Improved Algorithm with YOLOv5s for Obstacle Detection of Rail Transit

Shuangyuan Li[1], Zhengwei Wang[2], Yanchang Lv[3], Xiangyang Liu[4]
Information Construction Office, Jilin Institute of Chemical Technology, Jilin, China[1,3]
School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, China[2,4]

***Abstract*—As an infrastructure for urban development, it is particularly important to ensure the safe operation of urban rail transit. Foreign object intrusion in urban rail transit area is one of the main causes of train accidents. To tackle the obstacle detection challenge in rail transit, this paper introduces the CS-YOLO urban rail foreign object intrusion detection model. It utilizes the improved YOLOv5s algorithm, incorporating an enhanced convolutional attention CBAM module to replace the original YOLOv5s algorithm's backbone network C3 module. In addition, the KM-Decoupled Head is proposed to decouple the detection head, and SIoU is applied as the loss function. Tested on the WZ dataset, the average accuracy increased from 0.844 to 0.893 .The research method in this paper provides a reference basis for urban rail transit safety detection, which has certain reference value.**

***Keywords*—Railroad track intrusion detection; CBAM (Convolutional Block Attention Module) attention; activation function; decoupling probe; loss function**

I. Introduction

The safety environment of urban rail transit is crucial for ensuring the safety of rail transit operations and the well-being of passengers.

With the gradual improvement of safety requirements in the urban rail transportation industry, the number of safety accidents in rail transportation has significantly reduced. However, due to natural and human factors leading to foreign object intrusion on railroad tracks, these incidents exhibit sudden and unpredictable characteristics. Relying solely on manpower cannot ensure comprehensive and timely detection. Therefore, it is necessary to establish a corresponding railroad track foreign object intrusion detection system for real-time monitoring of the running lines. This system aims to provide accurate alarm information to duty personnel, ensuring the safety of train operations. The detection of foreign object intrusions on urban rails has been a focal point of research, and an effective intrusion detection method for urban rail is of great significance for ensuring the overall safety of urban rail operations.

At present, there are two main methods for urban rail transit intrusion detection: contact detection and non-contact detection. Contact detection requires a large amount of hardware as a support and is troublesome to install, which is not easy to use on a large scale, and at the same time, when the equipment detects the intrusion of foreign objects it cannot be disposed of in time, which will seriously affect the safety of urban transportation. However, computer vision is an efficient non-contact intrusion detection method that is widely used in different transportation industries. It has the advantages of easy maintenance and intuitive results. However, the complexity and variability of urban environments and disturbances such as bad weather can lead to false alarm problems. Fortunately, with the development of deep learning algorithms, it has been possible to achieve high detection accuracy and reduce the false alarm rate to a certain extent. Deep learning algorithms excel at automatically learning complex features and patterns from large amounts of data, allowing computer vision systems to better distinguish between true intrusions and normal changes in the environment. However, deep learning algorithms are slow, take up a lot of memory, and require the support of a high-performance computer. There are many cameras in cities, but fewer cameras are used for intrusion detection in urban rail transit, and it is not practical to use a large number of high-performance computers. An efficient method for urban rail intrusion detection is needed in complex transportation scenarios.

II. Literature Review

Intrusion detection is an active research topic in the field of urban rail transit security. At present, some universities and research institutions in the United States are also conducting in-depth research in this area, such as the University of California, Berkeley, Yale University and so on. They have obtained more accurate and practical algorithmic models by introducing deep learning techniques and combining a large number of data sets for training, and target detection algorithms based on convolutional neural networks have come into being and have been gradually applied to a variety of urban environments, which are now broadly categorized into single-stage detection algorithms and two-stage detection algorithms. Common single-stage detection algorithms include RetinaNet [1], SSD [2] and YOLO [3] series. The algorithm regards localization and classification as a regression problem, realizes end-to-end detection, and has a faster detection speed, but its anchor mechanism based method [4], generates a large number of candidate frames in detection, and the number of candidate frames in which the target is detected is small. Common two-stage detection algorithms such as R-CNN [5], Fast R-CNN [6] have been widely used in this field. They aim to improve the detection performance by reducing the redundancy in the candidate frames generated by the anchor mechanism. Two-stage detection algorithms first screen out all positive samples and subsequently generate regions of interest (ROIs), and then in the second stage of these two-stage detection algorithms, the bounding frames generated in the previous stage are further

This work was supported in part by the Scientific Research Foundation of Jilin Province under Grant JJKH20230305KJ.

refined by performing region classification and positional adjustments on the regions of interest. The whole process requires iterative detection, classification and position refinement, and although two-stage detection algorithms tend to be slower compared to single-stage detection algorithms, they usually achieve higher detection accuracy. Zuoming [7] proposed a CNN model for road extraction that optimizes the extracted data to obtain comprehensive road features. Jiguang Dai [8] introduced a multi-scale CN N based road extraction method for remote sensing images. They used sub-image training model and combined it with residual linking to solve the resolution reduction and gradient vanishing problems in the extraction process. X. Zhang [9] developed a FCN network which utilizes a spatially consistent integration algorithm to determine the weights of the loss function used to extract the road regions. Xiangwen Kong [10] introduced a SM-Unet semantic segmentation network with a strip pooling module to enhance the road extraction performance. Hao Qi [11] proposed an MBv 2-DPPM model considering segmentation accuracy and speed. He and Ren [12] proposed a train obstacle detection method based on improved R-CNN. A new parallel upsampling structure and a context extraction module were added to the architecture to improve the accuracy of R-CNN to 90.6%. In order to enhance the ability to recognize small target objects, He [13] applied an enhanced Mask RCNN railroad obstacle detection method and proposed a new feature extraction network that incorporates a number of multiscale improvement techniques. The two-level target detection technique is the basis of the above method. Multiscale feature fusion by extending the sensitive domain and fusing shallow and deep features can improve the target recognition ability to some extent. However, real-time detection is not possible due to the area suggestion network. Zhang [14] proposed a high speed rail intruder detection algorithm based on YOLOv3 network. By improving this algorithm with FPN structure and extracting features with switchable hollow convolution, the false alarm rate of target detection is reduced, but the frames per second (FPS) is low, which does not satisfy the requirement of real-time detection. Literature [15] proposed a lightweight adaptive multi-scale feature fusion target detection network based on YOLOv3 to improve the performance of small target detection in complex environments, but there is a leakage detection phenomenon for occluded objects Literature [16] proposed a traffic sign recognition algorithm based on YOLOv5, comparing the detection results of the current common algorithms for the traffic signs. Overall, YOLOv5s algorithm performs better and faster in urban track detection. However, there are still some challenges, such as dealing with complex traffic situations in the city and enhancing the detection speed of the model, so the yolov5s algorithm model is chosen as the experimental base model and improved.

