

YOLOv11-SFF: An Improved Algorithm for Detecting Foreign Objects on Mine Conveyor Belts

Wenxi Zhang¹, Jianming Zheng¹, Zhongkai Ma^{1,2}

¹ Jilin University of Chemical Technology, Jilin, Jilin, 132022, China

² Jilin Urban and Rural Planning Research Institute, Changchun, Jilin, 132000, China

Abstract: To solve the problem of detecting foreign objects on mine conveyor belts with different backgrounds and variations in target scales, difficulty recognising elongated targets, and missed targets or false negatives for small foreign objects, this paper proposes YOLOv11-SFF, a conveyor belt foreign object detection algorithm based on an improved version of YOLOv11n. YOLOv11-SFF first introduces a C3k2_S module based on direction-aware strip convolutions to enhance the model's ability to learn the long-term dependence structure of thin targets. Strip convolution performs horizontal and vertical convolution operations. Secondly, there is an FSPPF module that combines the FReLU activation function to augment deep feature representation and aggregate multi-scale contextual information for improved recognition, enabling the model to detect small objects. Finally, the FAFM Feature Alignment and Fusion Module is added in the neck network. Deformable convolutions and content-aware attention are used to achieve cross-scale feature space alignment and adaptive fusion, thus improving the detection accuracy of small, long and low-contrast objects on the conveyor belt. Experimental results demonstrate that, compared to the baseline model, the precision, mAP@50 and mAP@50-95 of YOLOv11-SFF increase by 2.4, 2.6 and 1.9 percentage points respectively, validating the effectiveness of YOLOv11-SFF in the task of detecting foreign objects on mine conveyor belts.

Keywords: Mine Conveyor Belt; Foreign Object Detection; YOLOv11-SFF; Strip Convolution; Feature Alignment and Fusion.

1. Introduction

Mine conveyor belts are vital equipment in coal mine transport systems, and their operational status directly affects production efficiency and operational safety [1]. During actual operation, anchor bolts and large foreign objects may become lodged on the conveyor belt due to production activities or unforeseen circumstances; if these are not cleared in a timely manner, they can easily cause equipment failures and, in severe cases, may even lead to safety incidents [2] [3]. Traditional manual inspection methods are hampered by factors such as insufficient underground lighting, dust interference and complex backgrounds, which obstruct visibility; consequently, the detection efficiency and reliability of such methods struggle to meet the requirements of intelligent coal mining [4].

In recent years, with the advancement of artificial intelligence, object detection algorithms have gradually been applied to research into the detection of foreign objects on mine conveyor belts [5]. Guo et al. [6] proposed a low-light coal mine conveyor belt foreign object recognition algorithm based on an improved YOLOv8 model. This approach enhances image quality through low-light image enhancement and introduces a dynamic serpentine convolutional enhancement model to improve feature extraction capabilities for elongated foreign objects. Furthermore, by combining a lightweight neck network and an improved loss function, it reduces model complexity whilst optimising detection performance. Wang et al. [7] constructed a foreign object detection model for coal mine conveyor belts based on YOLOv11. By improving the structure of the backbone feature extraction, spatial pyramid pooling and detection branches, they enhanced the model's ability to recognise small-sized and low-resolution targets. Lin et al. [8] proposed the UD-DETR lightweight detection

network to address the issues of a large number of parameters and high computational complexity in RT-DETR. By improving the backbone convolutional layers and the feature pyramid structure, they reduced the computational load of the model whilst enhancing the detection performance for small foreign objects.

The aforementioned research has achieved some success in the field of foreign object detection on mine conveyor belts; however, shortcomings remain in the extraction of features from elongated objects, the preservation of features for small objects, and the fusion of multi-scale features. On the one hand, elongated objects account for a low proportion of effective pixels in images and are easily confused with conveyor belt textures or shadowed areas; on the other hand, the features of small and low-contrast objects are not prominent and are easily weakened during propagation through deep neural networks. Furthermore, features at different levels differ in terms of spatial resolution and semantic representation; when fused, feature misalignment is likely to occur, leading to false negatives and false positives.

