Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 6, No. 3, 2026

YOLO-based Lightweight Drill Detection for Coal Mines

Dongyu Zhang

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, Henan, China

Abstract: Aiming at the challenges of insufficient illumination, heavy dust interference, large scale variation of targets, and limited computational resources of edge devices in underground coal mine drilling scenarios, this paper proposes a lightweight object detection model named YOLO-Drill. The proposed model is developed based on the YOLO framework with several improvements. First, a Lightweight Drill Feature Block (LDFB) is designed by integrating depthwise separable convolution and directional perception mechanisms, which reduces computational complexity while enhancing the representation ability of slender targets. Second, a Drill-Aware Feature Pyramid Network (DAFPN) is constructed to achieve efficient multi-scale feature fusion through bidirectional cross-layer interaction and attention weighting, thereby improving the detection performance of small and occluded targets. In addition, a scene-adaptive data augmentation strategy is introduced to enhance the robustness of the model under low-light, high-dust, and low-contrast conditions. Finally, the CIoU loss is adopted to optimize bounding box regression and improve localization accuracy. Experimental results on the self-constructed Drill Scene Dataset (DSD) demonstrate that the proposed YOLO-Drill achieves 85.2% mAP@0.5 and 51.0% mAP@0.5:0.95, which outperform the baseline by 6.6% and 5.8%, re-spectively. Meanwhile, the model maintains a lightweight structure with only 6.0M parameters and 19.9 GFLOPs, achieving a real-time inference speed of 87.9 FPS. The results verify that the proposed method effectively balances detection accuracy and computational efficiency, showing strong applicability in underground coal mine drilling scenarios.

Keywords: Coal mine underground, Drill detection, Lightweight network, Multi-scale feature fusion, Edge computing.

1. Introduction

Coal mine underground drilling is a core operation in coal production, directly affecting the safety and efficiency of mining activities. Real-time and accurate detection of drilling targets (such as drill body, drill pipe, drill bit, and drill tail) is a fundamental prerequisite for achieving intelligent monitoring of underground operations, preventing equipment failures, and ensuring the safety of workers [1]. However, the harsh underground working environment poses significant challenges to object detection tasks, severely limiting the effectiveness of traditional detection models [2].

At present, mainstream object detection models represented by YOLO series [3], Faster R-CNN [4] and CenterNet [5] have achieved excellent performance in general scenarios. However, they exhibit obvious limitations when applied to underground drilling target detection in coal mines. On the one hand, underground environments are characterized by insufficient illumination, severe dust interference, low image contrast, and significant noise, which result in blurred target features and poor distinction between targets and background [6], leading to frequent false detections and missed detections. On the other hand, underground detection devices are typically edge computing systems with limited computational power and memory, making it difficult for traditional models with large parameter sizes and high computational complexity to meet real-time detection requirements [7]. In addition, drilling targets are mostly slender structures, and conventional feature extraction and multi-scale fusion methods struggle to effectively capture their contour and morphological characteristics, resulting in low localization accuracy.

To address the above issues, this paper proposes a lightweight object detection model, YOLO-Drill, based on the YOLO framework and tailored for underground drilling target detection. The main improvements of the proposed model include: designing a lightweight drill feature extraction block (LDFB) to enhance slender object feature representation while reducing model complexity; proposing an improved multi-scale feature fusion structure (DAFPN) to handle large scale variations of underground targets; introducing a scene-adaptive optimization strategy to improve robustness under harsh conditions; and optimizing the loss function to enhance localization accuracy of drilling targets.

The main contributions of this paper are summarized as follows: (1) a lightweight feature extraction module LDFB is designed, integrating depthwise separable convolution and directional awareness mechanisms to achieve both computational efficiency and task-specific feature extraction; (2) an improved multi-scale feature fusion structure DAFPN is proposed, which enhances detection performance for small and occluded targets through bidirectional cross-layer interaction and attention weighting; (3) a scene-adaptive data augmentation strategy for harsh underground environments is constructed, improving model adaptability to low-light and high-dust conditions without increasing inference burden; (4) the CIoU loss is adopted to optimize bounding box regression, enhancing localization accuracy for slender drilling targets while balancing training stability and detection precision.

The remainder of this paper is organized as follows: Chapter 2 reviews related work on object detection in coal mine scenarios and multi-scale feature fusion; Chapter 3 presents the overall architecture and key module designs of the YOLO-Drill model; Chapter 4 validates the effectiveness of the proposed method through ablation and comparative experiments, along with visualization analysis; Chapter 5 concludes the paper and discusses limitations and future research directions.

2. Related-work

2.1. Object Detection in Underground Coal Mine Scenarios

In recent years, with the advancement of coal mine intelligence, an increasing number of researchers have applied object detection technologies to underground scenarios, focusing on addressing challenges such as harsh environments and poor real-time performance [8]. Existing studies can generally be divided into two categories: improvements to traditional models and the design of customized lightweight models.