This paper proposes an improved method for foreign object intrusion detection in urban rail transit using YOLOv5s network as the base network. The method is designed to meet the need for real-time and accurate detection in urban rail transit environment. The YOLOv5s network model consists of several parts including Input, Backbone, Feature Fusion, Neck and Head. To enhance the input data, a mosaic [17] data enhancement method is used at the Input. This technique randomly selects four images from the image library and then rearranges them to generate a new image. By using the mosaic technique, the YOLOv5s network can benefit from the increased variation and diversity in the training data, which helps to improve its performance in object detection tasks; the proposed method also includes an improvement to the channel attention mechanism used in the CBAM [18] module. This improvement yields a new attention mechanism module called SCBAM. The SCBAM module aims to enhance the information interaction between different channels, further improving the network's ability to capture relevant features and improving the overall performance of tasks such as object detection. The proposed SCBAM convolutional attention module replaces the original C3 module, which effectively improves the model's ability to specifically characterize and recognize small target objects. Three detection layers are used in the YOLOv5s detection head [19], which are responsible for predicting large, medium and small targets, respectively. In addition, the network structure, output characterization method and boundary regression loss function of the YOLOv5s architecture were modified to improve its performance. These modifications aim to reduce the false detection and leakage rates in urban rail transit intrusion detection. By enhancing the feature extraction capability of the algorithm, the improved YOLOv5s network architecture provides more accurate and reliable detection results for long-range and small intrusion objects.

III. CS-YOLO Algorithm

A. YOLOv5s Structural Model

YOLOv5s target detection framework has a great performance advantage in large and medium-sized target detection, with real-time monitoring speed up to 140 fps and relatively high recognition accuracy. YOLOv5s consists of 3 main parts: backbone network, bottleneck layer and detection head, and the model structure is shown in Fig. 1.

1) Backbone network: The YOLOv5s algorithm combines the C3 and SPPF modules to enhance its performance. The C3 module focuses on reducing computation and increasing inference speed, optimizing the model's efficiency. On the other hand, the SPPF [20] module conducts multi-scale feature extraction on a single feature map, contributing to improved model accuracy. By leveraging these modules, the enhanced YOLOv5s algorithm achieves a balance between computational efficiency and accurate object detection. This makes it well-suited for various applications requiring real-time, precise recognition, and tracking of objects, thereby enhancing detection accuracy and speed for small and medium-sized objects on the track. C3 contains three standard convolutional layers and several bottleneck modules, the number of which is determined by the configuration file. The number of bottleneck modules is determined by the product of the n and depth_multiple parameters of YAML.C3 is the main module for learning the residual features and is divided into two branches, one using the specified bottleneck layer and three standard convolutional layers, and the other passing through only one of the basic convolutional modules, and finally the two branches are merged. After the Concat

operation, the activation function in the standard convolution module is SiLU. This is a function that applies the Sigmoid linear unit by *elements, which is* characterized as take-anywhere, continuous, and smooth, not a monotonic function, and is suitable for representing nonlinear features. The principle of the SPPF module is basically the same as that of spatial pyramid pooling [21] but uses a different design of pooling kernels. SPP uses 4 pooling kernels by default in YOLOv5s: 5×5, 9×9, 13×13, and 1×1. SPPF uses two pooling kernels by default in YOLOv5s: 5×5 and 1×1.Spatial pyramid pooling allows fusing feature maps at different scales and in the SPP layer Apply different pooling operations to multiple scales to capture information from different receptive fields. By pooling features at different scales, the SPP layer increases the perceptual field of the network, allowing it to efficiently process inputs of different sizes. This approach allows the network to process inputs of various sizes without the need to resize or crop the image to a specific size beforehand. The SPP layer enhances the flexibility and adaptability of the network in processing inputs of different resolutions.

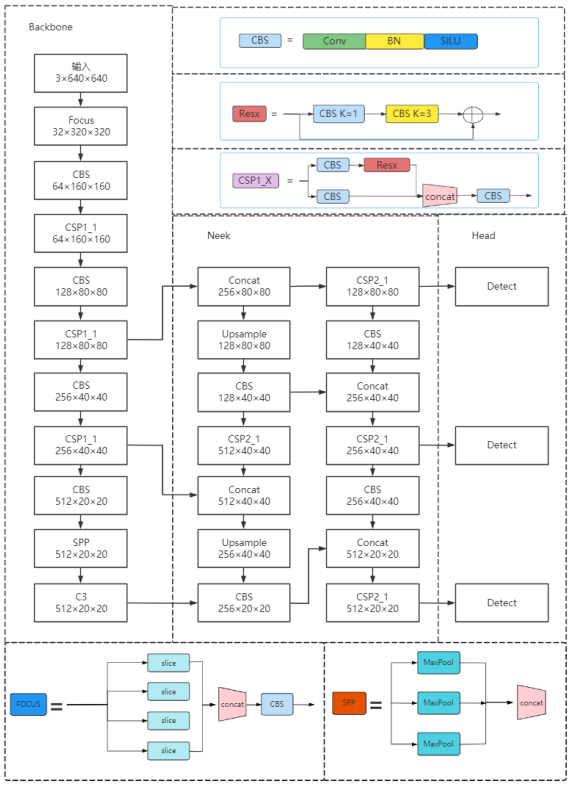


Fig. 1. YOLOv5s algorithm structure diagram.

2) Neck network: The YOLOv5s algorithm contains two key components, the Feature Pyramid Network (FPN) [22] and the Path Aggregation Network (PAN) [23]. The FPN utilizes a feature pyramid structure to integrate high-level semantic information with low-level features. This allows semantic knowledge to be passed top-down within the network. By combining the high-resolution details of low-level features with the contextual understanding of high-level features, FPN enhances the overall semantic representation of the network. On the other hand, PAN facilitates bottom-up transfer of localization information so that low-level information is propagated to higher levels in the network. PAN achieves this by fusing feature information from feature maps of different sizes. By integrating FPN and PAN, the YOLOv5s algorithm benefits from the transfer of semantic information and the effective utilization of multi-scale features. By integrating FPN and PAN, the YOLOv5s algorithm benefits from both the delivery of semantic information and the effective utilization of multi-scale features.

