

Article

Foreign Object Detection on Conveyor Belts in Coal Mines Based on RTA-YOLOv11

Liwen Wang *, Kehan Hu, Xiaonan Shi and Junhe Chen

School of Artificial Intelligence and Computing, Xi'an University of Science and Technology, Xi'an 710054, China; lmkin123@163.com (K.H.); shi_xn@xust.edu.cn (X.S.); changer_he@163.com (J.C.)

* Correspondence: wangliwen0429@xust.edu.cn

Abstract

To address the challenges of limited detection accuracy and the difficulty of deployment on edge devices caused by dust obstruction, low illumination, and complex background interference in coal mine conveyor belt foreign object detection, this paper proposes an improved algorithm model, RTA-YOLOv11, based on the YOLOv11 framework. First, a Receptive Field Enhancement Module (RFEM) is utilized to expand the field of view by fusing multi-scale perception paths, strengthening the network's semantic capture capability for subtle targets. Second, a Triplet Attention mechanism is introduced to suppress environmental noise and enhance the saliency of low-contrast foreign objects through cross-dimensional joint modeling of spatial and channel information. Finally, a lightweight detection head based on MBConv is designed, utilizing inverted bottleneck structures and re-parameterization strategies to compress redundant parameters and improve deployment efficiency on edge devices. Experimental results indicate that the mAP@0.5 of the improved RTA-YOLOv11 model is 4.0 percentage points higher than that of the original YOLOv11, with an inference speed of 79 FPS and a reduction in parameters of approximately 22%. Compared with algorithms such as Faster R-CNN, SSD, and YOLOv8, this model demonstrates a superior balance between accuracy and speed, providing an efficient and practical solution for intelligent mine visual perception systems.

Keywords: foreign object detection; YOLOv11; multiscale feature enhancement; small object detection; industrial vision intelligence

1. Introduction

Coal mine underground conveyor systems serve as critical hubs for energy transportation and operate for extended periods under extreme working conditions, including high humidity, low illumination, heavy dust, and complex background interference, which leads to a high level of safety risk [1]. Once foreign objects such as anchor rods, gangue, or tools intrude into the conveyor belt system, they can easily trigger a series of cascading failures, including belt tearing, transfer machine blockage, and motor overload, posing serious threats to personnel safety and equipment integrity, and causing substantial economic losses due to maintenance and production downtime. Therefore, achieving efficient and reliable foreign object detection has become an indispensable component of intelligent mine construction [2].

Conventional underground monitoring methods mainly rely on manual inspection, fixed cameras, and traditional image processing-based anomaly recognition. However, under low-contrast and high-interference environments, traditional approaches suffer



Academic Editor: Christos Bouras

Received: 5 January 2026

Revised: 23 January 2026

Accepted: 25 January 2026

Published: 29 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

from limited feature representation, poor generalization capability, and insufficient real-time performance [3], making them inadequate to meet the requirements of continuous, high-frequency, and accurate monitoring in modern mines.

In recent years, with the continuous demonstration of the significant advantages of the use of deep learning technology in the field of target detection and the continuous development of intelligent monitoring systems for coal flow transportation links in underground coal mines, many studies have focused on multiangle perception of the coal flow itself and the transportation state, and this method has become the core technical support for underground monitoring systems [4]. It has been widely applied in multiple subfields. For example, improvements in engineering deployment have been achieved in multiple mining areas through belt deviation and coal stacking fault diagnosis based on machine vision [5], damage detection based on conveyor belt images [6], coal flow contour positioning and registration based on point cloud data [7], intelligent centralized control of the main coal flow conveying system [8], no-load state identification [9], and coal flow anticongestion regulation in fully mechanized mining faces based on integrated decision-making [10]. However, most current research focuses on the conveyor chain itself or on the overall operating state of the transportation system, and the precise identification and determination of the location of “foreign objects on the surface of the coal flow” remain weak. Foreign objects are a potential threat and may cause serious consequences such as jamming, tearing, and false stopping. Especially in complex underground environments in which the target size is small, the background interference is strong, and various types of foreign objects are present [11], the detection of foreign objects imposes strict requirements on the model’s ability to represent small targets, restore image details, and reduce speed. Therefore, enhancing the accuracy, robustness, and adaptability of deployment of foreign object detection is not only a key point for reinforcing existing intelligent coal flow monitoring systems but is also among the core issues that must be dealt with to achieve higher reliability in multitask video perception systems in underground coal mines.

In terms of dataset construction, foreign object detection in underground coal mines often relies on real photos taken in specific mining areas. Owing to the high cost of collection and the difficulty of annotation [12], the number of public datasets containing such photos is limited, and the problem of category imbalance is serious, further aggravating the difficulty of model training. Researchers generally optimize model performance by introducing feature enhancement modules, attention mechanisms, lightweight convolutional structures, and improvements in loss functions [13]. However, most existing methods focus on performance improvement in a single module or at a fixed scale, making it difficult to simultaneously balance accuracy, efficiency and stability.

The primary research objective of this study is to design a high-precision detection model tailored for the complex underground environments of coal mines, specifically focusing on addressing the feature loss of small-sized foreign objects under low-illumination and dust-obscured conditions. Simultaneously, the research aims to mitigate the impact of limited computational resources on edge devices’ model performance, establishing an intelligent visual recognition model characterized by strong robustness and ease of industrial deployment.

On the basis of the foregoing considerations, this paper proposes a lightweight improved RTA-YOLOv11 algorithm for the detection of foreign objects in underground conveyor belts that is based on the YOLOv11 network. This method conducts systematic optimization around three core structures. It introduces a multibranch receptive field enhancement module (RFEM) [14] in the backbone and integrates dilated convolution and pooling paths to expand the receptive field range and enhance the semantic modeling ability for small targets. The triplet attention mechanism [15] is inserted into the main trunk

and neck structures to enhance the edge region response through cross-attention guidance and improve the recognition ability of low-contrast targets. In the detection head section, a combination of the MBConv (Mobile Inverted Bottleneck Convolution) and SE (Squeeze-and-Excitation Networks) modules [16] is adopted, and a reconfigurable parametric design that compresses the reasoning path is introduced, yielding a significant improvement in operational efficiency while maintaining end-to-end detection characteristics.

Experimental research results indicate that the RTA-YOLOv11 model proposed in this paper exhibits superior performance in foreign object detection tasks within complex underground backgrounds. Research findings reveal that the introduction of the RFEM to expand the effective receptive field significantly enhances the network's capability to capture minute targets; meanwhile, the Triplet Attention mechanism effectively suppresses false detections caused by dust and light interference through cross-dimensional joint modeling. Performance evaluations show that the algorithm achieves an mAP@0.5 of 89.3% on the relevant dataset, representing a 4.0 percentage point improvement over the original YOLOv11, while reaching an inference speed of 79 FPS and a reduction in parameters of approximately 22%.

This paper is organized according to the logical sequence of a research overview, model improvement principles, experimental analysis, and conclusions. First, related work regarding foreign object detection in underground coal mines is introduced, including traditional image processing methods and recent research progress based on deep learning. Subsequently, the network architecture of RTA-YOLOv11 and its primary improvement modules—the Receptive Field Enhancement Module, the Triplet Attention mechanism, and the lightweight detection head—are elaborated in detail. Then, the detection effectiveness of the model is validated through ablation studies and comparative analyses with mainstream algorithms. Finally, the research work of this paper is summarized, and potential future research directions are discussed.

2. Related Work

2.1. Foreign Object Detection Algorithms Based on Traditional Digital Image Processing

Prior to the application of deep learning in underground detection, traditional digital image processing techniques were the primary methods available for identifying foreign objects on coal conveyor belts. The basic process relies on manually defined rules to extract low-level visual features such as grayscale, texture, and edges from images and then uses algorithmic logic to locate abnormal regions in the images. This approach is characterized by its simplicity and low computational overhead.

Otsu [17] proposed an adaptive threshold selection method based on gray-level histograms, enabling automatic segmentation of foreground and background and providing a fundamental basis for early-stage foreign-object region extraction. Canny [18] developed a multi-stage edge detection framework that achieves stable edge extraction through Gaussian filtering, non-maximum suppression, and double-threshold hysteresis, and has been widely applied in foreign-object contour detection. Haralick et al. [19] extracted texture features such as contrast, correlation, and energy from gray-level co-occurrence matrices for foreign-object surface representation and classification. Lowe [20] introduced the Scale-Invariant Feature Transform (SIFT), constructing local feature descriptors invariant to rotation, scale, and illumination for robust foreign-object matching and recognition in complex environments. Sain [21] systematically formulated the statistical learning theory framework, providing a theoretical foundation for traditional feature modeling and classification algorithms.

Although the above methods are practical in single-background, well-lit, and foreign object-dominant scenarios, they rely heavily on manually designed features and lack deep

semantic expression capabilities, making them ill-suited for complex underground scenarios such as those with low-light conditions, multiple object occlusions, and background interference. In actual deployment, traditional algorithms are sensitive to the conditions under which the images are acquired, have weak generalization capabilities, and exhibit high rates of false positives and false negatives while also failing to meet the real-time requirements of high frame rates and edge-side deployment.

2.2. Foreign Object Detection Based on Deep Learning

In industrial production, deep learning foreign object detection technology is suitable for complex scenarios because of its powerful feature extraction capability. Li Gang et al. [22] addressed the problems of weak cross-layer interaction and insufficient semantic transmission by constructing an improved Hyper-YOLO framework that integrates DyCLAHE, Laplacian enhancement, ECA attention, and SimSPPF modules, thereby strengthening feature aggregation and improving the stability of small-object detection. Dong Haifeng et al. [23] developed a conveyor belt foreign-object recognition method based on an improved RegNet, introducing multi-scale dilated fusion attention to enhance feature representation under complex conditions. Jiang Feng [24] incorporated coordinate attention and adaptive anchor estimation into YOLOv5 to improve detection accuracy and stability in complex underground environments and further enhanced performance through illumination enhancement. Huang et al. [25] proposed the MSAM-YOLOv5 algorithm, which integrates multi-scale attention modules with depthwise separable convolutions to achieve high-speed real-time detection of foreign objects on conveyor belts in complex underground coal mine environments at 44 FPS.