In terms of improving traditional models, some researchers have optimized existing frameworks based on the characteristics of underground environments. For example, Chen et al. introduced a spatial attention mechanism into the YOLO framework to address uneven illumination and noise interference in underground images, effectively enhancing target perception in low-quality images; however, the model still incurs considerable computational cost, limiting its application on edge devices [9]. Zhou et al. tackled image degradation caused by high dust conditions by incorporating dehazing enhancement and cross-modal feature fusion into the YOLO model, improving detection accuracy in complex environments, but at the expense of increased model complexity [10]. In addition, some studies enhance detection heads or feature modeling strategies to improve recognition performance under occlusion and low-contrast conditions, often sacrificing inference speed [11].

2.2. Multi-scale Feature Fusion Technology

Multi-scale feature fusion is a key technique in object detection, effectively addressing the problem of large variations in object scale. Traditional feature fusion methods can be broadly categorized into unidirectional and bidirectional fusion. The Feature Pyramid Network (FPN) proposed by Lin et al. achieves top-down feature fusion [12], transferring high-level semantic features to lower layers to improve small object detection. However, it lacks sufficient interaction between shallow and deep features, resulting in underutilization of fine-grained details.

To address this issue, various improved multi-scale fusion structures have been proposed. Path Aggregation Network (PANet) [13] introduces a bottom-up fusion path on top of FPN, enabling bidirectional feature interaction and improving detection accuracy for medium and small objects. BiFPN [14] further optimizes fusion paths by removing redundant feature layers, enhancing efficiency, though its complex structure increases computational cost. Additionally, attention mechanisms such as CBAM have been incorporated into feature fusion, enabling the model to emphasize informative features and suppress background noise, thereby improving feature representation capability.

In the field of underground coal mine object detection, multi-scale feature fusion techniques have also been widely applied. Existing studies indicate that introducing multi-scale feature interaction can improve detection performance for small and occluded targets to a certain extent. However, due to the complexity of underground environments—such as low illumination, heavy dust interference, and the slender shape of targets—existing fusion methods based on FPN and PANet still exhibit limitations in practical applications. On the one hand, although multi-scale fusion enhances semantic representation, it has limited capability in capturing fine-grained structural features, making it difficult to accurately describe slender objects such as drill pipes. On the other hand, most existing fusion structures have high computational complexity and lack lightweight optimization tailored for underground edge devices, thereby limiting their real-time performance and deployment efficiency.

2.3. Lightweight Neural Network Design

Lightweight neural network design is a key approach to addressing the computational constraints of underground edge devices. The main strategies include model pruning, quantization, and lightweight architecture design. Among them, lightweight network architectures such as MobileNet [16], ShuffleNet [17] and EfficientNet [18] have become mainstream research directions due to their ability to achieve a good balance between accuracy and efficiency.

MobileNet replaces standard convolutions with depthwise separable convolutions, significantly reducing the number of parameters and computational complexity. MobileNetV2 further enhances feature extraction capability and efficiency by introducing inverted residual structures and linear bottlenecks. ShuffleNet addresses the issue of inter-channel information isolation through channel shuffle operations, improving model performance under low parameter budgets. EfficientNet adopts compound scaling to jointly adjust network depth, width, and resolution, achieving better performance with fewer parameters.

In underground coal mine object detection applications, lightweight network structures have been increasingly adopted. For example, Luo et al. proposed a lightweight detection method based on YOLOv5 by improving the backbone and feature fusion structure, significantly reducing model parameters and computational complexity while maintaining detection accuracy, making it more suitable for real-time underground scenarios [19]. Fan et al. proposed the CM-YOLOv8 model, achieving approximately 40% model compression through pruning and lightweight design strategies, while maintaining high detection accuracy in fully mechanized mining face scenarios [20]. However, existing lightweight methods mainly focus on model compression and general object detection, and still lack sufficient capability in representing slender drilling targets.

In summary, current technologies for underground drilling target detection in coal mines still suffer from poor environmental adaptability, difficulty in balancing accuracy and efficiency, and insufficient specialization for slender targets. Multi-scale feature fusion and lightweight neural network design provide effective technical support for addressing these challenges; however, integrated designs that jointly consider scenario characteristics, target properties, and engineering deployment requirements are still lacking. Therefore, this paper proposes the YOLO-Drill model by integrating lightweight feature extraction, customized multi-scale fusion, scene-adaptive enhancement, and loss function optimization to address key technical challenges in underground drilling target detection and provide technical support for intelligent monitoring of drilling operations.

3. Method

3.1. Overall Architecture of YOLO-Drill

To address challenges such as insufficient illumination, complex backgrounds, and limited computational resources in underground drilling target detection tasks, this paper

proposes a lightweight object detection model, termed YOLO-Drill, based on the YOLO framework. The overall architecture of the model is illustrated in Fig. 1.