3) Head networks: The head networks as the detection part of the model and is mainly used to predict objects of different sizes from the extracted multi-scale feature maps. The output anchor frame mechanism extracts a priori frame scales by clustering and constrains the location of the prediction boundaries. The first is an 8-fold downsampled output with respect to the input image, which has a small perceptual range, preserves low-level high-resolution features, and facilitates the detection of small objects. The second is a 16-fold downsampled output with respect to the input image, which has a medium perceptual range and is suitable for detecting medium objects. The third is an output downsampled 32 times with respect to the input image, which has a large perceptual range and is suitable for detecting large objects.

B. Improved CBAM-based Attention Mechanism Module

The CBAM module is an attention mechanism that enhances spatial [24] and channel [25] attention while minimizing parameters. Its performance in classification detection on public datasets has shown improvement. However, when applied to the urban railroad track intrusion detection dataset in this paper, challenges arise. These challenges include a greater variety of vehicle classifications, a scarcity of samples, and the presence of vehicles with similar features. Consequently, the classification detection performance falls short of achieving the anticipated results. Fig. 2 illustrates the structure of the CBAM module.

The original channel attention mechanism employed global pooling operations, which included average pooling and maximum pooling on the input feature map along the width and height dimensions. These pooling operations aggregated information from all spatial locations within each channel. Following pooling, the generated features underwent processing through a multilayer perceptron (MLP). The MLP performed element-wise operations [26] to learn the importance weights for each channel. Finally, the output of the MLP passed through the Sigmoid [27] activation function to generate the final channel attention feature map. The Sigmoid function maps the values to a range between 0 and 1, representing the importance or activation level of each channel. In summary, the original channel attention mechanism comprises a pooling operation, an MLP for weighting, and a Sigmoid operation that produces the final channel attention feature map.

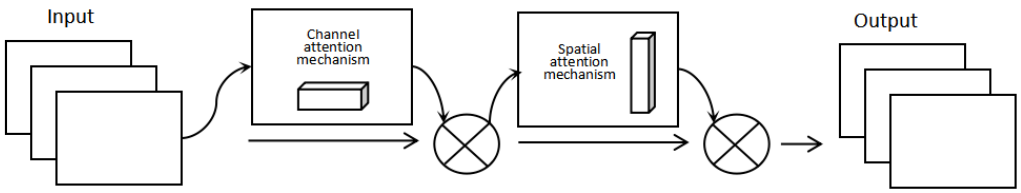


Fig. 2. CBAM structure diagram.

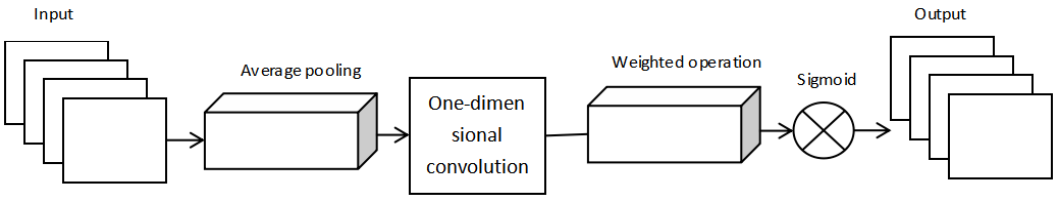


Fig. 3. SCAM structure diagram.

When information in the channel is globally and maximally pooled, this operation tends to neglect information interaction within the channel. In this paper, this paper optimizes and enhances the channel attention component of CBAM by discarding the maximum pooling operation. Instead, this paper represent the compressed features by aggregating the spatial information of the feature map through the average pooling operation. Drawing inspiration from the concept of ECA [28], this paper replaces the MLP network with a one-dimensional convolution operation in the improved version of the channel attention mechanism. Following one-dimensional convolutional processing, the ability to interact information between different channels is strengthened, and the output is subsequently simplified by a sigmoid function. In summary, as illustrated in Fig. 3, the improved channel attention module (SCAM) achieves local cross-channel interaction information of size K. This enhancement boosts the model's feature extraction capability, particularly for small and medium-sized objects.

C. KM-Decoupled Head

When recognizing targets on urban tracks, the task becomes more challenging due to the presence of multiple types of targets. To address this challenge and enhance the accuracy of localization and classification, this paper introduces a decoupled detection head known as KM-Decoupled Head. The primary idea behind this approach is to separate the feature channels for localization and classification tasks, specifically for bounding box coordinate regression and object classification. This decoupling enables a more precise estimation of bounding box coordinates, ultimately improving localization accuracy. Additionally, it ensures that the features used for classification are less influenced by changes in the localization task, leading to improved object classification. In essence, the goal of KM-Decoupled Head is to enhance target prediction in scenarios involving multiple types of targets and occluded targets. By decoupling the feature channels used for localization and classification, it enhances the accuracy of both tasks, resulting in more efficient target identification on urban tracks.

As illustrated in Fig. 4, the KM-Decoupled Head, a decoupled detection head, follows a specific process. First, a 1×1 convolution [29] is applied to the input feature map to reduce the number of channels and parameters generated. Subsequently, the output feature map is split into two branches to address the classification and localization tasks separately. For the classification branch, features are extracted using a 3×3 deep convolution [30]. The number of channel bits in the feature map is then adjusted to match the predicted number of target categories through a 1×1 convolution. On the other hand, the localization branch also employs a 3×3 deep convolution for feature extraction. After feature extraction, the resulting feature map is divided into two parts. One part predicts the center coordinates of the bounding box, along with the height and width of the bounding box. In summary, the KM-Decoupled Head employs a 1×1 convolution to reduce the number of channels and subsequently splits the feature map into separate branches for classification and localization tasks. The classification branch adjusts the feature map channels based on the predicted target categories, while the localization branch predicts the bounding box coordinates and confidence scores for target identification.

The decoupled structure of the KM-Decoupled Head offers several advantages over the coupled detector head, which integrates multiple types of information into a single feature map. Firstly, the decoupled design effectively avoids potential conflicts between different task requirements, thereby enhancing localization and classification capabilities. By separating the feature channels for each task, the model can focus on learning distinct representations for precise localization and accurate classification. Secondly, the decoupling head preserves information in each channel through depth and breadth operations, ensuring that valuable information is not weakened or diluted during processing steps. Additionally, this approach helps reduce computational complexity, thereby accelerating network convergence. Finally, due to the depth and breadth operations, the decoupling head achieves faster inference. In conclusion, the decoupled structure of the KM-Decoupled Head improves localization and classification by mitigating conflicts, effectively preserving channel information, reducing computational complexity, and achieving faster inference speed through depth and breadth operations.