Hao Shuai et al. [26] proposed a target detection algorithm based on CAM-YOLOv5 to address the difficulty of detection of foreign object images in underground conveyor belts due to coal dust interference, high-speed blurring, and uneven illumination. By introducing an attention mechanism and a deep backpropagation optimization strategy, the feature expression ability and the accuracy of detection of foreign object targets were effectively enhanced. Faced with the challenges of large differences in the sizes of foreign objects and high real-time requirements in complex coal mine scenarios, Mei Xiaohu et al. [27] constructed the Stair-YOLOv7-tiny model by improving YOLOv7-tiny, introducing a multi-level feature fusion structure that enhances the model's expression ability for medium- and low-resolution targets and improves the detection of small targets. Tang Jun et al. [28] designed the Faster-YOLOv7 network to address the issues of low accuracy of foreign object detection and slow model inference speed in low-light environments. This network, which integrates CLAHE image enhancement, the lightweight backbone of Mobilenetv3, and the Alpha-IoU loss function, significantly improves operational efficiency while ensuring detection accuracy. Wang Chao et al. [29] proposed an image enhancement strategy that integrates HSV space and the Retinex algorithm to address the issue of unclear foreign object features caused by excessive dust and weak light in underground coal mines. Combined with an optimized RetinaNet model structure, this strategy effectively improved the accuracy of detection of small target foreign objects in low-contrast images. Dai et al. [30] proposed a novel machine vision detection method that integrates an adaptive weighted multi-scale Retinex enhancement algorithm with a Multi-Layer Perceptron (MLP) network optimized by the Grey Wolf Optimizer (GWO), achieving an industrial field identification accuracy of over 95% for large foreign objects on coal mine belt conveyors.

Rongbin Yao's team improved the YOLOX model by rotating the decoupling head and the attention mechanism to solve the problem of false detection of the angle of slender foreign objects on the belt. The accuracy rate exceeded 93% in the laboratory environment, and the reasoning time was 25 ms [31]. Bingxin Luo et al. optimized the YOLOv5 algorithm,

using the combination of dark channel defogging and a lightweight design to solve the problem of the detection of noncoal foreign objects on high-speed belts in coal mines. The detection accuracy was 94.9%, and the parameters were reduced by 43.1% [32]. Yuanbin Wang et al. improved the SSD algorithm, balancing the detection accuracy and speed of coal mine belts through depth-separable convolution and GIoU loss, increasing the accuracy rate to 90.2% and the frame rate to 41 fps [33]. Di Miao et al. [34] designed a conveyor-belt foreign object detection system based on an improved NanoDet model, achieving 94.3% detection accuracy and 30 FPS real-time performance on ARM + Android edge devices. Zhang et al. [35] proposed an improved YOLOv4-based lightweight foreign object early-warning model for conveyor belts, reaching 93.73% accuracy and 70.1 FPS on a self-built dataset, and further demonstrated that increasing frame rate effectively alleviates motion blur. Dawid Szurgacz et al. [36] employed infrared thermal imaging to analyze temperature distribution, enabling early warning of mechanical failures caused by foreign objects on conveyor belts and providing complementary multimodal detection capability.

Overall, in the detection of foreign objects on underground coal mine conveyor belts based on deep learning, YOLOv11 demonstrates unique advantages over other mainstream algorithms in its detection speed, its accuracy, and its ability to generalize, providing a more effective solution to the problem of foreign object detection in industrial production.

3. YOLOv11 Network Improvements

YOLOv11 is an evolved version of the YOLO series, continuing the lightweight concept of YOLOv7/YOLOv8. Its overall structure includes an input end, a backbone network, a feature fusion layer and a decoupling detection head. The input end adopts multiple enhancement strategies, such as Mosaic stitching, color perturbation and scale transformation, and the image size is uniformly 640×640 . The backbone network is based on CSPDarknet and is constructed in five stages from P1 to P5. Downsampling is achieved through convolution with a step size of 2. Internally, CSP modules are stacked, featuring residual connections and channel division structures. The feature fusion component stacks FPN and PAN paths to construct top-down and bottom-up multiscale information flows and combines 1×1 convolution and bilinear interpolation to achieve up-down sampling. The detection head adopts an anchor-free decoupled structure, with independent classification and regression branches. It outputs three-scale feature maps corresponding to P3, P4, and P5. The final prediction includes category, position (cx, cy, w, h), and confidence information. The entire network uniformly uses the SiLU activation function and BatchNorm normalization, and the convolution is configured with padding = 1 to maintain consistent feature sizes.

On the basis of the original YOLOv11 structure, the research focus of this paper is on addressing practical engineering problems commonly encountered in underground conveyor belt foreign object detection, such as small-target semantic ambiguity, edge degradation, and deployment latency, by performing systematic optimization and adaptive improvements on the model. This design logic, through the organic integration of the convolutional layers, the attention mechanism, and the detection head, constructs the improved model RTA-YOLOv11.

Specifically, to alleviate the difficulty of adequately perceiving small objects in complex backgrounds, the RFEM receptive field enhancement module is introduced to expand the feature perception range; to address severe background interference and low target contrast in underground environments, the Triplet Attention mechanism is integrated to suppress irrelevant features and enhance salient regions from the spatial, channel, and cross-dimensional perspectives; meanwhile, in response to the strict constraints on model scale and inference latency in practical deployment, a lightweight MBConv detection head is

adopted to compress redundant computation while preserving representational capability. These modules collaboratively optimize detection performance from the perspectives of feature perception, feature selection, and inference efficiency, forming the overall design logic of RTA-YOLOv11.

As shown in Figure 1, in the Backbone architecture, to expand the model's receptive field and enhance its ability to capture features of small objects, this paper reconfigures the original YOLOv11 SPPCSP module and designs a receptive field enhancement module that integrates dilated convolutions with multiscale pooling paths. This structure enhances the context-aware range through parallel 3×3 dilated convolutions and 5×5 max pooling branches; simultaneously, it introduces feature reordering operations to compress redundant information flows, thereby improving the stability of deep semantic distributions while maintaining structural lightweightness. This module is deployed at the tail end of the backbone, explicitly reinforcing the deep modeling capability for foreign object regions < 32 px and expanding the adaptability to multiscale foreign objects.

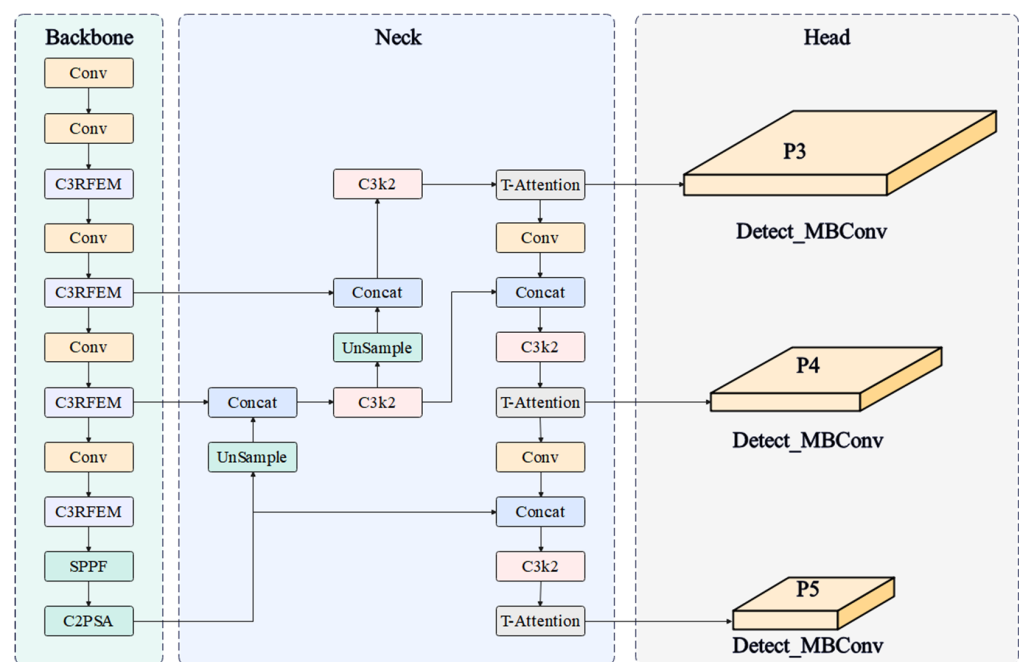


Figure 1. Overall network architecture of RTA-YOLOv11.

In the feature propagation path between the backbone and the neck, the triplet attention mechanism is combined to construct a cross-channel and vertical spatial information interaction pathway, thereby enhancing the model's ability to respond to low-contrast foreign object edges. This mechanism establishes lightweight attention coupling along the X, Y, and C axes, avoiding structural disturbances caused by single-channel enhancement while retaining key edge contours under background interference. The attention embedding position selection aligns with the semantic strength of the feature layer, covering both high-level semantic mapping in the backbone and reinforcing the guidance process of residual information for small objects in the top-down path, further optimizing the distribution perception of prominent foreign object regions.

To address the inference redundancy issue of the detection head in actual deployment, the original YOLOv11 Detect Head is modified to make it more lightweight using the *MBConv* and *SE* modules to construct dynamic compression detection channels. Combined with a reparameterizable design, the modified detection head maintains flexibility in channel selection and attention response during the training stage and achieves calculation

path merging through structural folding during the inference stage, effectively reducing inference latency. Additionally, the decoupled branches within the detection head adopt a space-channel hybrid strategy to increase the regression accuracy of long and narrow objects such as anchor rod contours while mitigating gradient imbalance caused by uneven channel allocation.

3.1. Convolutional Layer Optimization for Feature Extraction in the Backbone Network

Analysis of underground mining images faces three challenges: lighting, irregular occlusion, and wide-scale distribution. Small foreign objects are covered by only a small number of pixels in shallow convolutions, while deep convolutions lose edge textures because of stride accumulation. To address this contradiction, this paper replaces the traditional C3 bottleneck with the multibranch receptive field-enhanced convolution module *RFEM* in the *P3* and *P4* layers of the YOLOv11 backbone and in this way reconstructs the feature extraction path from three dimensions: the geometric field of view, computational complexity, and residual stability. The *RFEM* employs weight-sharing dilated convolutions to capture cross-scale dependencies while maintaining controllable parameter scale, thereby enhancing the discrimination capability for small objects, as shown in Figure 2.