To address these issues, this paper proposes the YOLOv11-SFF model for foreign object detection on mine conveyor belts, based on YOLOv11n. The model incorporates improvements in three areas: feature extraction for elongated objects, deep multi-scale feature enhancement [9], and multi-scale feature fusion [10], with the aim of enhancing the model's detection performance for small, low-contrast and elongated objects in complex mine environments.

The main contributions of this paper are as follows:

- (1) The C3k2_S module is proposed, which employs directional strip convolution to enhance the model's ability to extract directional features of elongated foreign objects [11];
- (2) The FSPPF multi-scale feature enhancement module is introduced, utilising multi-level pooling and the FReLU

activation function to improve deep-layer contextual modelling capabilities [12];

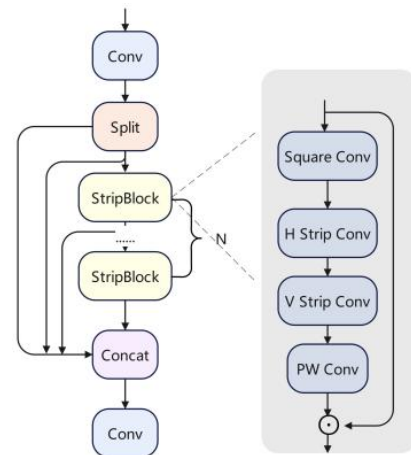
2. The YOLOv11 Model

Overall, YOLOv11 demonstrates good real-time detection performance; however, in the task of detecting foreign objects on mine conveyor belts, its feature extraction and fusion capabilities could be further optimised when dealing with slender foreign objects such as anchor bolts, small targets and complex background interference. YOLOv11 offers a variety of model sizes. Taking into account the requirements for real-time performance and deployment costs in on-site mine detection tasks, this paper selects the lightweight YOLOv11n as the baseline model and modifies it in accordance with the characteristics of foreign objects on conveyor belts.

Figure 1. YOLOv11 network architecture

3.1. C3k2_S Directional Feature Enhancement Module

convolution. Although it is capable of extracting texture and edge information from objects, in the context of foreign object detection on mine conveyor belts, slender foreign objects account for a low proportion of effective pixels in the image and are easily affected by interference from conveyor belt textures and shadowed areas. As conventional convolutions are constrained by a fixed receptive field, they lack the ability to model long-range dependencies for elongated objects; consequently, they struggle to adequately capture the continuous structural features of such foreign objects, which in turn further compromises the model's performance in localising and recognising this class of objects.



This structure first extracts local contextual information via small-sized square convolutions, then utilises horizontal and vertical strip convolutions to capture long-range dependencies of the target in the horizontal and vertical directions, respectively. Finally, channel information fusion is achieved via 1×1 pointwise convolutions, which are then multiplied element-wise with the input features to obtain a direction-enhanced feature representation. The feature extraction process can be represented as:

$$Y = \text{Conv}_{1 \times 1} \left(\text{Conv}_{k \times 1} \left(\text{Conv}_{1 \times k} \left(\text{Conv}_{k \times k}(X) \right) \right) \right) \quad (1)$$

$$F_{out} = X \odot Y \quad (2)$$

Where X denotes the input features, Y denotes the feature weights obtained via directional enhancement, and $\odot$ denotes element-wise multiplication.

3.2. FSPPF Multi-scale Feature Enhancement Module

using multi-scale pooling in the SPPF module, small and low-contrast targets, which are necessary for detecting foreign objects on mine conveyor belts, only have a very small number of effective pixels; their fine-scale features are easily lost in deep neural networks, and thus the network cannot build a strong semantic representation for weak-textured targets.

The new model has replaced the first SPPF block with an FSPPF block, and the structure of this block is shown in Figure 3. Multi-scale pooling is used to collect information from many receptive fields at different scales in the FSPPF module, and then, based on the features of this information, an adaptive combination method is employed to pay more attention to the key target area. First, a 1x1 convolution is employed to reduce the dimension of the input features; then, several max-pooling operations are carried out to obtain contextual information at various receptive fields, and finally, multi-scale feature fusion is achieved by convolutions. FSPPF adds an FReLU activation function to help learn the local spatial structure of the features better. The Steps of Calculation are as follows:

$$X_0 = \text{FReLU}(\text{BN}(\text{Conv}_{1 \times 1}(X))) \quad (3)$$

$$X_i = \text{MaxPool}(X_{i-1}), \quad i=1,2,3 \quad (4)$$

$$Y = \text{FReLU}(\text{BN}(\text{Conv}_{1 \times 1}(\text{Concat}(X_0, X_1, X_2, X_3)))) \quad (5)$$