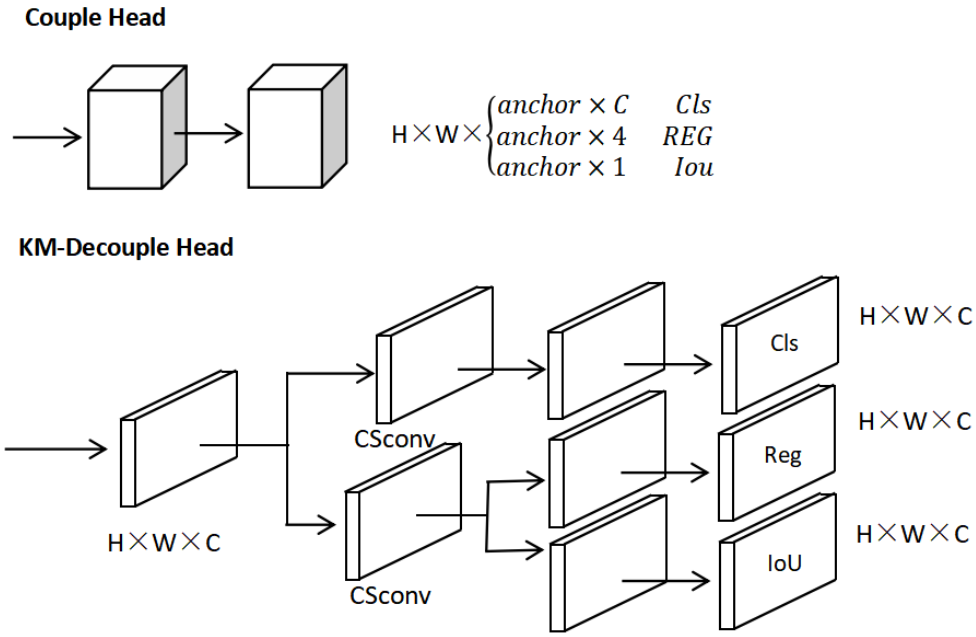


Fig. 4. Coupling head with KM-decoupled head.

D. Loss Function SIoU

The loss function in the YOLOv5s model is CIoU[31] :

$$L_{CIoU} = 1 - IoU + \frac{(b,\ b^{gt})p^2}{c^2} + \alpha v \quad (1)$$

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

$$\alpha = \frac{v}{(1-IoU)+v} \quad (3)$$

where IoU is the ratio of the intersection and union of the predicted and actual frames, b represents the center of the predicted point, b^{gt} represents the center of the actual frame, p denotes the Euclidean distance, c denotes the length of the diagonal of the bounding box formed by the predicted frame and the actual frame, α represents the weighting coefficient, and v represents the difference in the aspect ratio of the predicted frame and the actual frame.

CloU does not consider the direction of mismatch between the actual frames and the predicted frames, leading to slower convergence and lower efficiency. Therefore, we introduce a more balanced loss function, SloU [32], which increases the vector angle between regressions. We also redefine the cost function (penalty indicator), effectively reducing the degrees of freedom of regression, speeding up the convergence of the network, and further improving the accuracy of regression. SloU consists of four cost functions.

- Angle_Loss Angle_Cost minimizes the number of variables associated with distance. The formula is:

$$\lambda = 1 - 2 * \sin^2\left(\arcsin\left(\frac{c_h}{\sigma}\right) - \frac{\pi}{4}\right) \quad (4)$$

where σ is the distance between the center point of the real frame and the predicted frame, and ch is the height difference between the center point of the real frame and the predicted frame.

- Distance_LossDistance_Cost explores the distance of different bounding boxes at different centers as much as possible. The formula is:

$$\Delta = \sum_{t=x,\ y}(1 - e^{-\gamma pt}) \quad (5)$$

$$p_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2,\ p_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2 \quad (6)$$

"cw" and "ch" refer to the width and height of the minimum bounding rectangle of the actual box and the predicted box, respectively. "$b_{c_x}^{gt}$" and "$b_{c_y}^{gt}$" represent the coordinates of the center of the actual box, while "b_{c_x}" and "b_{c_y}" represent the center coordinates of the predicted frame.

- The shape loss Shape_Cost represents the deviation of the center of the predicted frame from the center of the real frame in an effort to obtain the optimal predicted frame. The formula is:

$$\Omega = \sum_{t=w,\ h}(1 - e^{-w_t})^{\theta} \quad (7)$$

$$\omega_w = \frac{|w - w^{gt}|}{\max(w,\ w^{gt})} \quad (8)$$

$$\omega_h = \frac{|h - h^{gt}|}{max(h,\ h^{gt})} \quad (9)$$

"w" and "h" represent the width and height of the predicted box, while "wgt" and "hgt" represent the width and height of the actual box,θ is the degree of concern for shape loss.

- Iou_Cost is the ratio of the intersection and the concatenation between the predicted and real boxes. The formula is:

$$\mathrm{IoU} = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|} \quad (10)$$

- the regression loss function SIoU is:

$$\mathrm{Loss}_{\mathrm{SIoU}} = 1 - \mathrm{IoU} + \frac{\Delta+\Omega}{2} \quad (11)$$

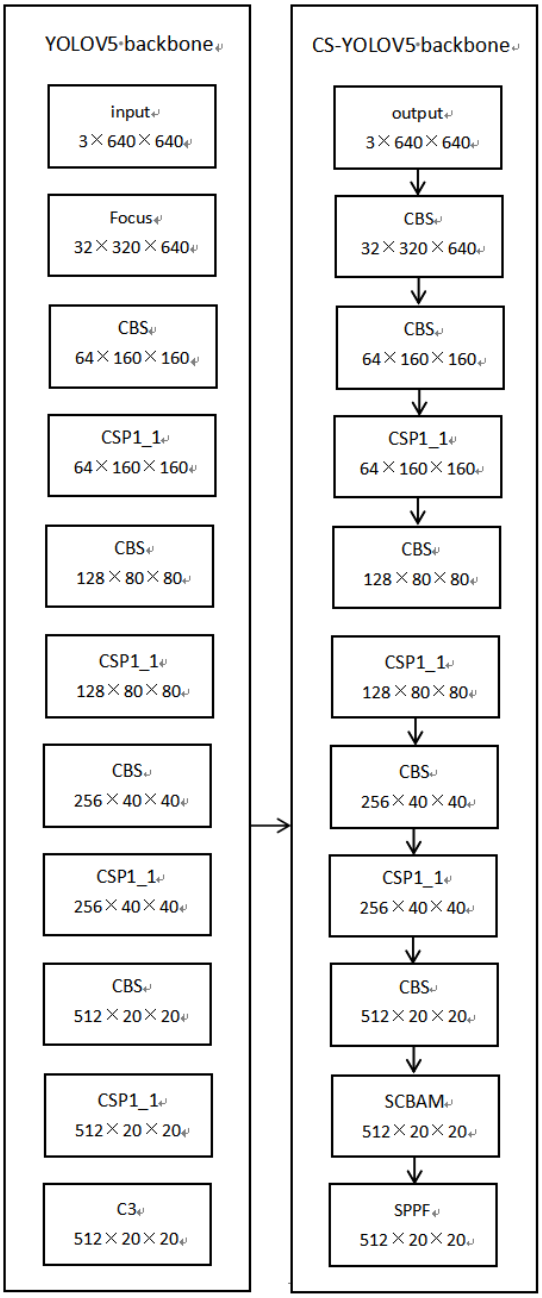


Fig. 5. Improved YOLOv5s backbone network.