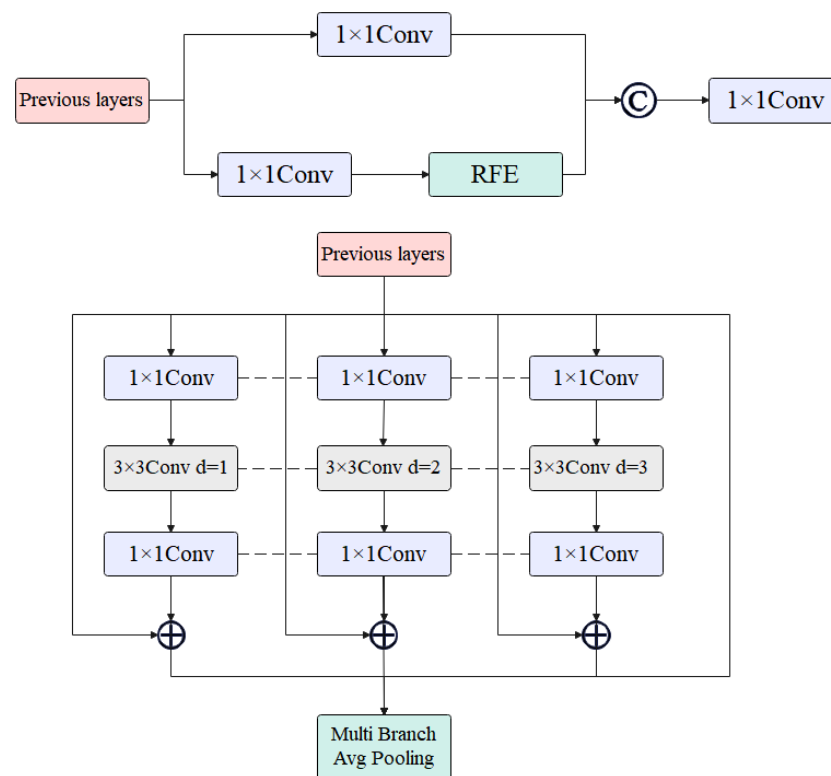


Figure 2. Structure diagram of the receptive field enhancement module.

The RFEM divides the input features into three shared weight expansion convolutional branches; the convolution kernel size, expansion rate set, and effective receptive diameter on a two-dimensional plane satisfy the following equation:

$$RF(d_i) = (k - 1)d_i + 1 \quad (1)$$

Among them,

$$k = 3, d_i \in \{1, 2, 3\} \quad (2)$$

Thus, a single convolution can provide a field of view of 3 pixels, 5 pixels, or 7 pixels, indicating a broadened shallow context. The output of a single branch convolution is uniformly written as follows:

$$U_i = W *_{d_i} F, i \in \{1, 2, 3\} \quad (3)$$

where W is a shared 3×3 kernel weight and $*_{d_i}$ denotes a convolution operation with an expansion rate of d_i . Each branch undergoes global average pooling— 1×1 convolution—*Softmax* to generate channel weights.

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^3 \exp(s_j)}, i = 1, 2, 3 \quad (4)$$

The greater the weight is, the more critical the scale is in the current input context. Soft gate fusion is then performed on an element-by-element basis.

$$U = \sum_{i=1}^3 \alpha_i \odot U_i \quad (5)$$

The fusion results are superimposed on the input features after batch normalization and *SiLU* activation mapping, forming an identity residual.

$$Y = \sigma(BN(U)) + F \quad (6)$$

Since the residual path depth is always 1, the gradient can pass through the back-propagation unchanged and is not dissipated by the deepening of the network. Owing to weight sharing, the total number of convolution parameters required for the three branches remains at

$$P_{RFEM} = k^2 CC_0 \quad (7)$$

The replaced C3 layer includes three independent convolutions of 3×3 , with approximately $3k^2 CC_0$. This means that *RFEM* can enable the high-resolution *P3* layer to achieve “fine-grained + wide background” dual perception without requiring a significant increase in memory or computational power while allowing the medium-resolution *P4* layer to capture global contours while retaining local edges. It thus provides a more balanced multiscale feature representation for subsequent attention mechanisms and detection heads.

The embedding method is implemented at the model configuration file level, where the C3 labels of nodes *P3* and *P4* are replaced with the custom *RFEM*; other anchor box sizes, loss weights, and training hyperparameters can remain unchanged. During inference, the sparse convolution kernel merges different dilation rate maps into a single node heterogeneous operation, with actual latency and memory overhead nearly equivalent to those of the original network. It effectively improves YOLOv11’s detection rate for small-sized targets while maintaining the integrity of large-sized object contours, demonstrating the effectiveness of convolution field reconstruction in underground multiscale foreign object detection tasks.

3.2. Attention Mechanism

Images of foreign objects on conveyor belts in coal mines contain complex noise sources such as floating dust, light streaks, and belt wrinkles. Relying solely on channel domain or single spatial domain attention weights often makes it difficult to respond to both row and column interference areas at the same time.

Triplet attention originates from the rotate-to-attend idea, as shown in Figure 3. Its core lies in mapping the “channel-row” and “channel-column” coupling relationships to

convolution-friendly tensor arrangements through two 90° view transformations, then using a set of shared 7×7 convolution kernels to extract three-dimensional context, and finally using sigmoid gating to reweight the input features element by element.

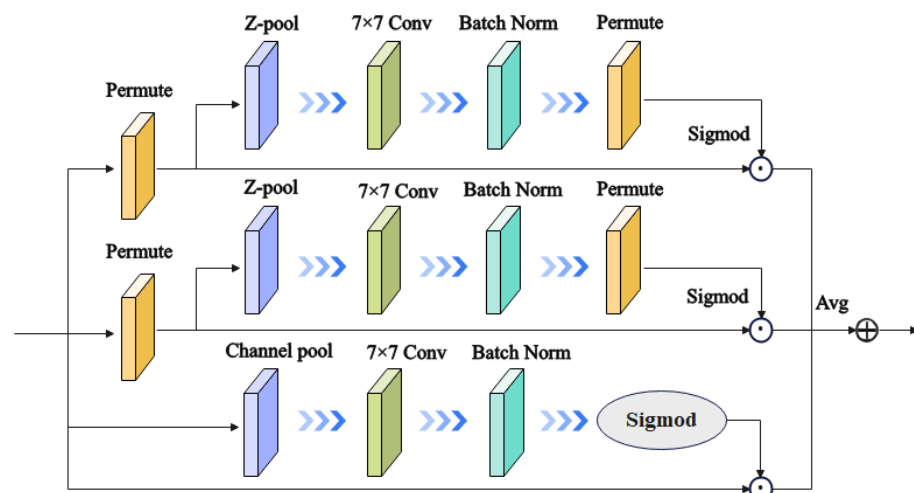


Figure 3. Triplet Attention mechanism structure.

On this basis, the overall computational procedure and network implementation framework are illustrated in Figure 4. Let the input feature be $X \in R^{C \times H \times W}$. First, a 90° view transformation is performed along the height and width axes in sequence:

$$\begin{aligned} X^{(1)} &= \text{Rotate}_H(X) \\ X^{(2)} &= \text{Rotate}_W(X) \\ X^{(3)} &= X \end{aligned} \quad (8)$$

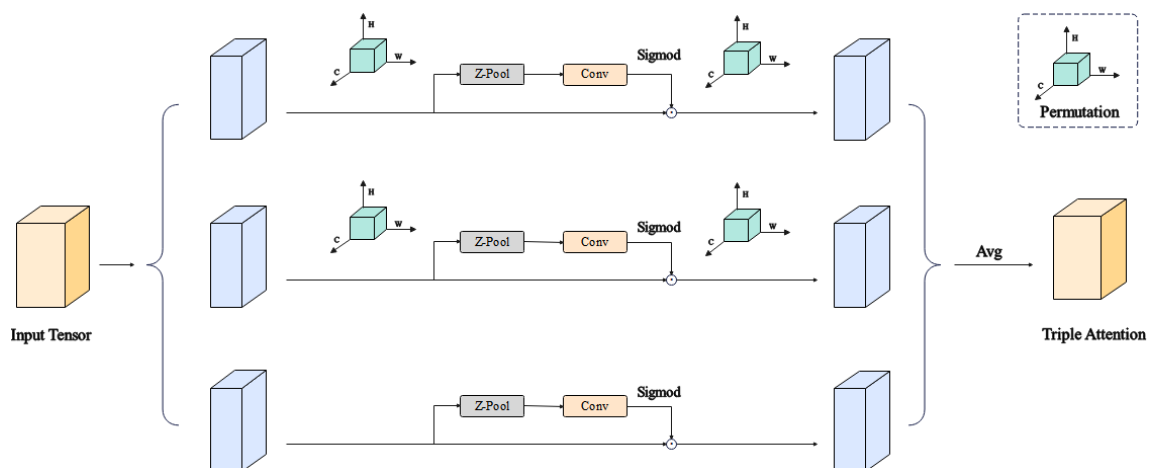


Figure 4. Rotation and gating process of the Triplet Attention mechanism.

Three cross-dimensional arrangements are obtained. Each arrangement first performs max pooling MP (kernel $k_p = 3$, stride 1) to increase the global perception domain and then uses a shared 7×7 convolution kernel $W_{7 \times 7}$ to extract the context.

$$Z^{(i)} = \sigma(W_{7 \times 7} * MP(X^{(i)})), \quad i = 1, 2, 3. \quad (9)$$

In Equation (9), $*$ denotes two-dimensional convolution. The rotation path output is reset by inverse rotation and added elementwise with the branch that retains the original axis arrangement to generate the attention mapping:

$$A = \text{Rotate}_H^{-1}(Z^{(1)}) + \text{Rotate}_W^{-1}(Z^{(2)}) + Z^{(3)} \quad (10)$$

The cross-dimensional weight map is obtained, and the input is then relabelled in a gated manner.

$$Y = X \odot A \quad (11)$$

This is elementwise multiplication. The overall process introduces the weight tensor only once, and its number of parameters is

$$P_{TA} = k^2 C, k = 7 \quad (12)$$

The complexity of the process is significantly reduced compared to attention schemes such as SE and CBAM that require two layers of fully connected or independent spatial convolutions. The multiplication and addition operations are performed in a corresponding manner.

$$FLOPs_{TA} = HWck^2 \quad (13)$$

At typical resolutions on layers $P3$, $P4$, and $P5$, it accounts for only approximately 1.8% of the total main convolution. Since both the forwards and inverse transforms are interpreted as tensor dimension views, no additional memory transfer is needed. Let the weight function for dust occlusion decay along the row direction be $\beta_h(y)$, and the weight function for light strip decay along the column direction be $\beta_w(x)$. The channel saliency modulation can then be approximated as

$$\tilde{S}(x, y) = S(x, y)[\alpha_c(x, y) + \beta_h(y) + \beta_w(x)] \quad (14)$$

In this way, the original foreign object significance distribution is derived. The above equation reflects the compensatory effect of three-dimensional coupled attention in complex underground light fields. Since the gated mapping uses an identity residual structure, the gradient propagation shortest path is

$$L_{grad} = \min(1, L) = 1 \quad (15)$$

Regardless of the total depth of the network, the attention weights converge stably in the early iteration stages. Triplet attention performs fine-grained cross-dimensional weighting on the convolutional output through the rotation–sharing–gating process defined by Equations (1)–(8), enhancing the contrast of dust-obscured areas and foreign object edges without significantly increasing parameters or latency, thereby improving the accuracy of foreign object detection on conveyor belts in coal mines.