FReLU incorporates local spatial conditional information into the activation process, and is expressed as follows:

$$f(x_{c,i,j}) = \max(x_{c,i,j}, T(x_{c,i,j})) \quad (6)$$

$$T(x_{c,i,j}) = x_{c,i,j}^\omega \cdot p_c^\omega \quad (7)$$

X denotes the input features, X_0 denotes the features after the first convolution, Y is the output features of the FSPPF module, $x_{c,i,j}$ is the pixel value of the input feature at position (i, j) in the c th channel, $T(\cdot)$ denotes the local spatial condition function, and p_c^ω denotes the learnable parameters of this window.

FSPPF adds context modelling in the feature fusion stage to improve upon SPPF, and it is relatively light in terms of computation. Therefore, there is a loss of fine-grained information, but the detection performance of the model for multi-scale and low-contrast objects in complex scenes has been improved.

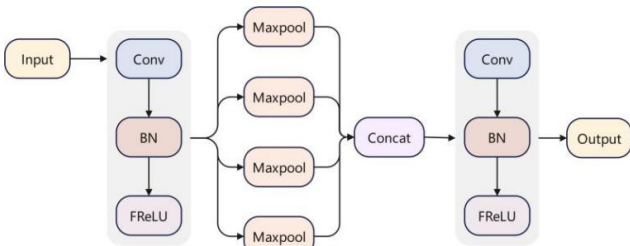


Figure 3. Structure of the FSPPF module

3.3. FAFM Feature Alignment and Fusion Module

At the stage of feature fusion, the old ways of concatenation and addition at different levels are not good at handling the semantic differences and spatial structure inconsistencies of these features. Otherwise, there will be a mismatch in the feature space of different scales; the effect of feature fusion will be reduced, small targets will not be detected, and false alarms will occur in complex environments. Therefore, a FAFM (Feature Alignment and

Fusion) module has been added to the Neck part. The Structure is as follows: As shown in Figure 4, there are two stages in the adaptive feature alignment and enhancement: deformable convolution alignment and content-guided attention fusion.

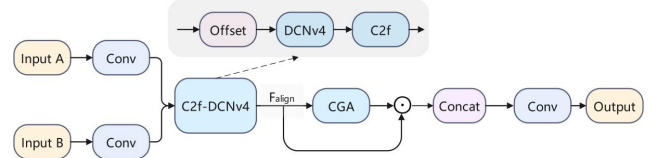


Figure 4. FAFM module architecture

During the feature alignment stage, deformable convolution is employed to achieve spatial position correction. This is primarily achieved by embedding DCNv4 into the C2f module to form C2f-DCNv4, which performs spatial position correction on the input features. Unlike standard convolution, which uses a fixed sampling grid, deformable convolution learns spatial offsets to dynamically adjust the sampling positions of the convolution kernels, thereby aligning features from different levels within the target key regions and mitigating the issue of spatial misalignment between features at different levels. Let the input features be denoted as, then the alignment process can be expressed as:

$$F_{align}(p) = \sum_k w_k \cdot F(p + p_k + \Delta p_k) \quad (8)$$

Where w_k denotes the convolution weights, p_k denotes the standard convolution sampling position, Δp_k denotes the offset obtained through learning.

Feature alignment has been carried out, and now FAFM will use a Content-Guided Attention (CGA) mechanism to incorporate these features further and determine attention weights at the three levels of channel, spatial and pixel. Attention is paid to the relationship among all feature channels, the location of the important area is determined by spatial attention, and at the pixel level, some weights are given to the features. Therefore, the response of object edges and important texture areas will be improved, and interference from all kinds of background noise will be reduced.