E. Backbone Network Improvement

The enhanced SCBAM attention mechanism module replaces the original C3 module in the YOLOv5s backbone. It strengthens inter-channel information fusion using the channel attention mechanism and improves inter-channel information fusion through the spatial attention mechanism. The spatial attention mechanism is focused on detecting the target's location. By combining these two mechanisms, the output information becomes more concentrated on key features, enabling the model to prioritize important features during target detection. This enhances the feature extraction capability and ultimately improves the model's accuracy in detecting objects of various sizes. Additionally, this paper replaces the Focus slicing operation with a 6x6 convolution and substitutes the SPP module with the SPPF module to address the speed issue caused by parallel operations in the original model. The enhanced backbone network diagram is illustrated in Fig. 5

IV. Experimental Analysis

A. Experimental Environment and Parameter Settings

The specific experimental environment is shown in Table I.

TABLE I. Experimental Environment

Configuration name	Version/parameters
Operating system	Windows 11
Video storage	16GB
GPU	RTX4050
Memory	64GB
Python	3.7
Deep learning framework	pytorch1.11.0

During training, the network is trained using the SGD optimizer with an initial learning rate of 0.2. The learning rate is adjusted using the cosine annealing strategy. The batch size is set to 16 and the training process is performed for a total of 300 epochs.

B. Experimental Data Set

To assess the superiority of the improved target detection algorithm proposed in this paper, experiments were conducted on urban rail foreign object intrusion using the custom dataset, WZ-dataset.

As there is no publicly available dataset for foreign objects intruding on tracks in cities. This article develops a new dataset WZ-dataset. This dataset comprises 10 common types of foreign objects that may intrude on urban rail tracks, including pedestrians (person), bicycles (bike), electric bicycles (ebike), three-wheeled vehicles (tricycle), bottles, bags, cars, buses, trucks, and books. The WZ-dataset contains a total of 2,000 images, distributed across training, validation, and testing sets in an 8:1:1 ratio. Specifically, there are 1,600 images in the training set, 200 images in the validation set, and 200 images in the test set. The image size is consistently maintained around 600×800 pixels. To ensure a balanced sample size, each type of target is evenly distributed. The CS-YOLO algorithm model proposed in this paper is trained using the training and validation sets of the WZ-dataset (see Fig. 6). The model's performance is then evaluated by measuring the final average accuracy and detection speed on the test set.

Fig. 6. Example of WZ dataset.

C. Performance Index

The performance metrics, including Precision (P), Recall (R), Mean Accuracy (mAP), Parameters, Inference Time per Image, and Frames per Second (FPS), are utilized in the experiments to evaluate the proposed algorithm's performance. The calculation formula for these performance metrics is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{13}$$

$$\text{mAP} = \frac{\sum_{i=1}^{N} \text{AP}_i}{N} \text{AP} = \int_0^1 \text{P(R)dR} \tag{14}$$

$$\text{FPS} = \frac{\text{TotalTime}}{\text{NumFigure}} \tag{15}$$

When evaluating object detection models, TP refers to true positives, representing instances where the model predicts correctly. FP stands for false positives, indicating instances where the model predicts incorrectly. Average Precision (AP) measures the area under the precision-recall curve, offering a comprehensive evaluation of the model's performance at various thresholds concerning both precision and recall. AP calculates the average precision for each category, and Mean Average Precision (mAP) is the average of the APs across all categories. mAP provides an overall assessment of the model's performance across multiple object categories. mAP0.5 represents the average precision calculated at an IoU threshold of 0.5, considering a bounding box prediction correct if the Intersection over Union (IoU) is greater than or equal to 0.5. For mAP0.5:0.95, average accuracy is calculated across IoU thresholds ranging from 0.5 to 0.95 in steps of 0.05. Frames per Second (FPS) represents the number of images processed and detected by the network model per second. It reflects the speed and efficiency with which the network model performs object detection. A higher FPS value indicates that the model can process images more quickly, enabling real-time or near real-time applications.

D. Experimental Analysis of SCBAM Attention Module

The SCBAM (Spatial and Channel Bottleneck Attention Module) is an enhanced version of the CBAM (Convolutional Block Attention Module). This paper aims to assess the effectiveness of the SCBAM attention module through five sets of comparative experiments on the WZ dataset. The experimental results presented in Table II demonstrate the efficacy of the SCBAM attention module. On the WZ dataset, the average accuracy mAP0.5 improved by 3.4%, and mAP0.5:.95 improved by 6.3%. Importantly, these enhancements were achieved while maintaining good detection speed, ensuring real-time detection performance. These findings strongly support the effectiveness of the SCBAM attention module proposed in this paper.

TABLE II. COMPARISON TESTS OF DIFFERENT ATTENTION MECHANISMS ON THE WZ DATASET

Method	mAP@0.5	mAP@0.5:.95	FPS
Baseline	0.844	0.532	85
+SENet	0.856(+1.2)%	0.556 (+2.4)%	77
+CA	0.865 (+2.1)%	0.505 (-2.7)%	90
+ECA	0.867 (+2.3)%	0.574 (+4.2)%	73
+CBAM	0.872 (+2.8)%	0.587 (+5.5)%	83
+SCBAM	0.878 (+3.4)%	0.595 (+6.3)%	91

E. Ablation Experiments

CS-YOLO is an enhanced version of the YOLOv5s network model, featuring improvements in various aspects, including the attention mechanism, output characterization method, backbone network, and bounding box regression loss function. In this paper, a series of ablation experiments are conducted to assess the impact of the proposed modules on the algorithm's performance. These experiments analyze the performance optimization achieved by individual modules, as well as the combined effect when different modules are introduced in different orders. The objective is to determine the effectiveness and contribution of each module in enhancing the overall performance of the algorithm. The results of the ablation experiments are presented in Table III.

