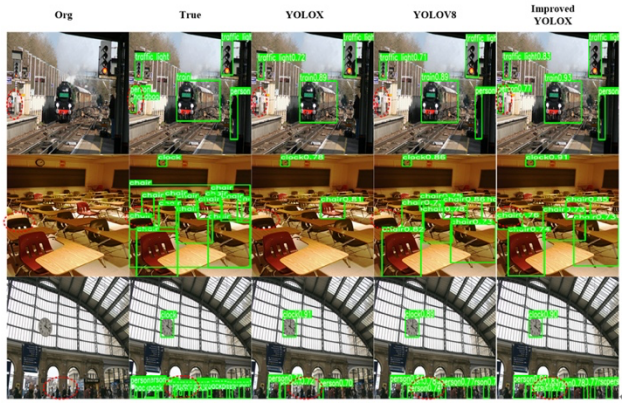


**Fig. 6.** Visualization of the comparative experiment on the COCO2017 dataset

Table 3 shows that each proposed module applied on the model has a certain improvement in performance. A heat

map, as shown in Fig. 7 below, is also generated to visualize intuitively the key areas of concern of the models.

**Table 3.** Ablation experiments (bold font in the table indicates the optimal indices)

| YOLOX-DarkNet | +SARM | +SASFF | +DCBS | mAP(%) | Param(M) |
|---|---|---|---|---|---|
| √ | | | | 47.4 | **9.0** |
| √ | √ | | | 47.8 | 9.41 |
| √ | √ | √ | | 49.1 | 9.47 |
| √ | √ | √ | √ | **49.3** | 9.75 |

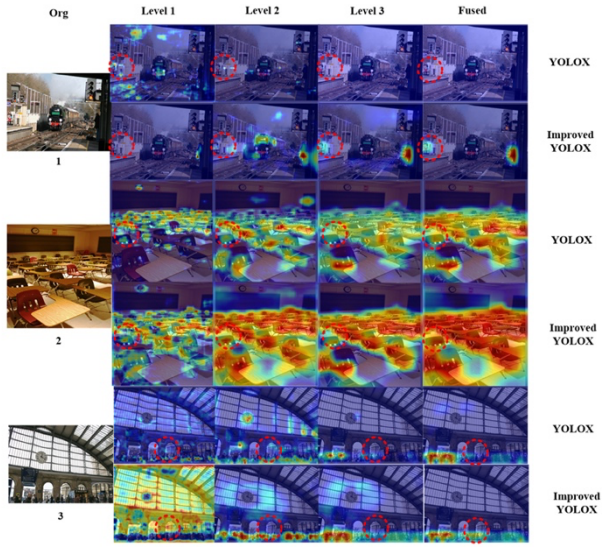


**Fig. 7.** Comparison of the heat maps of ablation experiments

Fig. 7 shows that compared with the improved YOLOX, the YOLOX algorithm has focused on the edge-side pedestrians at the red dotted line in the first- and second-level features in the first image but not on the third-level and fused features on the edge-side pedestrians at the red dotted line in the figure. The features at all levels in the second image have focused on the red dotted line, but the weights have remarkably decreased. Moreover, although the third features in the first and second layers of the image focus on the red dashed line, the features have considerably weakened after subsequent fusion. The improved YOLOX algorithm is effective in extracting features in the above cases. Therefore, compared with the preimprovement structure, the improved structure has higher performance for people at edges, the classroom scene with duplicate feature targets, and the confusing crowded scene.

### 4.4 Analysis of the experimental results for the open-truck transport state database

In this study, the labelme annotation tool is used for annotation, wherein 1267 training images and 370 validation images, constituting five categories, are included. The object classes are frozen bottom, red mine, empty car, with coal, and head.

**Table 4.** Comparative experiment on the Gondola transport status dataset (bold font in the table indicates the optimal indices)

| Method | Backbone | Size | FPS (3060ti) | MAP:50-95 |
|---|---|---|---|---|
| YOLOV5-S | CSPResNet50 | 640 | 84.4 | 69.6 |
| YOLOV8-S | ELAN | 640 | 87.6 | 73.8 |
| YOLOX-DarkNet | CSPDarkNet50 | 640 | 88.2 | 75.1 |
| Improved YOLOX | CSPDarkNet50 | 640 | 92.3 | 77.2 |

The relationship among the accuracies of the YOLOV5-S, YOLOV8-S, YOLOX-DarkNet, and improved YOLOX models on the open-vehicle transport state dataset is validated in Table 4. The accuracy of the improved YOLOX model has improved by 17.6%, 3.4%, and 2.1% compared with that of the other models.