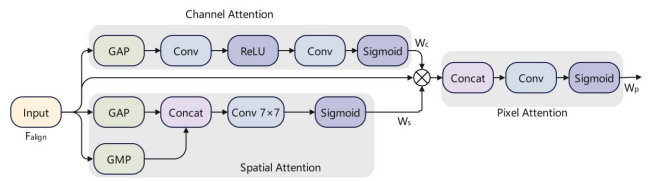


Figure 5. Structure of the Content-Guided Attention Mechanism

The specific structure of CGA is shown in Figure 5. Let the aligned features be denoted by; the channel attention weights can be expressed as:

$$W_c = \sigma(\text{Conv}_2(\delta(\text{Conv}_1(\text{GAP}(F_{align})))) \quad (9)$$

The spatial attention weights can be expressed as:

$$W_s = \sigma(\text{Conv}([\text{GAP}(F_{align}), \text{GMP}(F_{align})])) \quad (10)$$

The pixel attention weights are then constructed as follows:

$$W_p = \sigma(\text{Conv}([W_c \otimes F_{align}, W_s \otimes F_{align}])) \quad (11)$$

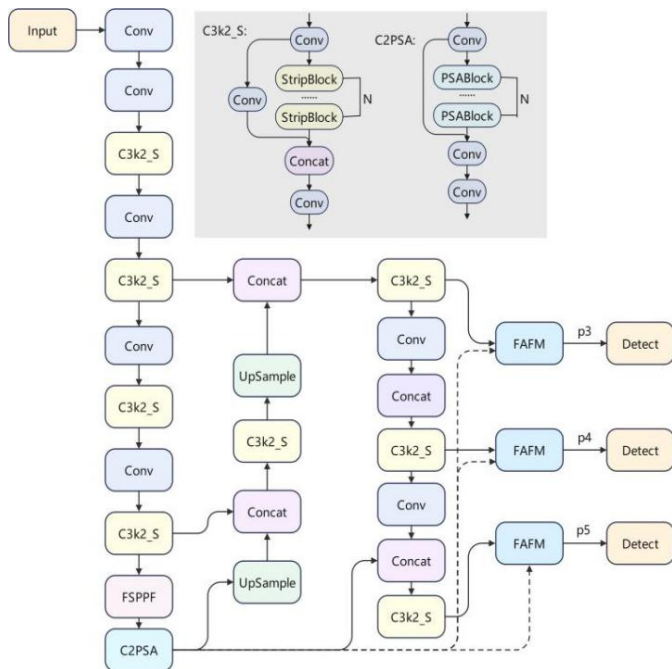
The final fused features are:

$$F_{out} = F_{align} \odot W_p + F_{align} \quad (12)$$

Where $\sigma(\cdot)$ denotes the sigmoid function, $\delta(\cdot)$ denotes a non-linear activation function, and $\odot$ denotes element-wise

multiplication.

3.4. Overall Structure of YOLOv11-SFF



4. Experiments and Analysis of Results

Table 1.

Table 1. Experimental parameter configuration

4.2. Experimental Evaluation Metrics

$$P = \frac{TP}{TP + FP} \quad (13)$$

In the equation, TP represents the number of correct predictions of foreign objects; FP represents the number of instances where background or non-target objects were predicted as foreign objects; FN represents the number of actual foreign objects not detected by the model; and N represents the number of target classes in the dataset.

4.3. Analysis of Ablation Experiments

Experiment shows the results of the original YOLOv11n model without any modifications; In Experiments 2–4, replacing the C3k2 module with the C3k2 S module

increased the precision to 91.3 percent, with mAP@50 increased to 92.9%, indicating that C3k2_S effectively enhances the model's ability to capture directional features of slender foreign objects, whilst the FLOPs increased to 9.0 G, suggesting that this module incurs a certain computational overhead; Upon replacing the SPPF module with the FSPPF module, mAP@50 and mAP@50-95 increased to 93.0% and

60.5%, whilst FLOPs remained consistent with the baseline model, indicating that the FSPPF module enhances the model's ability to represent deep, multi-scale features without increasing computational complexity; Following the introduction of the FAFM module, mAP@50 increased to 92.6%, whilst precision showed a decline, indicating that FAFM has a certain beneficial effect on feature fusion.