This paper introduces three improved methods incorporated into the YOLOv5s network model (indicated by a checkmark in Table III). These methods enhance the detection accuracy on the WZ dataset to varying degrees. The CS-YOLO algorithm proposed in this paper achieves a detection speed of 78 FPS, outperforming the original YOLOv5s algorithm. On the WZ dataset, this algorithm demonstrates a 4.9% improvement in mAP0.5 and an 8.9% improvement in mAP0.5:95. Experiments conducted on the WZ dataset validate the effectiveness of the proposed algorithm in urban rail transit detection. These experiments showcase the algorithm's capability to address the challenge of foreign object intrusion detection in complex urban rail transit environments.

F. Comparison Experiments

Table IV in this paper presents experimental results comparing the CS-YOLO algorithm with several other existing algorithms on the WZ dataset. The compared algorithms include the original YOLOv5s, SSD, Faster R-CNN, YOLOv3, YOLOv4-tiny, YOLOv4, YOLOv5m, YOLOX-tiny, YOLOX-S, YOLOv6-tiny, and YOLOv7-tiny. The experimental results comprehensively analyze the performance of the CS-YOLO algorithm in comparison with these popular algorithms. For detailed information about the experimental results and algorithm performance comparison on the RS dataset, please refer to Table IV in this paper.

The experimental results in Table IV demonstrate that the CS-YOLO algorithm proposed in this paper achieves higher detection accuracy on the WZ dataset compared to other mainstream algorithmic models. Notably, the parameters of the algorithm proposed in this paper are reduced by 130%, while the detection accuracy is improved by 2.7% compared to the YOLOv5m network with a similar structure. Although YOLOV4-tiny and YOLOv7-tiny exhibit higher detection speeds (111 FPS and 119 FPS, respectively), their detection accuracies are relatively low at 75.5% and 83.3%, making them unsuitable for complex urban rail transit environments. In conclusion, the CS-YOLO algorithm proposed in this paper demonstrates the highest detection accuracy while maintaining good real-time performance, offering a significant advantage over other algorithms under consideration.

TABLE III. CS-YOLO ABLATION EXPERIMENTS ON THE WZ DATASET

Group	SCBAM	KM-DHead	SIoU	mAP@0.5	mAP@0.5:.95	Parameters	FPS
1				0.844	0.532	7.1M	85
2	√			0.878 (+3.4)%	0.595 (+6.3)%	7.1M	91
3		√		0.878 (+3.4)%	0.553 (+2.1)%	7.4M	74
4			√	0.848 (+0.4)%	0.558 (+2.6)%	7.1M	94
5	√	√		0.884 (+4.0)%	0.590 (+5.8)%	7.4M	85
6		√	√	0.878 (+3.4)%	0.592 (+6.0)%	7.4M	80
7	√		√	0.881 (+3.7)%	0.589 (+5.7)%	7.1M	78
8	√	√	√	0.893 (+4.9)%	0.621 (+8.9)%	7.4M	78

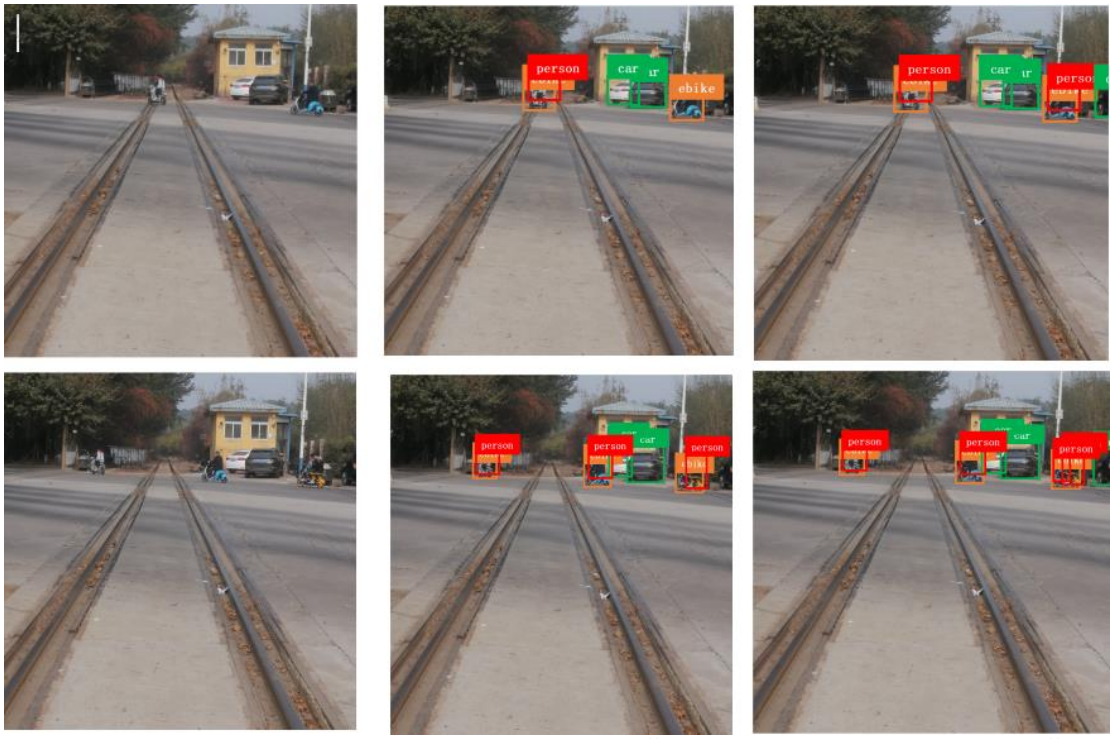
TABLE IV. COMPARATIVE EXPERIMENTS OF COMMON ALGORITHMS ON THE WZ DATASET

Method	mAP@0.5(%)	mAP@0.5:.95(%)	Parameters(M)	Inference(ms)	FPS
YOLOv5s	0.844	0.532	7.1	11.3	85
SSD	0.865	0.639	100.2	23.1	55
Faster R-CNN	0.711	0.453	136.2	46.6	23
YOLOv3-spp	0.833	0.512	9.56	12.3	81
YOLOv4-tiny	0.755	0.502	5.9	10.9	111
YOLOv4	0.817	0.528	244.8	55.2	35
YOLOv5m	0.866	0.636	20.9	19.9	70
YOLOX-tiny	0.822	0.565	5.1	16.8	43
YOLOX-S	0.848	0.591	9.0	23.5	25
YOLOv6-tiny	0.866	0.611	15.0	22.5	81
YOLOv7-tiny	0.833	0.567	6.1	8.3	119
CS-YOLO	0.893	0.621	7.4	14.8	78

In this paper, the detection performance before and after the improvement is visually compared, as illustrated in Fig. 7. The comparison results demonstrate that the CS-YOLO algorithm exhibits excellent detection performance in both sets of images. The results further highlight that the CS-YOLO algorithm effectively addresses misdetection and omission issues when detecting fuzzy and small targets in complex urban rail backgrounds compared to the initial YOLOv5s algorithm. With an FPS of 78, the CS-YOLO algorithm achieves higher detection accuracy by more accurately identifying common foreign object features while maintaining real-time detection speed. This capability meets the need for real-time and accurate detection in complex urban rail scenes.