3.3. Optimization Based on the Detection Head

In the actual deployment environment of foreign object detection on underground conveyor belts, inference devices are often power-constrained embedded GPUs or FPGAs. When the detection head becomes overloaded with computational tasks, a frame rate bottleneck can occur in the entire foreign object warning chain. To address this issue, this section replaces the original HeadBlock of YOLOv11, which consists of a three-layer standard convolution stack, with MBConv-Head, as shown in Figure 5.

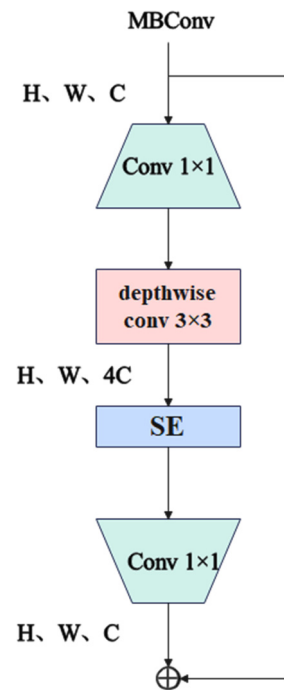


Figure 5. Structure of the lightweight MBConv detection head.

The improvement approach follows a five-step closed-loop process consisting of pointwise convolution with dimension expansion, depthwise separable convolution, an SE module, pointwise convolution with dimension reduction, and a shortcut connection. First, “pointwise convolution (1×1 convolution) for dimension expansion: a 1×1 pointwise convolution is applied to the input feature map to expand the number of channels” is performed; this can be specifically written as follows:

$$F_e = W_{1 \times 1}^\uparrow * F \quad (16)$$

$$C_e = tC_{in}, t = 6 \quad (17)$$

Channel expansion provides subsequent convolutions with a richer expressive base. Subsequently, “deep separable convolution” is performed; it includes deep convolution and pointwise convolution. Deep convolution extracts spatial information channel by channel, whereas pointwise convolution performs weighted fusion in the channel dimension. The use of a combination of the two significantly reduces the computational load and the parameter count.

In the detection head implementation, this can be merged into a single deep convolution:

$$F_d = W_{3 \times 3}^{dw} * F_e \quad (18)$$

Its riding amount is

$$MAC_{dw} = 9HWC_e \quad (19)$$

Compared with traditional convolution, the computational load is reduced by a factor of two, and the subsequent embedding of the squeeze-and-excitation (SE) module helps the model focus more on important feature information in the channel dimension, resulting in the following:

$$s = \sigma \left(W_{1 \times 1}^\downarrow \delta(GAP(F_d)) \right) \quad (20)$$

$$F_s = F_d \odot s \quad (21)$$

At this step, the spatial dimension is compressed using SiLU; this step prompts the SE module to automatically focus on more important channels by weighting channel features. Next, “pointwise convolution (1×1 convolution) dimension reduction” is performed: 1×1 pointwise convolution is used to restore the number of channels to a dimension close to the input, achieving feature fusion and compression, as follows:

$$F_o = W_{1 \times 1}^{\downarrow} * F_s \quad (22)$$

If the input shape matches, shortcut connection is enabled: when the input and output shapes are the same, they are added together to achieve feature reuse and alleviate gradient disappearance, i.e.,

$$Y = F_o + F \quad (23)$$

At this point, the complete workflow of “feature extraction: 1×1 -dimensional expansion channel $\rightarrow$ deep convolution to extract spatial features $\rightarrow$ pointwise convolution to fuse channel features; attention mechanism: SE compression–excitation; feature fusion and output: 1×1 -dimensional reduction + shortcut” is closed. The number of parameters in the entire MBConv-Head can be estimated using the inverted bottleneck formula

$$P_{AB} = 2C_{in}C_e + 9C_e + \frac{2C_e^2}{r}, r = 4 \quad (24)$$

A comparison of the weights of the three-layer convolutional HeadBlock shows that the reduction rate reaches approximately 70% at this point; the multiply–accumulate operation decreases synchronously according to Equation (3).

In summary, Detect-MBConv incorporates the MBConv structure, working principle, and advantages into the YOLOv11 detection head through a five-step process of “dimension elevation—depth separation—channel attention—dimension reduction—shortcut”. While significantly reducing the computational load and storage requirements, it maintains high accuracy, providing a real-time, lightweight, and highly reliable solution for detection of foreign objects on conveyor belts in coal mine shafts.

4. Experiments

4.1. Experimental Environment and Dataset

4.1.1. Experimental Environment

This experimental platform uses Ubuntu 22.04 as the operating system, with 60 GB of system memory, an RTX4090D graphics card (NVIDIA Corporation, Santa Clara, CA, USA), 24 GB of graphics memory, PyTorch framework version 2.3.0 + cu121, and Python version 3.12, as shown in Table 1. Owing to the limited size of the dataset, this paper employs data augmentation strategies on the CUMT-HelmeT dataset to expand the dataset size and prevent overfitting. These include adjustments to the brightness, cropping, translation, rotation, and overlay of the images.

Table 1. Configuration of the experimental environment.

Experimental Parameter	Configuration
CPU	12 vCPU Intel® Xeon® Platinum 8352 V CPU @ 2.10 GHz (Intel Corporation, Santa Clara, CA, USA)
GPU	RTX4090D
CUDA	12.1
Python	3.12 (Ubuntu 22.04)
PyTorch	2.3.0

4.1.2. Model Hyperparameter Settings

The training hyperparameters are configured as follows. The network is optimized using the AdamW optimizer with an initial learning rate of 0.001 and a weight decay coefficient of 0.0005. The first 5 epochs are used as a warm-up stage to stabilize network convergence. The momentum parameters of AdamW are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

All input images are resized to a resolution of 640×640 , and the training batch size is set to 32. The loss function consists of CIoU Loss for bounding box regression, BCE Loss for classification, and Distribution Focal Loss (DFL), with weighting factors of $\text{box} = 7.5$, $\text{cls} = 0.5$, and $\text{dfl} = 1.5$, respectively, to balance localization accuracy and classification stability. During training, the IoU threshold is set to 0.5, the IoU threshold for non-maximum suppression (NMS) is 0.6, and the confidence threshold is 0.25. Network parameters are initialized using the Kaiming Normal initialization method.

To ensure the fairness of the comparative experiments, the algorithm models selected in this paper were trained and tested synchronously under the aforementioned consistent environmental configurations and hyperparameter strategies, thereby eliminating interference from external factors so that the experimental results effectively reflect the performance differences of various algorithms in the underground coal mine production environment.

4.1.3. Dataset

The dataset used in this paper is the CUMT-BelT public dataset [37] for foreign object detection on underground conveyor belts in coal mines, released by the China University of Mining and Technology. The images were collected from actual underground conveyor belt operations in coal mines and cover typical foreign object types such as gangue and anchor bolts, as shown in Figure 6. To construct the training samples, the original images were manually annotated using the LabelMe tool (version 5.1), resulting in approximately 5200 images; the images were then divided into training, validation, and test sets at a 7:2:1 ratio.

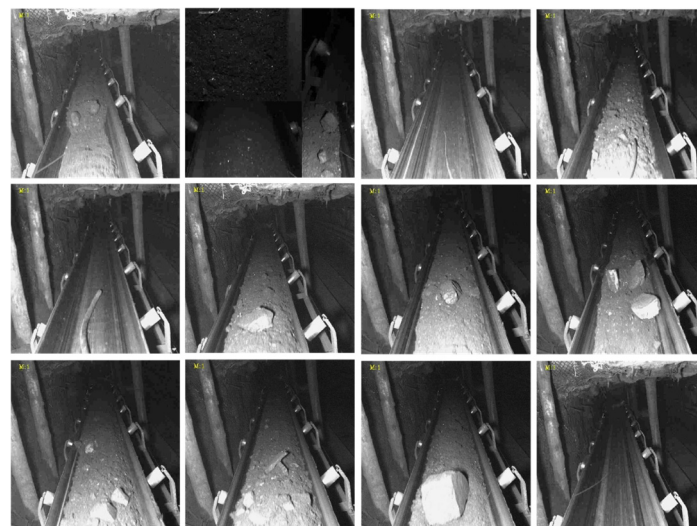


Figure 6. Sample images from the underground coal mine dataset.

4.1.4. Dataset Preprocessing

The original image dataset of CUMT-BelT underwent systematic preprocessing. Since the original dataset contains images of underground scenes with low light, severe occlusion, and complex backgrounds, the target objects have significant size differences, and small targets represent a high proportion of the targets, use of the original samples directly

could easily lead to model overfitting or to unstable target detection results due to the complex environment.

To enhance the model's robustness and generalization while effectively evaluating the training results, the original data require augmentation. Using tools such as OpenCV (version 4.12.0), the images were screened for completeness and uniformly resized to 1280×720 . The aspect ratio was maintained using proportional scaling and black border padding. Referencing existing methods, the images were then subjected to enhancement strategies, including horizontal flipping, rotation within $\pm 10^\circ$, scaling and cropping, HSV color perturbation, and addition of Gaussian noise. The execution probability for each operation was set to 0.5. For samples in which the image dimensions changed during the enhancement process, the bounding box coordinates in the label files were synchronously remapped and normalized. Through this preprocessing workflow, the sample size of the dataset was further expanded, providing effective support for the model's object detection performance in complex scenarios.

4.2. Evaluation Indicators

In this paper, precision (P), recall (R), and mean average precision (mAP) are used as metrics for evaluation of the model. $mAP@0.5$ is calculated by setting the intersection over union (IoU) between the predicted bounding box and the ground truth bounding box to 0.5 and then computing the average AP for each category. P and R are both calculated at an IoU threshold of 0.50.

Precision (P) is a metric that measures the proportion of predicted bounding boxes classified as positive that are actually true targets. Its formula is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (25)$$

where TP refers to the number of samples predicted as positive that are actually positive, and FP refers to the number of samples that are predicted as positive but are actually negative. The higher the accuracy rate is, the lower the false-positive rate is, indicating that the model has stronger discrimination capabilities.