Table 2. Results of the ablation experiments

Experiment	C3k2_S	FSPPF	FAFM	P/%	mAP@50/%	mAP@50-95/%	FLOPs/G
1				89.1	91.7	59.6	6.3
2	√			91.3	92.9	59.4	9.0
3		√		91.3	93.0	60.5	6.3
4			√	89.0	92.6	59.4	6.6
5	√	√		89.9	93.7	61.4	9.0
6	√		√	88.6	92.5	59.3	9.3
7	√	√	√	91.5	94.3	61.5	9.3

In combination experiments 5–6, building upon the C3k2_S module, the addition of either the FSPPF or FAFM module resulted in varying degrees of improvement in mAP@50; when the three improved modules—C3k2_S, FSPPF and FAFM were used in combination, mAP@50 reached 94.3%, mAP@50-95 reached 61.5%, representing improvements of 2.6 and 1.9 percentage points respectively compared to the baseline model, although the FLOPs were 9.3G, detection accuracy was significantly enhanced, indicating that the synergistic effects of the three modules in feature extraction, feature enhancement and feature fusion can improve the model's ability to recognise foreign objects on mine conveyor belts and enhance its robustness.

4.4. FAFM Module Position Ablation Experiment

To further validate the fusion effectiveness of the FAFM module across feature layers at different scales, this paper introduces the FAFM module before the P3, P4 and P5 detection layers, respectively, and conducts position ablation experiments under a unified experimental environment. The experiments only alter the insertion position of the FAFM module, whilst the remaining network structure and training parameters remain consistent. The experimental results are shown in Table 3.

Table 3. Results of the FAFM Module Position Ablation Experiment

FAFM insertion point	P/%	mAP@50/%	mAP@50-95/%
P3 level	84.9	90.0	58.3
P4 level	85.8	91.4	59.4
P5 level	85.1	91.8	58.3
P3, P4, and P5 levels	89.0	92.6	59.4

As shown in the table, there are some differences in the detection performance of the model when the FAFM module is added to the feature layer at different scales. Adding the FAFM module at a single level, such as P3, P4 or P5, will increase the fusion of features at that scale, but the overall improvement will be relatively small. When the FAFM module was added to the beginning of all three detection stages (P3, P4 and P5) simultaneously, the P, mAP@50 and mAP@50-95 of the model reached 89.0%, 92.6% and 59.4% respectively, and the overall performance was better than that obtained by adding FAFM in a single stage. Based on the above results, it has been concluded that by adding the FAFM module at different levels of feature maps simultaneously, fine-grained spatial details, intermediate structural

information and deep semantic knowledge can be integrated more effectively to improve the object detection performance at all scales.

4.5. Comparative Analysis of Training Loss

To further analyse the impact of each improved module on the convergence of the model, this paper compares the changes in training loss at different times after improvement and shows the results in Figure 7. As shown in the figure, the loss values of all four groups of models gradually decrease with the increase of training iterations and stabilise in the later stages of training; thus, it can be concluded that the training process for each model is relatively stable.

The loss curve of YOLOv11n after adding C3k2_S is generally lower than that of the baseline model, so this structure has improved the model's learning of directional information for thin foreign objects. Therefore, based on the above, FSPPF is added to further reduce the loss, and it can be concluded that the multi-scale feature enhancement structure helps the model obtain richer deep-level features. After adding FAFM, the loss curve of YOLOv11-SFF was the lowest overall and remained relatively stable in the later stages; thus, it can be concluded that the synergy of the three modules has strengthened the model's ability to learn features of foreign objects on mine conveyor belts.

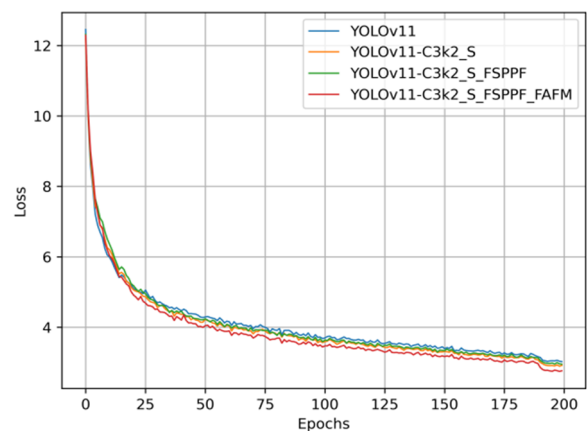


Figure 7. Training loss curves at different stages of improvement

4.6. Comparative Experimental Analysis

To further verify the detection results of the YOLOv11-SFF model, several other object detection algorithms on the same dataset have also been tested in this paper, such as different