This study chooses empty trucks, frozen bottom, coal with coal, snowy bottom, and red mine in the open-truck transport status dataset as the verification images to show the effect of the proposed algorithm intuitively. The YOLOV8 and YOLOX algorithms with high accuracy (Table 4) are used to conduct comparison experiments in the environment shown in Table 1. The detection results are visualized and analyzed.


**Fig. 8.** Visualization of the comparative experiment on the Gondola transport status dataset

The second column in Fig. 8 shows that the YOLOV8 and YOLOX algorithms cannot complete the detection in the frozen bottom scenario. Specifically, YOLOV8 identifies the frozen bottom scenario as an empty car, whereas YOLOX identifies the frozen bottom scenario as a car with coal. The third column shows that YOLOV8 can detect the case of a

car with coal. However, the detection frame is at the wrong position. Meanwhile, YOLOX experiences leakage detection. The fourth column shows that YOLOV8 can detect the snow scenario. However, the position of the detection frame is incorrect. Meanwhile, YOLOX can obtain the correct result. Column 5 shows that YOLOV8 identifies the red mine scenario as car with coal, and YOLOX has a deviation in the position of the detection frame. By contrast, the improved YOLOX can detect the correct results in the above scenarios.

## 5. Conclusions

This study started from the YOLOX model to improve the leakage and wrong detection phenomena and detection accuracy in the detection of foreign objects. Experiments proved that the proposed method improved detection accuracy with a small increase in the number of parameters and the detection of targets in open-car transport images to reduce the identification of the abnormal state of open cars manually. The following conclusions could be drawn:

(1) The global information of the feature map was enhanced by adding an efficient channel attention module to the neck structure of the YOLOX model, and accuracy improved by 0.8% as shown by the ablation experiments.

(2) Adding a new detection head to the detection component improved the accuracy of the detection of targets on different scales, as shown by the ablation experiments. Accuracy improved by 4%, and the convergence of the model accelerated.

(3) The method improved the accuracy of the improved YOLOX model compared with that of the other models by 17.6%, 3.4% and 2.1%, with a small increase in the number of parameters. This effect improved the accuracy of the detection of the running state of open cars.

This study performed analyses and experiments on application scenarios from publicly available and homemade datasets to propose an algorithmic model for target recognition and classification in foreign object and vehicle load detection. The constructed detection model is close to the actual coal mine logistics scenario while maintaining high efficiency and has a certain reference value for improving the efficiency of foreign object detection in carriages. However, given that this study adopted a small self-made dataset, the dataset suffers from the problem of limited and imbalanced categories, affecting the detection results of the algorithm. Moreover, the method in this study cannot obtain accurate results for targets with insignificant category characteristics and targets accounting for a small percentage of a dataset. Therefore, in follow-up work, we will conduct an in-depth study on such problems for targets with inconspicuous features and small samples in the dataset.



## References

[1] L. Wang, G. Liu, H. Lin, and Y. Zhao, "Design of quick detecting system for doped coal based on MSP430F449," *Transducer Microsyst. Technol.*, vol. 33, no. 8, pp. 97-100, Aug.2014 (in Chinese).

[2] Y. Tian, H. Yang, C. Hu, and J. Tan, "Moving foreign object detection and track for electric vehicle wireless charging based on Millimeter-Wave radar," *Trans. China Electrotech. Soc.*, vol. 38, no. 2, pp. 297-308, 2023 (in Chinese).

[3] Y. Chen, Z. Wang, and F. Zhou, "Infrared railway foreign objects tracking based on spatial location and feature generalization enhancement," *J. Beijing Univ. Aeronaut*. Astronaut., pp. 1-12, Apr. 2023. [Online]. Available: https://doi.org/10.13700/j.bh.1001-5965.2022.0974 (in Chinese).

[4] T. Ye, Z. Zhao, and Z. Zheng, "Research on the autonomous detection system for railway intrusion obstacles based on LAM-Net," *Chin. J. Sci. Instrum.*, vol. 43, no. 9, pp. 206-218, Apr.2022. DOI: 10.19650/j.cnki.cjsi.J2209185 (in Chinese).