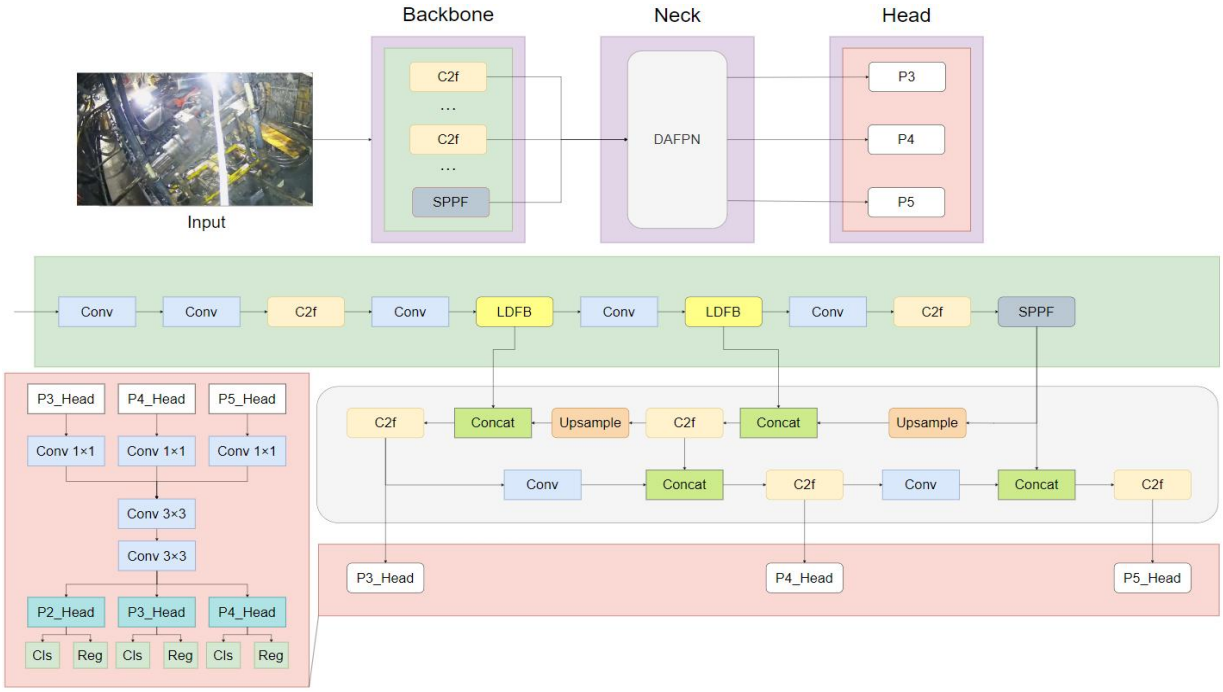

Figure 1. Overall Architecture of the YOLO-Drill Model

The model mainly consists of three components: a backbone network incorporating the lightweight feature extraction module (LDFB), an improved multi-scale feature fusion structure (DAFPN), and a detection head.

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the model first extracts multi-level features through the backbone network, then integrates multi-scale information via the feature fusion network, and finally outputs object categories and location information through the detection head.

3.2. Lightweight Feature Extraction Module (LDFB)

3.2.1. Problem Analysis

Object detection models deployed in underground coal mines are typically required to operate on edge devices with limited computational power and memory, imposing strict constraints on model complexity, parameter size, and inference speed. Traditional YOLO models employ feature extraction modules such as C2 and C3, which rely on standard convolutions and dense connections, resulting in parameter redundancy and computational inefficiency, making them unsuitable for efficient deployment on underground edge devices.

Meanwhile, drilling targets such as drill pipes, drill bits, and drilling rigs are mostly slender structures with strong directional characteristics. Standard convolution operations are insufficient to capture such directional features and are easily affected by noise such as dust and shadows. To address this issue, this paper proposes a lightweight feature extraction module for slender targets, termed Lightweight Drill Feature Block (LDFB), which integrates directional awareness mechanisms into depthwise separable convolution to achieve both lightweight computation and task-specific feature extraction.

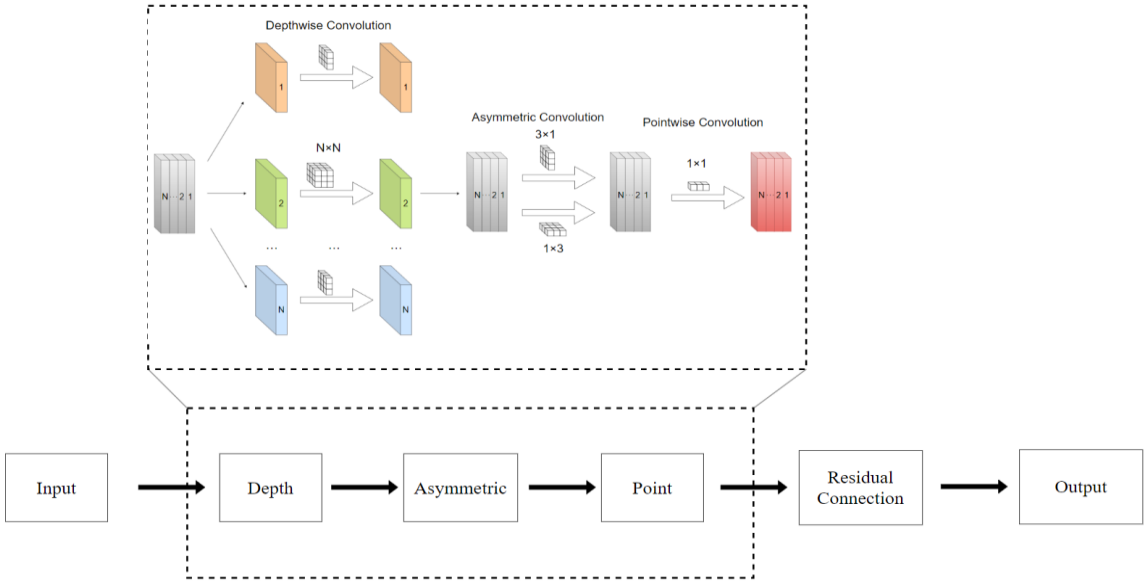

Figure 2. Structure of the Lightweight Drill Feature Block (LDFB)