versions of YOLO (YOLOv8, YOLOv12 and YOLOv26), D-FINE and RT-DETR. To ensure the fairness of the experiment, all the models to be compared will be trained in the same manner with data division and training parameters. The three evaluation indicators are Precision, Recall and mAP@50; the corresponding experiment results are shown in Table 4.

As can be seen from the table, YOLOv11-SFF achieved the highest mAP@50 and mAP@50-95 among the compared models, reaching 94.3% and 61.5% respectively, indicating that YOLOv11-SFF demonstrates superior overall performance in object classification and bounding box

localisation. Although the D-FINE and RT-DETR models outperform YOLOv11-SFF in terms of precision and recall, their computational cost is significantly higher than that of YOLOv11-SFF. Compared with other YOLO models, YOLOv11-SFF achieves higher mAP@50 and mAP@50-95 at the cost of increased computational complexity. Comparative experimental results indicate that YOLOv11-SFF strikes a good balance between detection accuracy and computational overhead, making it more suitable for foreign object detection scenarios on mine conveyor belts.

Table 4. Results of Comparative Experiments

method	P/%	R/%	mAP@50/%	mAP@50-95/%	FLOPs/G
YOLOv8	85.3	85.3	91.5	58.4	8.1
YOLOv9	92.5	83.3	93.1	58.8	7.6
YOLOv10	83.2	82.3	87.6	55.5	6.5
YOLOv11	89.1	84.3	91.7	59.6	6.3
YOLOv12	88.5	83.8	91.4	57.4	6.3
YOLOv26	86.9	84.1	90.8	57.1	5.2
D-FINE	92.3	85.6	91.6	61.4	21.4
RT-DETR	93	85.3	93.5	58.7	57.2
YOLOv11-SFF	91.5	85	94.3	61.5	9.3

4.7. Analysis of YOLOv11-SFF Detection Results

To further analyse the overall performance of the YOLOv11-SFF model during training, the improved model was trained and tested on our self-built dataset for 200 training epochs, and the changes in training loss and evaluation metrics were analysed; the experimental results are

shown in Figure 8. The experimental data show that the various loss functions declined rapidly during the early stages of training before stabilising as the number of iterations increased. Furthermore, the trends in validation set loss and training set loss were largely consistent, indicating the training and validation curves showed no obvious divergence indicative of severe overfitting. Metrics such as precision, recall, mAP@50 and mAP@50-95 improved initially before gradually converging.