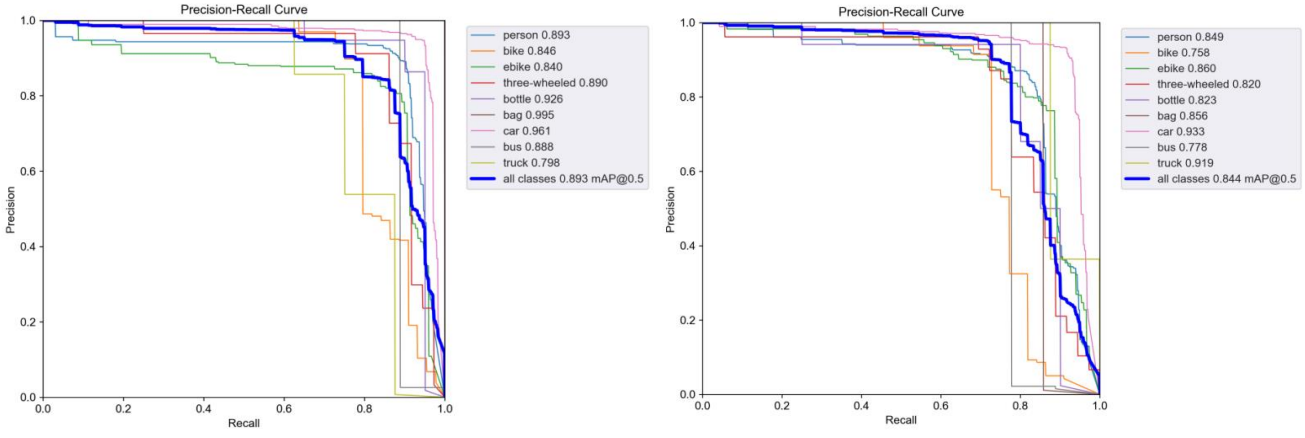
G. Analysis of Experimental Results

The accuracy recall curve of the initial YOLOv5s and CS-YOLO are shown in Fig. 8. The horizontal axis represents the recall rate, and the vertical axis represents precision. The inset box displays the detection precision for various common intruders, with the bolded blue line representing the average precision across all detection categories. Upon comparing the detection results of the two algorithms, it is evident that the CS-YOLO algorithm exhibits lower detection precision for tricycles and trucks compared to the initial YOLOv5s. However, it demonstrates higher detection precision for all other intrusions. The average accuracy is improved by 4.9% compared to the initial YOLOv5s algorithm, indicating the significance of the proposed improvements.



Original detection chart YOLOv5s detection effect CS-YOLO detection effect

Fig. 7. YOLOv5s and CS-YOLO detection effect comparison.



Graph of CS-YOLO detection results YOLOv5s detection result chart

Fig. 8. Graph of CS-YOLO and YOLOv5s detection results.

V. CONCLUSION

In this paper, we propose the CS-YOLO algorithm model for detecting foreign object intrusions in complex urban railways. The algorithm addresses the challenges of low accuracy and poor timeliness present in existing methods. Key enhancements include the introduction of the SCBAM attention mechanism module, replacing the original C3 module in YOLOv5s. Additionally, the algorithm features the KM-Decoupled Head decoupling head, utilizes SIoU as the loss function, and modifies the backbone network structure. In comparison to other mainstream target detection algorithms, This article improves the algorithm achieves superior detection accuracy, particularly in resolving issues of false positives and missed detections when dealing with concealed and small targets. Despite the enhanced accuracy, the algorithm maintains real-time detection speed, making it well-suited for detecting foreign objects on complex urban tracks. Future research should focus on optimizing the network for easier deployment on embedded GPU platforms, ensuring its applicability in real-world scenarios.

REFERENCES

[1] Qiu Zhiruo,Rong Shiyang,Ye Likun. YOLF-ShipPnet: Improved RetinaNet with Pyramid Vision Transformer[J]. International Journal of Computational Intelligence Systems,2023,16(1).

[2] Chen Renfei,Wu Jian,Peng Yong,Li Zhongwen,Shang Hua. Solving floating pollution with deep learning: a novel SSD for floating objects based on continual unsupervised domain adaptation[J]. Engineering Applications of Artificial Intelligence,2023,120.

[3] Li Jie,Li Sudong,Li Xiaoli,Miao Sheng,Dong Cheng,Gao Chuanping,Liu Xuejun,Hao Dapeng,Xu Wenjian,Huang Mingqian,Cui Jiufa. primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model.[J]. European radiology,2023,33(6).

[4] Said Yahia,Atri Mohamed,Albahar Marwan Ali,Ben Atitallah Ahmed,Alsariera Yazan Ahmad. Indoor Signs Detection for Visually Impaired People. Navigation Assistance Based on a Lightweight Anchor-Free Object Detector[J]. International Journal of Environmental Research and Public Health,2023,20(6).

[5] Zhang Wenming,Zhu Qikai,Li Yaqian,Li Haibin. MAM Faster R-CNN: Improved Faster R-CNN based on Malformed Attention Module for object detection on X-ray security inspection[J]. Digital Signal Processing,2023,139.

[6] Sang-Soo Baek,JongCheol Pyo,Yakov Pachepsky,Yongeun Park,Mayzonee Ligaray,Chi-Yong Ahn,Young-Hyo Kim,Jong Ahn Chun,Kyung Hwa Cho. Identification and enumeration of cyanobacteria species using a deep neural network[J]. Ecological Indicators,2020,115(C).

[7] She, Z.; Shen, Y.; Song, J.; Xiang, Y. Using the classical CNN network method to construct the automatic extraction model of remote sensing image of Guiyang road elements. Bull. Surv. Mapp. 2023, 4, 177‑182.

[8] Dai, J.; Du, Y.; Jin, G.; Tao, D. A Road Extraction Method Based on Multiscale Convolutional Neural Network. Remote Sens. Inf. 2019, 35, 28‑37.