[5] S. Hao, X. Zhang, X. Ma, S. Sun, H. Wen, and J. Wang, "Foreign object detection in coal mine conveyor belt based on CBAM-YOLOv5," *J. China Coal Soc.*, vol. 47, no. 11, pp. 4147-4156, Jul.2022. DOI: 10.13225/j.cnki.jccs.2021.1644 (in Chinese).

[6] L. Guan, L. Jia, and Z. Xie, "Research on lightweight model for railway intrusion Detection Integrating attention mechanism," *J. China Railw. Soc.*, pp. 1-10, Apr. 2023. [Online]. Available: http://kns.cnki.net/kcms/detail/11.2104.U.20220718.1414.002.html (in Chinese).

[7] Z. He, G. Chen, S. Wang, Y. Zhang, and W. Guo, "Detection of foreign object debris on night airport runway fusion with self-attentional feature embedding," *Opt. Precis. Eng.*, vol. 30, no. 13, pp. 1591-1605, Jul.2022 (in Chinese).

[8] L. Wang, H. Wan, Z. Liu, N. Qin, D. Huang, and Y. Zhang, "Method for Detecting Track Abnormal Objects in Foggy Weather Based on Inter-frame Differential Optimization Algorithm," *Urban Mass Transit*, vol. 25, no. 10, pp. 192-193+197, May.2022. DOI: 10.16037/j.1007-869x.2022.10.036 (in Chinese).

[9] W. Liu, J. Wen, Z. Zheng, Y. Dai, and H. Li, "DifferentNet: Neural network for foreign objects foreground detection in metro," *J. South China Univ. Technol. (Nat. Sci. Ed.)*, vol. 49, no. 10, pp. 11-21+40, Jun.2021 (in Chinese).

[10] G. Wang, Y. Sun, Y. Zhang, H. Lu, and W. Zhao, "Block detection and tracking algorithm of foreign objects debris in airport runway based on background alignment and difference," *J. Comput.-Aided Des. Comput. Graph.*, vol. 33, no. 3, pp. 413-423, Feb.2021 (in Chinese).

[11] L. Zhu and J. Hyyppa, "The use of airborne and mobile laser scanning for modeling railway environments in 3D," *Remote Sens*., vol. 6, no. 4, pp. 3075-3100, Apr. 2014. DOI: 10.3390/rs6043075.

[12] W. Zhangyu, Y. Guizhen, W. Xinkai, L. Haoran, and L. Da,, "A camera and LiDAR data fusion method for railway object detection," *IEEE Sens. J.*, vol. 21, Jun. 2021. DOI: 10.1109/JSEN.2021.3066714.

[13] T. Ye, B. Wang, P. Song, and J. Li, "Automatic railway traffic object detection system using feature fusion refine neural network under shunting mode," *Sensors*, vol. 18, no. 6, Jun.2018. DOI: 10.3390/s18061916.

[14] C. Zhang, H. Ding, Q. Shi, and Y. Wang, "Grape cluster Real-Time detection in complex natural scenes based on YOLOv5s deep learning network," *Agriculture*, vol. 12, no. 8, pp. 1242, Aug. 2022, doi: 10.3390/agriculture12081242.

[15] Z. Chen *et al.*, "Foreign object detection for railway ballastless trackbeds: A semisupervised learning method," *Measurement*, vol. 190, no. 1 pp.110757, Feb.2022. DOI: 10.1016/j.measurement.2022.110757.

[16] B. Guo, G. Geng, L. Zhu, H. Shi, and Z. Yu, "High-Speed railway intruding object image generating with generative adversarial networks," *Sensors*, vol. 19, no. 14, pp. 3075, Jul. 2019, doi: 10.3390/s19143075.

[17] Z. Feng, J. Yang, F. Li, Z. Chen, Z. Kang, and L. Jia, "An efficient foreign object recognition model in rail transit based on Real-Time railway region extraction and object detection," *J. Elect. Eng. Technol*., vol. 19, no. 6, pp. 3723-3734, Feb.2024. DOI: 10.1007/s42835-024-01805-y.

[18] M. Wang, Q. Y. Wang, Y. H. Zhang, Z. Xuan, F. Ning, and C. Feng, "Preserving differential privacy in neural networks for foreign object detection with heterogeneity-based noising among distributed devices," *J. Supercomput.*, vol. 80, no. 14, pp. 21447-21474, Aug. 2024. DOI: 10.1007/s11227-024-06243-1.