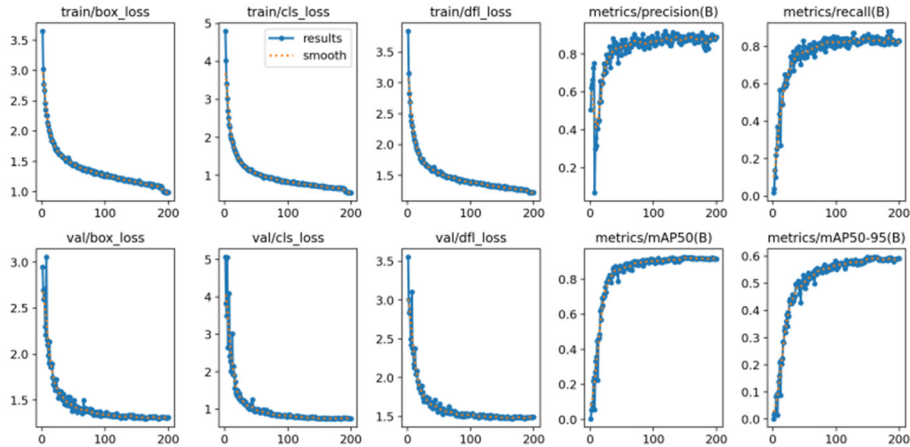


Figure 8. YOLOv11-SFF training performance curve

To further validate the detection performance of the improved model, experiments were conducted using both the YOLOv11-SFF model and the original YOLOv11n model on the same dataset, with detection comparisons carried out across scenarios involving multiple targets, small targets and low-resolution targets. The experimental results are shown in Figure 9 below. In Figure 9(a), in a multi-object scenario, the YOLOv11-SFF model’s detection bounding boxes for anchor bolts and large foreign objects fit the target areas more closely, and the overall confidence levels are higher than those of the YOLOv11n model; in Figure 9(b), the original YOLOv11n model exhibits low detection confidence for some small objects, whilst the YOLOv11-SFF model is able to detect multiple small objects with greater stability; In Figure 9(c), the original YOLOv11n model exhibits missed detections in

low-pixel-count scenarios, whereas the YOLOv11-SFF model successfully detects the corresponding low-pixel-count targets. The experimental results validate that the YOLOv11-SFF model demonstrates superior detection stability in scenarios involving multiple targets, small targets and low-pixel-count targets.

5. Conclusion

This paper proposes the YOLOv11-SFF detection model to address issues such as the difficulty in recognising elongated objects, the omission of small objects, and interference from complex environments in the detection of foreign objects on mine conveyor belts. The model utilises C3k2_S, FSPFP and FAFM to improve the feature extraction of elongated objects,

deep multi-scale feature representation and multi-scale

feature fusion respectively.

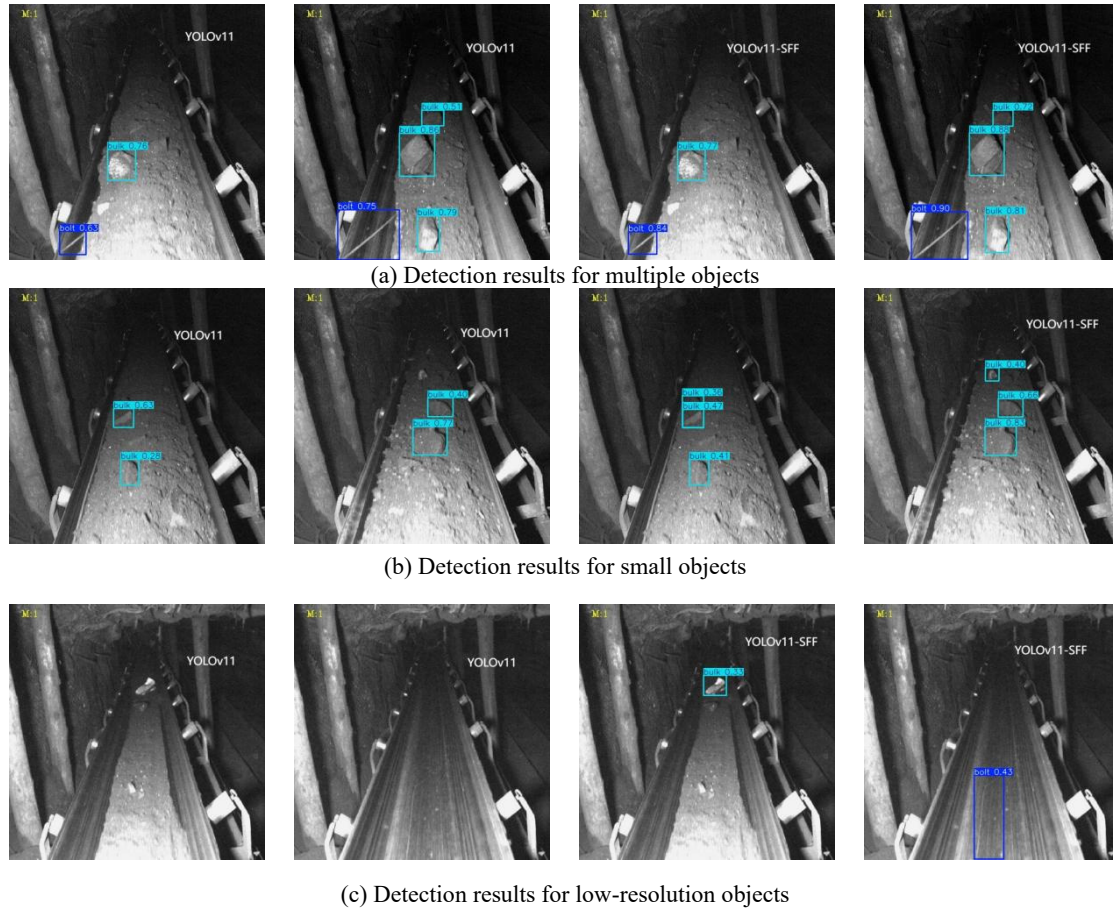


Figure 9. Comparison of detection results between YOLOv11n and YOLOv11-SFF under different scenarios