[9] Zhang, X.; Ma, W.; Li, C.; Wu, J.; Tang, X.; Jiao, L. Fully Convolutional Network-Based Ensemble Method for Road Extraction from Aerial Images. IEEE Geosci. Remote Sens. Lett. 2019, 17, 1777‑1781.

[10] Kong, X.; Wang, C.; Zhang, S.; Li, J.; Sui, Y. Application of Improved U-Net Network in Road Extraction from Remote Sensing Images. Remote Sens. Inf. 2022, 37, 97‑104.

[11] Qi, H.; Li, Y.; Qi, Y.; Liu, L.; Dong, Z.; Du, X. Research on Track and Obstacle Detection Based on New Lightweight Semantic Segmentation Network. J. Railw. Sci. 2019, 45, 58‑66.

[12] He, D.; Ren, R.; Li, K.; Zou, Z.; Ma, R.; Qin, Y.; Yang, W. Urban Rail Transit Obstacle Detection Based on Improved R-CNN. Measurement 2022, 196, 111277.

[13] He, D.; Qiu, Y.; Miao, J.; Zou, Z.; Li, K.; Ren, C.; Shen, G. Improved Mask R-CNN for Obstacle Detection of Rail Transit.Measurement 2022, 190, 110728.

[14] Zhang, J.; Wang, D.; Mo, G. High-speed rail foreign body intrusion detection algorithm based on improved YOLOv3. Comput.Technol. Dev. 2022, 32, 69‑74.

[15] Ye, T.; Zhao, Z.; Wang, S.; Zhou, F.; Gao, X. A Stable Lightweight and Adaptive Feature Enhanced Convolution Neural Network for Efficient Railway Transit Object Detection. IEEE Trans. Intell. Transport. Syst. 2022, 23, 17952‑17965.

[16] Yingning Gao , Weisheng Liu. Complex Labels Text Detection Algorithm Based on Improved YOLOv5[J]. IAENG International Journal of Computer Science,2023,50(2).

[17] Mech Mario,Ehrlich André,Herber Andreas,Lüpkes Christof,Wendisch Manfred,Becker Sebastian,Boose Yvonne,Chechin Dmitry,Crewell Susanne,Dupuy Régis,Gourbeyre Christophe,Hartmann Jörg,Jäkel Evelyn,Jourdan Olivier,Kliesch LeifLeonard,Klingebiel Marcus,Kulla Birte Solveig,Mioche Guillaume,Moser Manuel,Risse Nils,RuizDonoso Elena,Schäfer Michael,Stapf Johannes,Voigt Christiane. author Correction: MOSAiC-ACA and AFLUX - Arctic airborne campaigns characterizing the exit area of MOSAiC.[J]. Scientific data,2023,10(1).

[18] Liu Hui,Yang Guangqi,Deng Fengliang,Qian Yurong,Fan Yingying. MCBAM-GAN: The Gan Spatiotemporal Fusion Model Based on Multiscale and CBAM for Remote Sensing Images[J]. Sensing Images[J]. Remote Sensing,2023,15(6).

[19] Cao Lianyu,Zhang Xiaolu,Wang Zhaoshun,Ding Guangyu. Multi Angle Rotation Object Detection for Remote Sensing Image Based on Modified Feature Pyramid Networks[J]. International Journal of Remote Sensing,2021,42(14).

[20] Buffington Matthew L., Garretson Alexis, Kula Robert R., Gates Michael W., Carpenter Ryan, Smith David R., Kula Abigail A.R.. Pan trap color preference across Hymenoptera in a forest clearing[J]. Entomologia Experimentalis et Applicata,2020,169(3).

[21] Kase Kina,Nakanishi Yosuke,Murono Shigeyuki,Yoshizaki Tomokazu. Comparison of Plain and Contrast-enhanced Computed Tomography for the Detection Head-and-neck Abscess[J]. Practica oto-rhino-laryngologica. Suppl.,2018,152(0).

[22] Nishiyama T.,Kumagai A.,Kamiya K.,Takahashi K.. SILU: Strategy Involving Large-scale Unlabeled Logs for Improving Malware Detector[J]. Proceedings - IEEE Symposium on Computers and Communications,2020,2020-July.

[23] Nagaoka Hiroshi,Ohtake Makiko,Karouji Yuzuru,Kayama Masahiro,Ishihara Yoshiaki,Yamamoto Satoru,Sakai Risa. Sample studies and SELENE (Kaguya) observations of purest anorthosite (PAN) in the primordial lunar crust for future sample return mission[J]. Icarus,2023,392.

[24] Yu Junwei,Shen Yi,Liu Nan,Pan Quan. Frequency-Enhanced Channel-Spatial Attention Module for Grain Pests Classification[J]. Agriculture,2022,12(12).

[25] Lee Haeyun,Cho Sunghyun. Image Restoration Network with Adaptive Channel Attention Modules for Combined Distortions[J]. Journal of the Korea Computer Graphics Society,2019,25(3).

[26] Kim Dohyun,Park Daeyoung. Element-Wise Adaptive Thresholds for Learned Iterative Shrinkage Thresholding Algorithms[J]. IEEE Access,2020,8.

[27] Coronavirus - COVID-19; Hanchuan People's Hospital Reports Findings in COVID-19 (The Challenges Of Urgent Radical Sigmoid Colorectal

Cancer Resection In A COVID-19 Patient; A Case Report)[J]. Medical Letter on the CDC & FDA,2020.

[28] Wang Xinkai,Jia Xu,Zhang Miyuan,Lu Houda. Object Detection in 3D Point Cloud Based on ECA Mechanism[J]. Journal of Circuits, Systems and Computers,2023,32(05).

[29] Yang K. Research on the Application of Time Convolution Series in Futures Price Forecasting[C]//Wuhan Zhicheng Times Cultural Development Co. ...Proceedings of 3rd International Conference on the Frontiers of Innovative Economics and Management (FIEM 2022).BCP Business & Management. BCP Business & Management, 2022:151-155.DOI:10.26914/c.cnkihy.2022.069652.

[30] Donghan X,Zhi W,Chunlin C, et al. Depthwise Convolution for Multi-Agent Communication With Enhanced Mean-Field Approximation.[J]. IEEE transactions on neural networks and learning systems,2022,PP.

[31] Yifan J,Dexin G,Shiyu Z, et al. A real-time fire detection method from video for electric vehicle-charging stations based on improved YOLOX-tiny [J]. . Journal of Real-Time Image Processing,2023,20(3).

[32] Yonghong L,Cheng Z,Zhiqiang Z, et al. Research on detection method of Tubercle Bacilli based on the improved YOLOv5.[J]. Physics in medicine and biology,2023,68(10).