Recall that R reflects the proportion of true targets that the model can correctly detect and reflects the degree of false negatives. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \quad (26)$$

Among these variables, FN denotes the number of samples that are actually positive but are not detected. A higher recall rate indicates that the model has stronger coverage in target detection.

The mean average precision (mAP) is among the comprehensive evaluation metrics that are most commonly used in object detection tasks; it is used to measure the overlap between predicted boxes and ground-truth boxes. The $mAP@0.5$ metric used in this paper refers to the average of the average precision (AP) for all categories under an intersection over union (IoU) threshold of 0.5. It is calculated as follows:

$$AP_{50}^{(i)} = \int_0^1 p_i(r) dr \quad (27)$$

$$mAP_{50} = \frac{1}{N} \sum_{i=1}^N AP_{50}^{(i)} \quad (28)$$

The accuracy curve of the class at different recall rates represents the total number of classes and the average accuracy of the class under the condition. The higher this value

is, the stronger the model's target localization capability is and the better its detection performance is under that condition.

The number of model parameters (Parameters) is an important metric for measuring network complexity and storage overhead. It represents the total number of trainable parameters (such as convolution kernels and fully connected layer weights) in the model; these typically number millions (M) or tens of millions (10^7). A smaller number of parameters indicates a lighter-weight model that is easier to deploy on edge devices, but an excessively small number may compromise the model's expressive capability.

Inference speed (frames per second, FPS), a key metric for evaluating real-time model performance, indicates the number of image frames the model can process per second. FPS is influenced by factors such as model complexity, input resolution, and the computing platform. To meet real-time requirements in practical deployment, a trade-off must be made between detection accuracy and inference speed.

4.3. Ablation Experiment

The specific role of each individual structural module of the model in performance improvement is often difficult to quantify. Therefore, it is necessary to conduct a detailed breakdown and analysis of the combined methods item by item. On this basis, eight configuration combinations are constructed around the receptive field enhancement module RFEM, the triplet attention module Triplet Attention, and the lightweight detection head MBConv in this paper. In this way, the impact of the individual introduction and collaborative superposition of each module on the overall detection performance is systematically evaluated. The results are summarized in the following Table 2.

Table 2. Ablation experiment table.

Group	RFEM	T-Attention	Detect_MBConv	mAP@0.5	P	R	Parameters (M)	FPS
1				0.860	0.845	0.800	2.55	60
2	✓			0.875	0.855	0.812	2.31	63
3		✓		0.872	0.852	0.808	2.59	63
4			✓	0.869	0.848	0.806	2.26	67
5	✓	✓		0.890	0.860	0.818	2.34	70
6		✓	✓	0.885	0.858	0.816	2.56	76
7	✓		✓	0.888	0.859	0.817	1.98	73
8	✓	✓	✓	0.902	0.867	0.821	1.99	79

Note: "✓" indicates the inclusion of the corresponding module.

As shown in Table 2, to verify the actual effectiveness of each module in improving overall performance, a total of eight ablation experiments focusing on the receptive field enhancement module RFEM, the triplet attention mechanism Triplet Attention, and the lightweight detection head MBConv were performed. The experimental results are shown in Table 2.

In Group 1, no modules were incorporated into the original YOLOv11 model. The model achieved an mAP@0.5 of 0.860, an accuracy of 0.845, a recall rate of 0.800, a parameter count of 2.55 million, and an inference speed of 60 frames per second. In Group 2, in which the RFEM module was introduced, the mAP improved to 0.875, indicating that expanding the receptive field effectively enhances the model's ability to represent multiscale objects. In Group 3, in which triplet attention was introduced, the mAP reached 0.872, indicating that the cross-dimensional attention mechanism has a positive effect on the extraction of significant features. In Group 4, in which MBConv was used to improve the detection head, the model parameters were reduced to 2.26 million, the FPS was

improved to 67, and the mAP reached 0.869, demonstrating good advantages in terms of lightwightness and efficiency.

In Groups 5 to 7, which combine the two modules, the performance is further improved. In Group 5, in which *RFEM* and triplet attention are jointly introduced, the mAP improves to 0.890. In Group 6, in which triplet attention and MBConv are introduced, the model achieves an mAP of 0.885 while maintaining a high frame rate. In Group 7, the combination of *RFEM* and MBConv was adopted, and an mAP of 0.888 was achieved while compressing the parameters to a minimum of 1.98 M and increasing the frame rate to 73, thus balancing accuracy and speed.

Group 8 is a full-module fusion scheme that comprehensively introduces *RFEM*, triplet attention, and MBConv. It results in the best overall performance, with $mAP@0.5$ s reaching 0.902, accuracy and recall rates of 0.867 and 0.821, respectively, a parameter count of just 1.99 M, and an inference speed of 79 frames per second, fully validating the effectiveness of the three-module collaborative design.

Figure 7 shows the trend of the mean average precision ($mAP@0.5$) as the number of iterations increases during training for the original YOLOv11 and the improved RTA-YOLOv11 models. The models are used to evaluate the impact of the improvement strategy on the convergence speed and on the upper limit of detection performance.

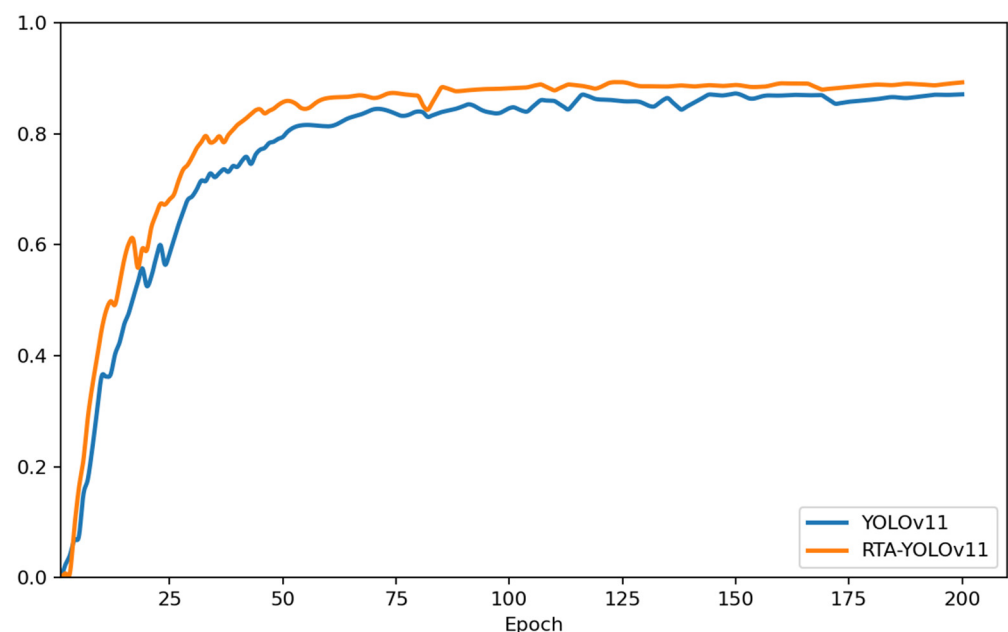


Figure 7. Accuracy comparison between the improved and baseline models.

The curve trends indicate that the improved RTA-YOLOv11 model converges significantly faster than the original model does in the early stages. It surpasses 0.5 mAP by the 15th iteration, far outperforming YOLOv11, which reaches this threshold around the 25th iteration. These findings indicate that the synergistic effects of the introduced receptive field enhancement, attention mechanism, and lightweight detection head in feature extraction and fusion significantly enhance the early discriminative ability of the model. As the number of training iterations increases, both models stabilize, but compared with YOLOv11, RTA-YOLOv11 maintains a higher detection accuracy, reaching a peak mAP of approximately 0.88 by the 100th iteration, an improvement of approximately 0.05. This further validates the effectiveness of the proposed improvements in enhancing convergence efficiency and detection performance.

To further validate the improvement in the model's actual detection performance through the enhanced strategy, typical test sample images were selected, and visualization detection by the original YOLOv11n model was compared with visualization detection by the improved RTA-YOLOv11 model. The results are shown in Figure 8.

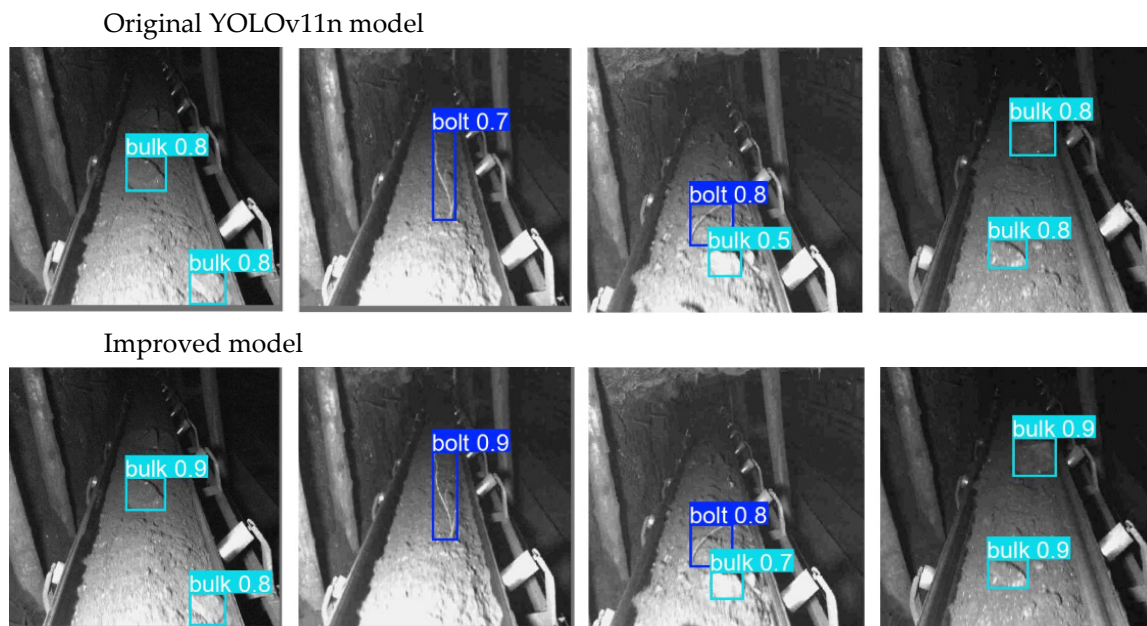


Figure 8. Prediction performance comparison between the improved and baseline models.

As shown in Figure 8, the original model exhibits issues such as low confidence scores and bounding box shifts when small targets such as bolts and bulk objects are detected, resulting in incomplete or inaccurate target detection for some objects. In contrast, the improved model not only enhances target confidence scores but also produces more compact and accurate bounding boxes, demonstrating superior feature representation and target discrimination capabilities.