3.2.2. Module Structure

As shown in Fig. 2, the LDFB module is built upon depthwise separable convolution and incorporates asymmetric convolution kernels and residual connections to significantly reduce computational cost while enhancing the extraction of edge and contour features of slender drilling targets.

The computational complexity of standard convolution is:

$$C_{std} = H \times W \times K^2 \times C_{in} \times C_{out} \quad (1)$$

The computational complexity of depthwise separable convolution is:

$$C_{dw} = H \times W \times K^2 \times C_{in} + H \times W \times C_{in} \times C_{out} \quad (2)$$

Where H and W denote the feature map size, K is the kernel size, and C_{in}, C_{out} are the input and output channels, respectively. Under the same configuration, the computational cost of LDFB is only about 1/8–1/10 of standard convolution.

Residual connections are introduced to prevent gradient vanishing in deep networks, ensuring stable feature extraction under lightweight design.

3.2.3. Module Advantages

First, the module achieves extreme lightweight design by significantly reducing parameters and computational complexity, making it suitable for deployment on underground edge devices.

Second, LDFB enhances target perception capability by incorporating directional convolution structures tailored for slender targets, effectively extracting edge and contour features while suppressing background noise such as dust and shadows.

Third, the module ensures high inference efficiency without introducing complex operators or additional branches.

Finally, residual connections improve feature stability, allowing robust feature extraction even under low illumination and high dust conditions.

3.3. Improved Multi-scale Feature Fusion Structure (DAFPN)

3.3.1. Problem Analysis

Underground drilling targets exhibit large scale variations, frequent occlusion, and dense distribution. Traditional FPN only performs top-down unidirectional fusion and lacks bidirectional interaction, resulting in insufficient feature representation for small and occluded targets.

To address this, a Drill-Aware Feature Pyramid Network (DAFPN) is proposed, which incorporates bidirectional cross-layer connections and lightweight channel attention mechanisms to achieve efficient multi-scale fusion and enhanced target feature representation.

3.3.2. Module Structure

As illustrated in Fig. 3, the DAFPN adopts a bidirectional feature fusion architecture, which enables feature interaction and enhancement across three scales: P3, P4, and P5.

In the top-down pathway, the feature map at the P5 level is first upsampled to a higher resolution and then concatenated with the feature maps from P3 and P4. This process transfers rich semantic information from deeper layers to shallower layers, thereby enhancing the representation of fine-grained details.

In the bottom-up pathway, the initially fused P3 features are downsampled via convolution to align with the spatial resolutions of P4 and P5, followed by further fusion with these higher-level features. This mechanism facilitates the feedback of detailed shallow-layer information to deeper layers, strengthening multi-scale semantic representation.

Through this bidirectional interaction, DAFPN enables sufficient information flow and complementarity across feature levels, effectively improving the detection performance for objects with significant scale variations.

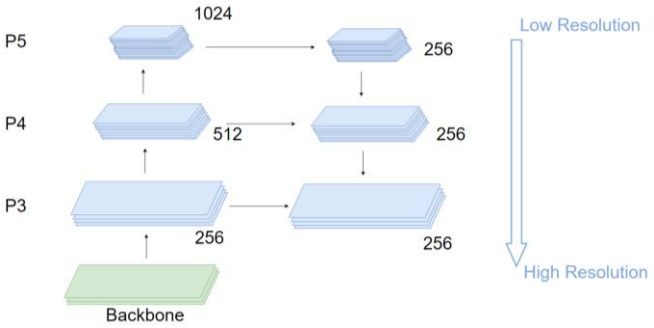

Figure 3. Structure of the Drill-Aware Feature Pyramid Network (DAFPN)

Through the above bidirectional fusion, P3 and P4 achieve comprehensive interaction between shallow details and deep semantics, effectively alleviating the issue of large scale variation in underground drill rods.

Meanwhile, a SE channel attention (lightweight Squeeze-and-Excitation) mechanism is introduced at feature fusion nodes, which adaptively learns to emphasize channels relevant to drilling targets while suppressing interference caused by dust, shadows, and other background noise.

The feature fusion process can be formulated as:

$$F_{out} = \sum_{i=1}^{N} w_i \cdot F_i \cdot A(F_i) \quad (3)$$

Where F_i denotes the input feature at the i-th scale, w_i represents the learnable fusion weight, and $A(F_i)$ is the attention coefficient. The attention mechanism automatically assigns weights to focus on regions containing drilling targets, thereby enhancing robustness under complex conditions.