According to the above experimental results, YOLOv11-SFF has increased the precision, $mAP@50$ and $mAP@50-95$ by 2.4%, 2.6% and 1.9% respectively, compared with the baseline model. Ablation and comparison experiments have shown that the improved model is also good. In the future, I will reduce the computational complexity of the model and improve its real-time capability.

References

- [1] Gao, R. (2025). Research on foreign object detection in coal mine conveyor belts based on an optimized YOLOv8 framework [Master's thesis]. Shanxi University. <https://doi.org/10.27284/d.cnki.gsxix.2025.001402>.
- [2] Wang, Y. (2022). Design of a large foreign object detection system for belt conveyors based on machine vision [Master's thesis]. China University of Mining and Technology. <https://doi.org/10.27623/d.cnki.gzkyu.2022.000183>.
- [3] Liu, J., & Wang, H. (2025). Research and application of video-based recognition technology for safety hazards in coal mine belt conveyors. *Coal Mine Machinery*, 46, 158–163. <https://doi.org/10.13436/j.mkjx.202507043>.
- [4] Wang, G., Zhao, G., & Ren, H. (2019). Analysis on key technologies of intelligent coal mine and intelligent mining. *Journal of China Coal Society*, 44, 34–41. <https://doi.org/10.13225/j.cnki.jccs.2018.5034>.
- [5] Jin, Z., Shi, L., Geng, Y., & et al. (2026). Research status and prospects of YOLO algorithms for object detection in fully mechanized coal mining faces. *Coal Science and Technology*, 1–24. <https://link.cnki.net/urlid/11.2402.TD.20260513.1444.004>.
- [6] Guo, Z., & Zhang, L. (2024). Foreign object recognition on coal mine conveyor belts under low-light conditions using an improved YOLOv8 algorithm. *Electronic Measurement Technology*, 47, 188–196. <https://doi.org/10.19651/j.cnki.emt.2416685>.
- [7] Wang, H., & Han, L. (2025). Foreign object detection on coal mine conveyor belts based on an improved YOLOv11 model. *Electronic Design Engineering*, 33, 66–70. <https://doi.org/10.14022/j.issn1674-6236.2025.14.013>.
- [8] Lin, Y., Zheng, J., & Li, Z. (2025). Small-object foreign object detection based on a lightweight UD-DETR network. *Informatization Research*, 51, 71–78.
- [9] Cheng, J., Bian, C., Zhang, J., & et al. (2025). Enhanced object detection using parallel pooling attention and multi-feature fusion. *Modern Electronics Technique*, 48, 59–67. <https://doi.org/10.16652/j.issn.1004-373x.2025.05.010>.
- [10] Ma, H., Liu, L., & Gao, K. (2024). Lightweight foreign object detection on coal conveyor belts via multi-scale feature fusion. *Journal of Safety and Environment*, 24, 4617–4629. <https://doi.org/10.13637/j.issn.1009-6094.2024.0888>.
- [11] Yuan, X., Zheng, Z. H., Li, Y., & et al. (2026). Strip R-CNN: large strip convolution for remote sensing object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, pp. 12259–12267). <https://doi.org/10.1609/aaai.v40i15.38217>.
- [12] Wang, S., Wu, M., Wei, X., & et al. (2025). An advanced multi-source data fusion method utilizing deep learning techniques for fire detection. *Engineering Applications of Artificial Intelligence*, 142, 109902.
- [13] Zhang, W., Zhang, X., Xu, X., & et al. (2026). MHFNet: multimodal hybrid fusion framework for misaligned SAR–optical ship detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 231, 151–166.