Additionally, in multiobject scenarios, the improved model significantly enhances the recognition capability for parallel objects, avoiding the missed detection of bulk objects that occurs when the original model is used. This further validates the effectiveness of the fusion of the enhanced receptive field module, the triplet attention mechanism, and the lightweight detection head in improving foreign object detection performance in complex underground scenarios.

4.4. Comparative Experiments

4.4.1. Lateral Comparison of Convolutional Layers

To verify the actual performance of the proposed RFEM in the backbone network and comprehensively evaluate the effects of different convolutional structures on the detection accuracy, model size and inference speed, the current mainstream lightweight and multiscale convolution modules (such as C3k2 DynamicConv, APConv, DBB, and SPDConv) were tested in lateral comparison experiments with the proposed RFEM. All improvements were made by replacing the standard convolutional structure in the YOLOv11 backbone network in the same way and maintaining a consistent training configuration on the CUMT-Belt dataset to ensure comparability of the results. The specific results of the comparisons are shown in Table 3 below.

Table 3. Lateral ablation experiment: comparison of convolutional layers.

Improvement Module	mAP@0.5	Precision	Recall	Number of Parameters (M)	FPS
C3k2 DynamicConv	0.804	0.845	0.715	3.00	58
APConv	0.837	0.846	0.745	2.53	61
DBB	0.812	0.808	0.719	3.22	54
SPDConv	0.823	0.833	0.726	2.22	67
RFEM	0.840	0.839	0.766	2.31	69

Table 3 presents the results of a side-by-side comparison that was performed after introducing different lightweight/multiscale convolution structures into the main convolution layer of YOLOv11. As shown in the table, both C3k2 DynamicConv and DBB achieve some gains in terms of receptive field expansion. However, owing to the computational overhead caused by the parameter scale and the convolutional reparameterization, their mAP@0.5s are 0.804 and 0.812, respectively, and their inference speeds are only 58 FPS and 54 FPS, respectively, making it difficult to meet real-time requirements. SPDConv effectively reduces the number of parameters (2.22 M) through spatial rearrangement strategies, increasing FPS to 67, but its detection accuracy remains lower than the 0.840 achieved by RFEM.

APConv, which relies on a separable and reusable convolution kernel design, improves the mAP to 0.837 under 2.53 million parameters, but its precision and recall metrics still lag slightly behind those of RFEM. In contrast, the proposed RFEM introduces cross-scale gated aggregation and deep separable residual structures in the enhanced receptive field module, maintaining a lightweight scale of 2.31 million parameters while achieving the highest detection accuracy (mAP@0.5: 0.840; precision: 0.839; recall: 0.766) while achieving an FPS of 69, thereby achieving a better balance between detection accuracy and real-time performance.

In summary, RFEM achieves a more balanced trade-off between detection accuracy, recall rate, and real-time inference speed, providing a more competitive convolutional layer improvement scheme for future small object detection tasks.

4.4.2. Horizontal Comparison of Attention Mechanisms

There are significant differences in the capabilities of different attention mechanisms with respect to information screening and feature enhancement, and their methods of introduction and their structural complexity also have different effects on model performance. Considering the sensitivity of foreign object detection tasks to multiscale detailed features, in this paper, experiments are conducted around three typical attention modules, and a lateral analysis is conducted by combining dimensions such as accuracy, recall rate, reasoning speed and parameter quantity. The experimental results are shown in Table 4.

Table 4. Horizontal ablation experiment: comparison of attention mechanisms.

Improved Attention Module	mAP@0.5	Precision	Recall	Number of Parameters (M)	FPS
SimAM	0.825	0.819	0.724	2.58	64
MCA	0.834	0.838	0.731	2.58	66
Triplet Attention	0.837	0.847	0.757	2.58	68

Table 4 compares the performance of different attention mechanism modules embedded in the YOLOv11 backbone network. As shown in Table 4, SimAM, as a single-channel attention mechanism with extremely few parameters, achieves only 0.825 mAP@0.5 and 0.724 recall while maintaining 2.58 million parameters and high inference efficiency (64 FPS), demonstrating relatively limited performance. Its main advantage lies in its simple struc-

ture and its minimal impact on inference speed; however, it faces limitations in terms of representational capability and receptive field adaptability. MCA enhances feature selection capabilities by introducing multichannel interaction and attention reallocation strategies without significantly increasing model complexity; it improves mAP to 0.834 and precision to 0.838, a notable improvement over SimAM, while maintaining FPS at 66 and balancing accuracy and efficiency. Triplet attention further introduces a three-dimensional joint modeling mechanism for spatial, channel, and directional dimensions, enhancing the expression of multidimensional feature dependencies. The experimental results show that this module achieves the highest values on all three metrics (mAP, precision, and recall) at 0.837, 0.847, and 0.757, respectively. Moreover, while maintaining the same number of parameters, the inference speed is improved to 68 FPS, demonstrating superior overall performance.

4.4.3. Transverse Comparison of Detection Heads

As a key component of the output stage of the target detection network, the structural design of the detection head directly affects the model's ability to distinguish the position and category of the target. In response to the potential accuracy and speed bottlenecks that processing at this stage may cause, this paper selects several common improvement schemes for detection heads at present and conducts a comparative analysis of their performances under various performance indicators. The test results are summarized in the following Table 5.

Table 5. Lateral ablation experiment: comparison of detection heads.

Improvement Plan for Detection Head	mAP@0.5	Precision	Recall	Number of Parameters (M)	FPS
Detect_AFPN3	0.824	0.818	0.692	4.21	62
Detect_Conv2Formers	0.797	0.869	0.694	2.39	68
Detect_PPA	0.847	0.819	0.744	3.20	71
Detect_MBCConv	0.829	0.889	0.685	2.26	78

Table 5 shows a side-by-side comparison of the results obtained using the four detection head improvement schemes. Among the metrics, Detect_PPA achieves mAP and recall values of 0.847 and 0.744, respectively. Although it has high accuracy, its parameter count reaches 3.20 M, resulting in high model complexity and making it difficult to adapt to resource-constrained scenarios. Detect_AFPN3 leverages a feature pyramid to enhance feature fusion capabilities, achieving an mAP of 0.824 and a precision of 0.818. However, it has a parameter count of 4.21 M and an inference speed of only 62 FPS, resulting in significant deployment costs. Detect_Conv2Formers introduces a transformer encoder that enhances the global receptive field; it has 2.39 million parameters and a fast speed of 68 FPS, but the mAP is only 0.797, limiting detection performance. In the comparison, Detect_MBCConv stands out. This architecture employs efficient depth-separable convolutions, with only 2.26 million parameters and an inference speed of 78 FPS. It achieves a high detection accuracy of 0.829 mAP while maintaining lightweight characteristics.

A comprehensive analysis of all the metrics reveals that Detect_MBCConv uses the smallest number of parameters, has the highest inference speed, and exhibits excellent detection performance, demonstrating the best overall performance among the tested solutions.

4.4.4. Comparison of Different Models

In the context of the continuous evolution of various detection algorithms, individual models exhibit distinct trade-off strategies between detection accuracy and deployment efficiency. To visually compare the adaptability of mainstream detection models using the

dataset employed in this study, this paper statistically analyses the core performance of several representative models, including the two-stage method, SSD series, and various versions of YOLO, in terms of mAP@0.5, parameter count, and inference speed (FPS). The detailed results are provided in Table 6 below.

Table 6. Multi-index comparative experiment.

Method	AP (%)		mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params (M)	FPS
	Bolt	Bulk				
Faster R-CNN	64.2	91.0	77.5	42.6	42.10	16.9
SSD	67.8	90.2	79.5	44.5	35.64	48.5
YOLOv7-tiny	72.1	91.3	81.2	46.7	6.03	57.4
YOLOv8	78.0	91.6	84.8	50.2	3.20	56.6
YOLOv10	79.5	91.7	85.6	51.0	2.72	65.0
YOLOv11	80.1	91.5	85.9	51.6	2.59	68.3
RTA-YOLOv11	86.6	92.1	89.3	54.8	1.99	76.2

Table 6 below presents a comparison of the performance of current mainstream object detection models on this dataset, evaluating metrics including mAP@0.5, parameter count, and inference speed (FPS) so as to comprehensively assess each model's ability to balance accuracy and deployment efficiency. Despite having a certain detection accuracy (77.5%), the traditional two-stage detector Faster R-CNN has a parameter count as high as 42.10 M and an FPS of only 16.93, making it difficult to meet real-time detection requirements. SSD improves speed to some extent (48.53 FPS) but still has relatively low detection accuracy (79.5%) and limited adaptability.

YOLOv7-tiny, an early lightweight version, compresses the model to 6.03 M, achieves 57.42 FPS, and improves the mAP to 81.2%. With structural iterations, YOLOv8, YOLOv10, and YOLOv11 achieve continuous improvements in accuracy to 84.8%, 85.6%, and 85.9%, respectively, with parameters optimized to 3.20 M, 2.72 M, and 2.59 M, respectively, and FPS exceeding 50, demonstrating stable and efficient performance.

The improved RTA-YOLOv11 model proposed in this paper achieves a better balance between detection accuracy and deployment efficiency, with mAP@0.5 and mAP@0.5:0.95 reaching 89.3% and 54.8%, respectively, the highest among all models. In terms of small object performance evaluation, the AP for the bolt category reaches 86.6%, which is a notable improvement over the baseline YOLOv11n. This growth reflects the perceptual advantage of the RFEM for small-sized and weak-feature targets, as well as the stability of the Triplet attention mechanism in extracting key contour features under complex backgrounds. Meanwhile, the model parameters are compressed to 1.99 M and the FPS is increased to 76.2, fully demonstrating the high level of unity between accuracy and lightweight design. This model maintains leading detection accuracy while reducing computational resource overhead, making it more suitable for real-time deployment in resource-constrained scenarios such as coal mines.

Figure 9 shows the visualization results of foreign object detection in typical coal mine scenarios using various earlier models and the RTA-YOLOv11 model proposed in this paper. A comparative analysis of three aspects of the detection (accuracy, stability, and robustness) was conducted. Faster R-CNN suffers from issues such as low confidence scores and misaligned bounding boxes, resulting in failure to effectively identify some foreign objects. This is particularly evident for images with low contrast or complex backgrounds; in these cases, false positives and false negatives are prominent, indicating weak generalization capabilities. YOLOv8 shows significant improvements over traditional two-stage models, achieving stable detection for some targets. However, it still struggles

with distant targets and small-sized foreign objects, resulting in some information loss and ambiguity in classification.