Compared with traditional FPN, DAFPN expands the receptive field through bidirectional cross-layer connections and improves effective feature weighting via target-aware attention, significantly enhancing the detection performance for small, occluded, and scale-variant targets.

3.3.3. Module Advantages

The DAFPN module offers several advantages:

First, it significantly enhances multi-scale feature representation. The bidirectional fusion design enables sufficient interaction across feature levels, allowing the model to better adapt to large scale variations in underground targets.

Second, the introduced lightweight attention mechanism strengthens target-related features. It adaptively highlights channels associated with drilling objects while suppressing interference from dust, shadows, and other complex background factors, thus improving feature robustness.

Third, DAFPN maintains a well-controlled computational overhead. The design avoids complex structures and high-parameter operators, achieving improved detection performance while preserving model lightweightness, making it suitable for deployment on resource-constrained underground devices.

Additionally, DAFPN improves small object detection performance. By enhancing the utilization of shallow high-resolution features, it preserves fine-grained details of small

targets such as drill bits, reducing missed detections and improving accuracy in complex scenarios.

3.4. Scene-Adaptive Optimization Strategy

3.4.1. Problem Analysis

The unique working environment in coal mines poses severe challenges to object detection, primarily reflected in four aspects: complex lighting conditions, as underground operations rely entirely on artificial lighting with uneven distribution, often resulting in large shadows and local overexposure that severely interfere with object details; severe dust interference, where coal dust generated during drilling causes image blurring and fogging, leading to the loss of object edges and texture information; low image contrast, as the coal–rock background closely resembles the color of drill pipes, making it difficult to distinguish between the target and the background; and significant noise interference, where equipment vibration and sensor contamination introduce image noise, further degrading image quality. These issues directly affect the feature extraction capability of models, leading to false detections and missed detections of targets. Therefore, it is necessary to design targeted scene-adaptive optimization strategies to improve detection performance in complex underground environments.

3.4.2. Data Augmentation Strategy

In response to the harsh visual environment in underground coal mines, such as uneven lighting, dust interference, low contrast, and noise pollution, a scene-adaptive data augmentation strategy is introduced during the model training phase to improve the model's robustness under real-world conditions.

The strategy first employs a low-light adaptive enhancement method, which combines Gamma nonlinear correction and dynamic brightness adjustment to perform adaptive image brightness processing. The transformation formula is as follows:

$$I_{out} = I_{in}^{\gamma} \tag{4}$$

Where $\gamma \in [0.5, 2.0]$. This method effectively brightens the dark regions underground while suppressing the over-saturation of the bright areas, addressing the issue of uneven lighting distribution.

To simulate the dusty underground environment, the strategy introduces dust simulation augmentation. Using mathematical morphology, a dust mask with random concentrations is generated and superimposed onto the training images with a certain probability. This simulates the visual degradation effects under different dust concentrations, enhancing the model's adaptability to foggy and blurred images.

For low contrast between the target and background, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to enhance the contrast of local image regions, emphasizing the edges and contours of the drill pipe and improving the distinction between the target and coal rock background.

Additionally, random noise perturbation is introduced by adding mild Gaussian noise and salt-and-pepper noise to simulate image degradation caused by underground equipment vibration or contamination. This further improves the model's generalization ability in complex environments.

These data augmentation methods are only applied during the training phase and do not add computational load during inference, thus maintaining detection efficiency while enhancing the model's performance under the harsh underground environment.

3.4.3. Strategy Advantages

The core advantage of the proposed scene-adaptive optimization strategy lies in its adaptation to the specific conditions of underground coal mines, addressing the practical needs of drilling target detection. It fills the gap in general data augmentation methods, which are poorly suited for underground environments. This strategy does not introduce additional model computation overhead and does not affect inference speed, while effectively solving the issues of blurred detection and missed detections due to low light and high dust. It complements the previously proposed lightweight feature extraction module and further enhances the model's ability to capture the features of slender targets (such as the drill pipe) in complex conditions. The model remains efficient and avoids redundant computation, aligning perfectly with the deployment requirements of edge devices in underground environments, while adhering to the overall lightweight design philosophy. This provides reliable technical support for future performance optimization.

3.5. Loss Function Optimization

To balance the stability of model training and the accuracy of drilling target localization, the YOLO-Drill model employs a loss function that combines classification loss, confidence loss, and bounding box regression loss. The overall form is:

$$L = L_{cls} + L_{obj} + L_{bbox} \tag{5}$$

The classification loss L_{cls} uses cross-entropy loss to distinguish drilling targets from the background, improving the model's accuracy in identifying target classes. The confidence loss L_{obj} also uses binary cross-entropy loss and is responsible for determining the presence of a target in the image, preventing the generation of unnecessary detection boxes in complex underground backgrounds.

Due to the slender shape and high localization requirements of the drill pipe, the bounding box regression loss L_{bbox} does not use the traditional smooth L1 loss but instead employs CIoU (Complete Intersection over Union) loss. CIoU provides more comprehensive constraints in terms of bounding box overlap, center distance, and aspect ratio, enabling more precise localization of slender targets such as the drill pipe and drill head while accelerating model convergence.

The simplicity of this loss function structure and the use of CIoU optimization ensure stable training while significantly improving target localization accuracy. This design aligns with the model's lightweight approach, ensuring easy convergence during training without increasing the inference burden, making it more suitable for practical deployment in underground coal mines.

4. Experimental Results and Analysis

4.1. Experimental Environment and Parameter Configuration

The experiments were conducted on an Ubuntu 24.04LTS operating system, with a 40-core Intel Xeon Gold 5218R CPU, 192GB RAM, and dual NVIDIA RTX A6000 GPUs (48GB memory per GPU). The experiments were based on Python 3.8.2 and the PyTorch deep learning framework. The batch size was set to 64, the image input size was 640×640, and the

initial learning rate was 0.01 for 300 training epochs. To ensure training efficiency and stability, the experiments ran in an isolated Conda virtual environment named YOLO-Drill.

4.2. Dataset Construction and Division

The experimental data was sourced from underground coal mine drilling monitoring videos, from which image datasets named Drill Scene Dataset (DSD) were created by frame extraction. The DSD construction process followed the principles of diversity, representativeness, and independence, with systematic design during sample collection, cleaning, set division, and target annotation to form a high-quality, generalizable dataset for experiments.

During the initial data collection phase, monitoring videos from actual coal mine deployments were used, covering various coal mines, tunnels, work shifts, and lighting conditions. The video material contained multiple complex conditions, including uneven lighting, dust interference, equipment occlusion, and cluttered backgrounds, ensuring sufficient visual diversity across environments.

In the image sample extraction and cleaning phase, Python scripts were used to sample frames at intervals of 20 to 30 frames, balancing sample diversity and avoiding excessive redundancy. Structural similarity index (SSIM) was applied to remove redundant frames, reducing interference from near-identical samples. Key action nodes from each drilling process (e.g., drilling start, drill pipe advancement, hole completion, drill pipe withdrawal) were selected to ensure comprehensive coverage of important states. After processing, the DSD dataset contained 14 typical drilling operation scenes, totaling 3495 high-quality image samples.

For dataset division, to better reflect the model's generalization ability to unseen scenarios, the dataset was divided without overlap of images across training, validation, and test sets. The training set included 10 scenes, the validation set contained 2 scenes, and the test set included 2 scenes, ensuring spatial layout, equipment configuration, and operational phases were distinct between sets, thus enhancing the objectivity and credibility of evaluation results.

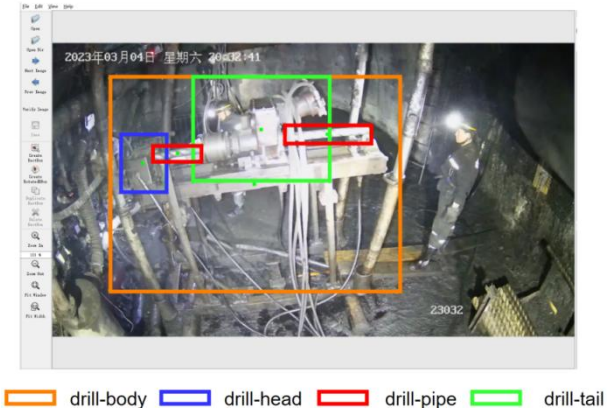

Figure 4. DSD Annotation Examples and Label Categories

For image annotation, the open-source annotation tool LabelImg was used to manually label all images in the dataset. Four target categories were defined: drill head (Drill_head), drill tail (Drill_tail), drill pipe (Drill_pipe), and drill body (Drill_body), covering the main components of drilling equipment. The annotations were stored in XML format, including target category and bounding box coordinates, providing precise supervision signals for model training. The DSD dataset included 2657 training images, 483 validation images, and 355 test images, with some sample images shown in Figure 4.

4.3. Evaluation Metrics

To comprehensively and objectively evaluate the model's detection accuracy and practical applicability, the following two categories of core evaluation metrics were selected:

4.3.1. Detection Accuracy Metrics

Precision: Measures the proportion of true positive samples among all positive predictions. The formula is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{6}$$

Where TP (True Positive) is the number of correctly detected targets, and FP (False Positive) is the number of misdetected non-targets.

Recall: Measures the proportion of true positive samples among all actual positive samples. The formula is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{7}$$

Where FN (False Negative) is the number of missed targets.

mAP@0.5: Average precision (AP) for all categories at an intersection over union (IoU) threshold of 0.5, a core accuracy metric in object detection.

mAP@0.5:0.95: The mean average precision from IoU thresholds of 0.5 to 0.95, reported at steps of 0.05, providing a more comprehensive performance measure across different localization precision requirements.

F1-Score: The harmonic mean of precision and recall, providing a balanced measure of the model's accuracy, calculated as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

4.3.2. Model Complexity Metrics

Parameters: Measured in millions (M), reflecting the model's storage overhead.

FLOPs: Measured in gigaflops (G), reflecting the model's computational complexity.