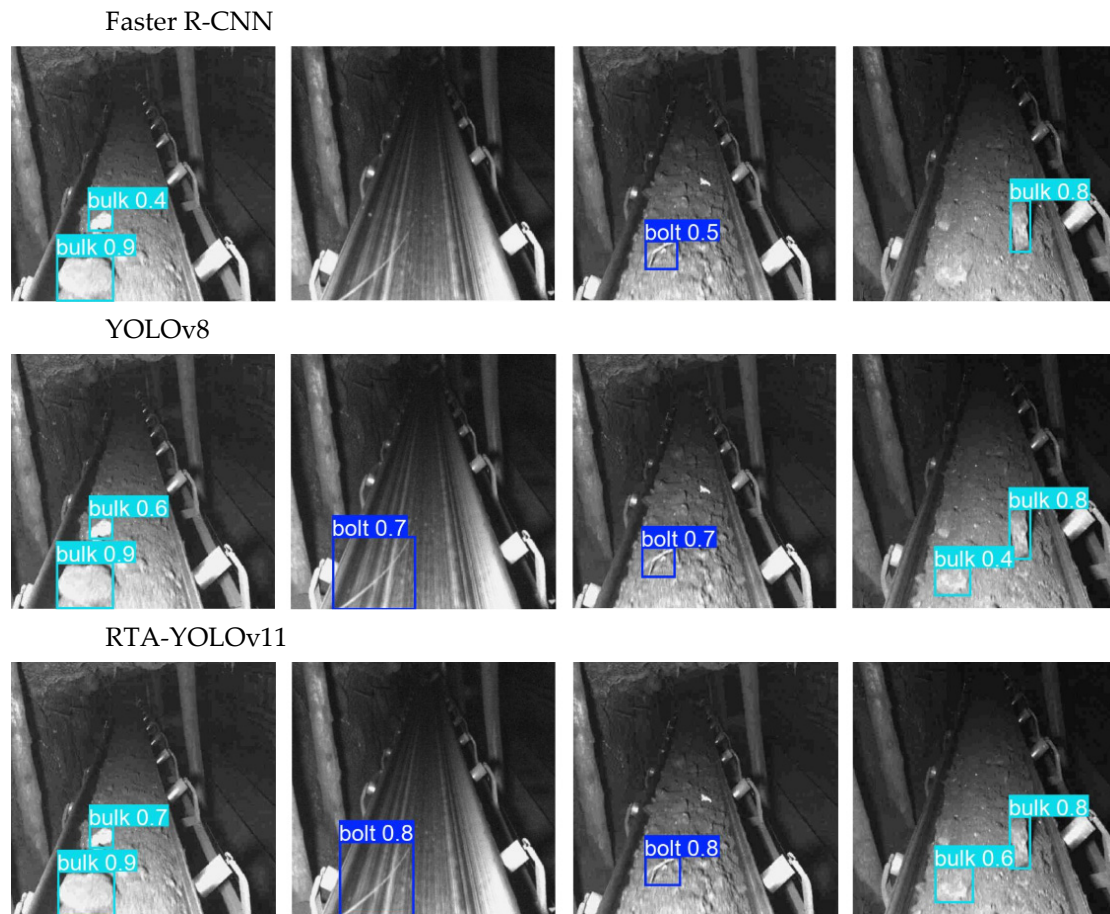


Figure 9. Comparison of prediction results between the improved model and other models.

RTA-YOLOv11 demonstrates more precise detection performance in the same scenario, successfully identifying all foreign objects and producing bounding boxes closely aligned with the actual boundaries. Its confidence scores are balanced and significantly higher than those of the other models, indicating excellent robustness and detail perception capabilities. RTA-YOLOv11 combines accuracy of detection with lightweight characteristics, enabling high-quality identification of foreign objects on conveyor belts in complex underground environments and demonstrating significant application potential.

Figure 10 shows the changes in the mean average precision (mAP@0.5) during the training process for each detection model. Faster R-CNN and SSD converge slowly in the early stages, with limited improvements in mAP, indicating a clear underfitting trend. YOLOv7-tiny exhibits a rapid upwards trend in its curve and a faster convergence rate, but the overall stability is average, with some fluctuations. YOLOv8 and YOLOv9 have smoother convergence processes, but YOLOv9 experiences oscillations after achieving high accuracy in the early stages, resulting in relatively poorer stability. YOLOv10 performs exceptionally well, maintaining high-accuracy output in the mid-to-late stages and exhibiting a stable and reliable training process. The improved RTA-YOLOv11 demonstrates superior convergence efficiency and higher detection accuracy across all stages, with a stable upwards trend in the overall curve. In the later stages, its accuracy improves further, and the mAP reaches its peak, indicating stronger object detection capabilities and better generalization performance. This model balances improved accuracy with controlled

parameter complexity, offering greater adaptability and maintaining a leading advantage throughout multiple training iterations.

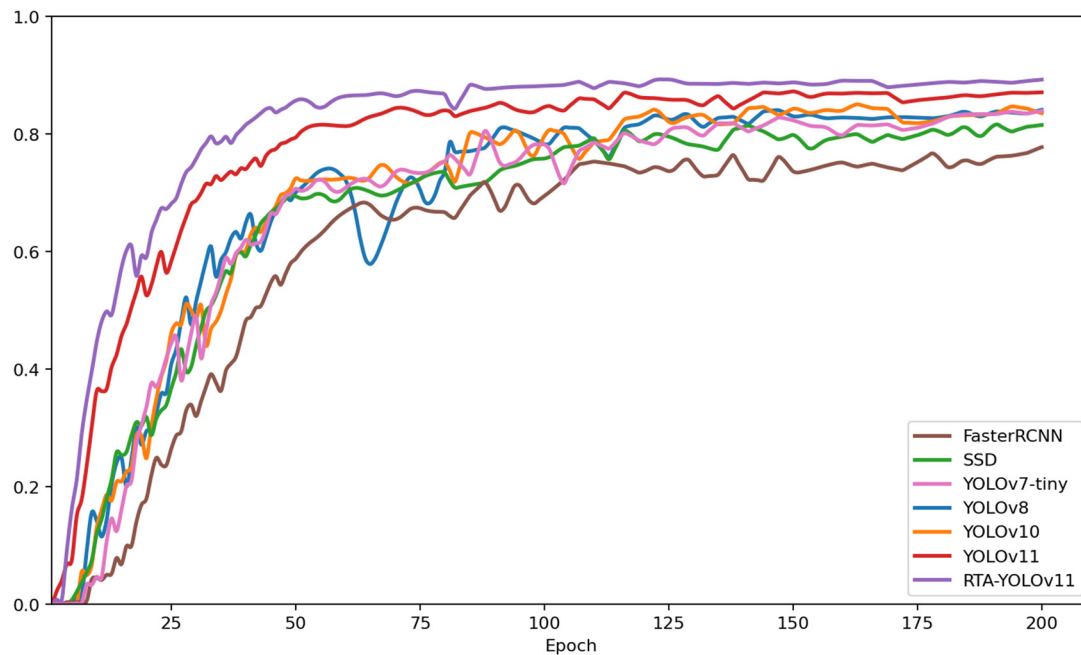


Figure 10. Accuracy comparison between the improved model and other models.

4.4.5. Generalization Verification

To further verify the generalization capability of RTA-YOLOv11 in different mine scenarios, the open-source dataset DsLMF+ [38] for fully mechanized coal mining faces was introduced for supplementary experiments. In this paper, the large_coal category in this dataset is taken as the anomaly detection target to verify the model's target recognition capability across scenarios.

Table 7 presents the performance of the algorithm on the DsLMF+ dataset. This dataset was collected from fully mechanized mining faces with complex backgrounds, where the texture features of large coal targets are similar to those of the background coal flow; as a result, the metrics of each model exhibit domain shift phenomena as the environment changes. Specifically, the mAP@0.5 increased from 78.6% for the YOLOv11n baseline model to 81.8% for the improved RTA-YOLOv11 model, an increase of 4.1%. This growth magnitude is basically consistent with the 4.0% improvement achieved by the model on the original dataset, indicating that the model's improvement strategies still provide stable contributions under various environments, such as different mining areas and geological conditions, reflecting the generalization capability of the algorithm.

Table 7. Performance comparison of models across datasets.

Method	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params (M)	FPS
YOLOv11	78.6	48.3	2.59	66.4
RTA-YOLOv11	81.8	51.6	1.99	74.1

4.5. Limitations and Failure Analysis

RTA-YOLOv11 displays stable detection performance under various operating conditions, but deviations still exist within its performance boundaries in specific physical environments. These deviations typically stem from the deep coupling of the physical features of the target and background noise. In areas with high dust concentrations and

poor lighting conditions, the surfaces of foreign objects are often covered with coal dust, causing their texture and color features to converge with those of the coal flow on the conveyor belt. This makes it difficult for the model to distinguish target edges during feature extraction; for instance, feature loss is more pronounced when the target size is smaller than 32 pixels or when the target is located on a critical downsampling path. Such a low signal-to-noise-ratio (SNR) environment not only increases the probability of false negatives but also causes the model to potentially identify coal blocks or shadow areas with similar textures as abnormal targets, resulting in detection bias.

Furthermore, the high-speed operation of the conveyor system poses a challenge to detection stability. The motion blur effect may cause changes in the geometric contours of foreign objects, or cause belt damage and abnormal coal accumulation to exhibit features similar to those of foreign objects, thereby triggering false positives. Dynamic shadows generated by light refraction are intertwined with these motion features, further increasing the complexity of model discrimination. These phenomena indicate that there is still room for improvement in the model's anti-interference capability under complex dynamic backgrounds.

5. Conclusions and Outlook

5.1. Main Research Findings

This paper addresses the challenges of modeling small objects, significant low-light interference, and limited deployment efficiency in the detection of foreign objects on underground conveyor belts. On the basis of the YOLOv11 network, an improved model, RTA-YOLOv11, was developed. Specifically, the model introduces a multibranch receptive field enhancement module (a residual feature enhancement module, RFEM) in the P3 and P4 layers of the backbone network, incorporates a triplet attention mechanism in the backbone and neck sections, and includes a lightweight fusion module in the detection head structure. RFEM combines dilated convolutions with multiscale pooling structures to expand the receptive field, effectively enhancing the response to shallow-layer features of small foreign objects. Triplet attention employs a cross-channel-spatial interaction modeling approach to guide the model to focus on key features in low-contrast regions, enhancing target discrimination capabilities under dust interference and cluttered background conditions. Detect_MBCConv combines MBCConv, SE channel attention, and a structurally reparameterizable design to achieve efficient inference path compression under end-to-end conditions, balancing model accuracy and deployment speed. Through ablation experiments and comparative experiments, it has been verified that each of the innovative modules is effective and that they exhibit significant synergistic effects. The improved RTA-YOLOv11 outperforms the original YOLOv11n, SSD, and Faster R-CNN models in terms of metrics such as mAP@0.5 and the recall rate. The improved model also has high computational efficiency, making it suitable for real-time detection scenarios in underground environments, providing a reliable solution for practical applications and contributing to enhanced safety and production efficiency in underground operations.