FPS: Measured in frames per second (fps), reflecting the model's real-time detection capability, providing the basis for edge deployment in underground environments.

4.4. Ablation Experiments

4.4.1. Experimental Design

To verify the effectiveness of the three core improvements: the LDFB lightweight feature extraction module, the DAFPN multi-scale fusion module, and the scene-adaptive enhancement strategy, we designed ablation experiments. Starting from the original YOLO baseline model, each module was sequentially added, and the model performance was compared under different combinations. The experiment strictly controlled variables by only adjusting the activation/deactivation status of the target modules, while keeping other hyperparameters and training processes consistent. The results are shown in Table 1.

Table 1. Impact of core module combinations on model performance

YOLO-Drill	+LDFB	+DAFPN	+Enhancement	mAP@0.5/%	mAP@0.5:0.95/%	F1-score	Parameters /M	FLOPs /G	FPS/ frame·s^{-1}
√	-	-	-	78.6	45.2	0.74	7.2	24.5	78.1
√	√	-	-	79.8	46.0	0.75	5.6	18.8	92.2
√	-	√	-	81.5	47.8	0.77	7.5	25.7	75.1
√	-	-	√	80.2	46.5	0.76	7.2	24.5	78.8
√	√	√	-	83.1	49.2	0.80	6.0	20.0	88.6
√	√	-	√	81.0	47.3	0.79	5.6	18.8	91.0
√	-	√	√	83.8	49.8	0.81	7.5	25.7	75.3
√	√	√	√	85.2	51.0	0.83	6.0	19.9	87.9

4.4.2. Ablation Experiment Results and Analysis

From the ablation experiment results in Table 1, it can be seen that each of the proposed improvements leads to varying degrees of performance enhancement, verifying their effectiveness.

First, after introducing the LDFB lightweight feature extraction module to the baseline model, the mAP@0.5 increases from 78.6% to 79.8%, while the parameter count decreases from 7.2M to 5.6M, FLOPs decrease from 24.5G to 18.8G, and FPS increases from 78.1 to 92.2. This result shows that the LDFB module, through the combination of depthwise separable convolutions and asymmetric convolution kernels, significantly reduces computational overhead while maintaining detection accuracy. This validates the design goal in Section 3.2, where the directional perception structure for elongated targets effectively enhances the edge feature extraction capability of drill rods, drill heads, etc., and the residual connections ensure stable feature extraction, allowing the model to maintain good detection performance even after being lightweight.

Next, after introducing the DAFPN multi-scale fusion module, mAP@0.5 increases to 81.5%, mAP@0.5:0.95 increases to 47.8%, and the F1-score increases to 0.77. This improvement confirms the analysis in Section 3.3: the traditional FPN's one-way fusion is insufficient to handle the significant scale differences of underground targets. DAFPN achieves full interaction between shallow detail features and deep semantic features through bidirectional cross-layer connections. Additionally, the lightweight channel attention mechanism enhances features related to drilling targets and suppresses interference such as dust and shadows, thus improving the detection of small-scale drill heads and occluded drill rods. Although this module slightly increases the number of parameters and computational load, the accuracy improvement it brings is significant for detection in complex underground scenarios.

Furthermore, after using the scene-adaptive enhancement strategy alone, mAP@0.5 increases to 80.2%, with no change in parameters or computational load. This verifies the discussion in Section 3.4: through low-light adaptive enhancement, dust simulation, CLAHE contrast enhancement, and random noise perturbation, the model's ability to adapt to harsh underground conditions with low light, high dust, and low contrast is significantly improved. Since this enhancement strategy is only applied during the training phase, it does not add any burden during inference, achieving a robust performance improvement at no cost.

Finally, when all three modules are used together, the model performance further improves. The combination of LDFB and DAFPN achieves 83.1% mAP@0.5, higher than the cumulative effect of using each module individually. This indicates that the lightweight feature extraction and multi-scale feature fusion form a good complement. The complete model (LDFB + DAFPN + Enhancement) achieves the best performance: mAP@0.5 reaches 85.2%, a 6.6% improvement over the baseline; mAP@0.5:0.95 reaches 51.0%, an improvement of 5.8%; and F1-score reaches 0.83. At the same time, the complete model's parameter count is only 6.0M, FLOPs are only 19.9G, and FPS is 87.9, achieving the best balance between accuracy and efficiency. This fully verifies the effectiveness and engineering practicality of the proposed improvement strategies in the coal mine drilling target detection task.

4.5. Comparison Experiment

4.5.1. Comparison Methods

To fully validate the effectiveness and superiority of the proposed YOLO-Drill model in the coal mine underground drilling target detection task, and to avoid bias from single architecture models, several mainstream models with different architectures were selected as comparison baselines to ensure the comprehensiveness, objectivity, and representativeness of the comparison experiment. The selected comparison models include classic two-stage detection models such as Faster R-CNN, which has a wide application and reference value in the field of object detection; single-stage anchor box models such as SSD, YOLOv8s, YOLOv10s, YOLOv11s, and YOLOv12s, which have strong real-time characteristics, and the YOLO series models have been continuously optimized with each version iteration, reflecting the current development of single-stage models; anchor-free models like CenterNet, which eliminates traditional anchor boxes and has certain advantages in small object detection; and real-time detection models like RT-DETR, which improves real-time detection efficiency through an end-to-end detection approach, making it suitable for engineering deployment scenarios.