5.2. Prospects for Future Work

Although this study yielded significant results, there is still room for optimization, and future work can be conducted from multiple angles. In terms of data, while the current dataset covers various underground scenarios, its diversity needs to be improved. To expand the dataset, future efforts can focus on collecting data from different coal mines where there are a range of geological conditions, lighting intensities, and types of foreign objects. Combining data augmentation techniques such as rotation and flipping can enrich the training data, enhance the model's generalization ability, and provide better adaptation

to complex real-world scenarios. In terms of model structure, while the improved YOLOv11 achieves significant gains in performance, further exploration of more efficient attention mechanisms and feature fusion methods is warranted to enhance operational efficiency, achieve precise foreign object identification, and improve detection accuracy and reliability.

Author Contributions: Conceptualization, L.W.; Methodology, L.W. and K.H.; Software, L.W. and K.H.; Validation, K.H.; Formal analysis, L.W. and K.H.; Investigation, L.W., K.H. and X.S.; Resources, L.W. and X.S.; Data curation, X.S. and J.C.; Writing—original draft, K.H. and J.C.; Writing—review & editing, L.W., K.H., X.S. and J.C.; Visualization, K.H. and J.C.; Supervision, L.W., X.S. and J.C.; Project administration, X.S. and J.C.; Funding acquisition, L.W. and X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

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

References

- Cheng, D.; Qian, J.; Guo, X.; Kou, Q.; Xu, F.; Gu, J.; Gao, Y.; Zhao, J. Review on key technologies of AI recognition for videos in coal mine. *Coal Sci. Technol.* **2023**, *51*, 349–365.
- Gao, H.; Zhao, P.; Yu, Z.; Xiao, T.; Li, X.; Li, L. Coal mine conveyor belt foreign object detection based on feature enhancement and transformer. *Coal Sci. Technol.* **2024**, *52*, 199–208.
- Fan, H.; Liu, J.; Cao, X.; Zhang, C.; Zhang, X.; Li, M.; Ma, H.; Mao, Q. Multi-scale target intelligent detection method for coal, foreign object and early damage of conveyor belt surface under low illumination and dust fog. *J. China Coal Soc.* **2024**, *49*, 1259–1270.
- Wang, D.; Wang, G.; Cao, R.; Wang, Z.; Lyu, H.; Tong, R. Interaction mechanism and model of coal mine safety management behaviors. *China Saf. Sci. J.* **2025**, *35*, 32.
- Hou, X. Design of Vision-Based Belt Conveyor Misalignment and Coal Spillage Fault Diagnosis System. *Coal Mine Mach.* **2024**, *45*, 164–166. [[CrossRef](#)]
- Han, Y. Research on Small-Scale Low-Contrast Conveyor Belt Damage Detection Algorithm Based on Single-Stage Network. Master's Thesis, Taiyuan University of Technology, Taiyuan, China, 2023. [[CrossRef](#)]
- Wang, W.; Li, K.; Zhao, S.; Wang, Y.; Lu, Z.; Li, L.; Guo, S. Research on Registration Method of Coal Flow Profile Point Cloud of Scraper Conveyor in Synthesized Mining Face. *Mod. Electron. Technol.* **2025**, *48*, 81–87.
- Bai, Z. Intelligent Centralized Control System for Main Coal Flow Belt Conveyor in Coal Mines. *Energy Energy Conserv.* **2025**, 185–187. [[CrossRef](#)]
- Cheng, Y.; Bai, S.; Dang, W.; Pan, L.; Wu, Z. Underload Detection of Underground Coal Mine Conveyor Belt. *Comput. Syst. Appl.* **2021**, *30*, 171–176. [[CrossRef](#)]
- Wang, Y.; Liu, J.; He, H.; Zhao, S.; Lu, Z.; Xu, L. Research on the Control Method of Coal Flow Congestion Prevention in Fully Mechanized Mining Face Based on Integrated Decision Making. *Coal Sci. Technol.* **2025**, 1–12.
- Han, X.; Yunlong, S. Research on AI-Based Foreign Object Recognition and Linkage Control of Underground Conveyor Belts. *Energy Energy Conserv.* **2023**, 154–158. [[CrossRef](#)]
- Cheng, D.Q.; Chen, J.; Kou, Q.Q.; Nie, S.; Zhang, J. Lightweight Super-Resolution Reconstruction Method Based on Hierarchical Features Fusion and Attention Mechanism for Mine Image. *Chin. J. Sci. Instrum.* **2022**, *43*, 73–84. [[CrossRef](#)]
- Lin, D. Design of Foreign Object Detection System for Coal Mine Belt Conveyors. *Autom. Appl.* **2025**, *66*, 138–141. [[CrossRef](#)]
- Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. Yolo-facev2: A scale and occlusion aware face detector. *Pattern Recognit.* **2024**, *155*, 110714. [[CrossRef](#)]
- Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; IEEE: Los Alamitos, CA, USA, 2021.
- Tan, M.; Quoc, L. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2019.
- Otsu, N. A threshold selection method from gray-level histograms. *Automatica* **1975**, *11*, 23–27. [[CrossRef](#)]

18. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *6*, 679–698.
19. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **2007**, *6*, 610–621. [[CrossRef](#)]
20. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
21. Sain, S.R. The nature of statistical learning theory. *Technometrics* **1996**, *38*, 409. [[CrossRef](#)]
22. Li, G.; Zhu, Y.; Yang, Q.; Zou, J.; Cai, T.; He, P.; Zhang, Y.; Zhao, Y.; Tian, X. A detection method for foreign objects on coal mine conveyor belts based on an improved Hyper-YOLO. *J. Mine Autom.* **2025**, *51*, 114–121.
23. Dong, H.F.; Lin, Z.H.; Li, Z.X. Research on Foreign Object Recognition Method for Coal Mine Conveyor Belt Based on Improved RegNet. *Inf. Technol. Informatiz.* **2025**, 188–191. [[CrossRef](#)]
24. Jiang, F. Research on Deep Learning-Based Monitoring System for Foreign Objects in Coal Flow on Belt Conveyors. Doctoral Dissertation, China University of Mining and Technology, Xuzhou, China, 2023. [[CrossRef](#)]
25. Huang, K.; Li, S.; Cai, F.; Zhou, R. Detection of large foreign objects on coal mine belt conveyor based on Improved. *Processes* **2023**, *11*, 2469. [[CrossRef](#)]
26. Hao, S.; Zhang, X.; Ma, X.; Sun, S.; Wen, H.; Wang, J. Foreign object detection in coal mine conveyor belt based on CBAM-YOLOv5. *J. China Coal Soc.* **2022**, *47*, 4149–4158.
27. Mei, X.; Lyu, X.; Lei, M. Foreign object detection of coal mine underground conveyor belt based on Stair-YOLOv7-tiny. *J. Mine Autom.* **2024**, *50*, 99–104.
28. Tang, J.; Li, J.; Shi, Q.; Yang, P.; Wang, R. Real-Time Foreign Object Detection for Belt Conveyors Based on Faster-YOLOv7. *Ind. Min. Autom.* **2023**, *49*, 46–52+66. [[CrossRef](#)]
29. Wang, C.; Hao, B.N.; Zhang, L.Y.; Yang, Z.F. Conveyor belt foreign object detection technology based on RetinaNet in underground coal mine. *Coal Mine Mach.* **2022**, *43*, 180–183.
30. Dai, L.; Zhang, X.; Gardoni, P.; Lu, H.; Liu, X.; Krolczyk, G.; Li, Z. A new machine vision detection method for identifying and screening out various large foreign objects on coal belt conveyor lines. *Complex Intell. Syst.* **2023**, *9*, 5221–5234. [[CrossRef](#)]
31. Yao, R.; Qi, P.; Hua, D.; Zhang, X.; Lu, H.; Liu, X. A foreign object detection method for belt conveyors based on an improved YOLOX model. *Technologies* **2023**, *11*, 114. [[CrossRef](#)]
32. Luo, B.; Kou, Z.; Han, C.; Wu, J.; Liu, S. A faster and lighter detection method for foreign objects in coal mine belt conveyors. *Sensors* **2023**, *23*, 6276. [[CrossRef](#)] [[PubMed](#)]
33. Wang, Y.; Wang, Y.; Dang, L. Video detection of foreign objects on the surface of belt conveyor underground coal mine based on improved SSD. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 5507–5516. [[CrossRef](#)]
34. Miao, D.; Wang, Y.; Yang, L.; Wei, S. Foreign object detection method of conveyor belt based on improved Nanodet. *IEEE Access* **2023**, *11*, 23046–23052. [[CrossRef](#)]
35. Zhang, M.; Cao, Y.; Jiang, K.; Li, M.; Liu, L.; Yu, Y.; Zhou, M.; Zhang, Y. Proactive measures to prevent conveyor belt Failures: Deep Learning-based faster foreign object detection. *Eng. Fail. Anal.* **2022**, *141*, 106653. [[CrossRef](#)]
36. Szurgacz, D.; Zhironkin, S.; Vöth, S.; Pokorný, J.; Spearing, A.J.S.; Cehlár, M.; Stempniak, M.; Sobik, L. Thermal imaging study to determine the operational condition of a conveyor belt drive system structure. *Energies* **2021**, *14*, 3258. [[CrossRef](#)]
37. Cheng, D.Q.; Xu, J.Y.; Kou, Q.Q.; Zhang, H.X.; Han, C.G.; Yu, B.; Qian, J.S. Lightweight network based on residual information for foreign body classification on coal conveyor belt. *J. China Coal Soc.* **2022**, *47*, 1361–1369.
38. Yang, W.; Zhang, X.; Ma, B.; Wang, Y.; Wu, Y.; Yan, J.; Liu, Y.; Zhang, C.; Wan, J.; Wang, Y.; et al. An open dataset for intelligent recognition and classification of abnormal condition in longwall mining. *Sci. Data* **2023**, *10*, 416. [[CrossRef](#)] [[PubMed](#)]

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