To ensure fairness in all comparison experiments and avoid performance deviation caused by different training strategies, hyperparameter settings, or evaluation standards, all comparison models were retrained on the Drill Scene Dataset (DSD) used in this paper. The training process followed the exact same hyperparameter configuration and evaluation metrics as the YOLO-Drill model, including image input size, batch size, learning rate, training epochs, and key hyperparameters, as well as precision, recall, mAP@0.5, F1 score, and other evaluation metrics. The performance comparison results of all models are recorded in Table 2, and to visually show the differences in the core accuracy indicators of each model, the mAP@0.5 values and F1 scores of all comparison models are presented in a line chart (Fig. 5), clearly illustrating the accuracy performance of different

models in the drilling target detection task and providing a reliable basis for subsequent performance analysis.

Table 2. Comparison of model performance

Models	mAP@0.5/%	mAP@0.5:0.95/%	F1-score	Parameters /M	FLOPs/G	FPS/ frame·s^{-1}
Faster R-CNN	75.3	42.1	0.72	41.2	180.1	12.1
SSD [22]	70.8	38.5	0.68	23.1	45.2	35.9
YOLOv3 [23]	76.5	43.0	0.73	61.5	65.3	45.3
CenterNet	77.2	44.5	0.74	32.0	70.1	28.2
YOLOv5s [24]	79.5	46.3	0.76	7.2	16.0	80.1
RT-DETR [25]	82.1	48.7	0.79	20.5	60.2	40.2
YOLOv8s [26]	83.0	49.5	0.80	11.2	28.3	85.8
YOLOv10s [27]	83.5	50.1	0.81	9.5	24.5	90.2
YOLOv11s [28]	83.8	50.4	0.81	9.2	23.9	92.8
YOLOv12s	84.2	50.8	0.82	8.8	22.1	95.7
YOLO-Drill (Ours)	85.2	51.0	0.83	6.0	19.9	87.9

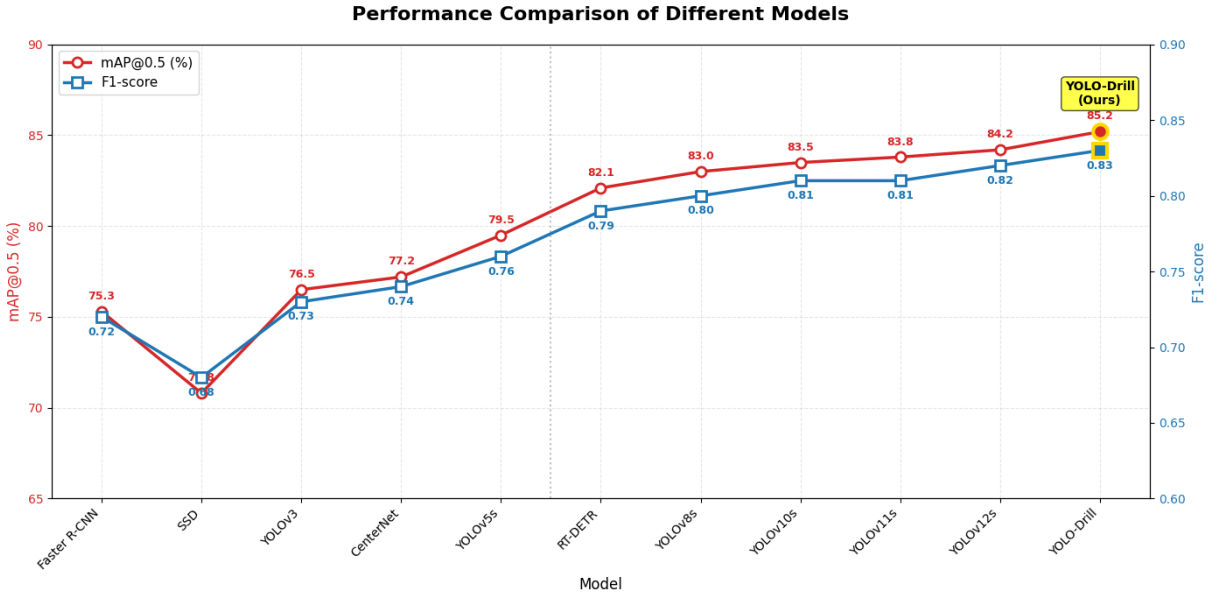

Figure 5. Performance Comparison of Different Detection Models

4.5.2. Experimental Results Analysis

From the comparison experiment results shown in Fig. 5, it can be seen that the proposed YOLO-Drill exhibits significant advantages in both detection accuracy and model efficiency. Specifically, according to Table 2, YOLO-Drill achieves the highest mAP@0.5 of 85.2% and mAP@0.5:0.95 of 51.0% among all the comparison models. Compared to mainstream lightweight models, YOLO-Drill shows improvements of 5.7%/4.7% over YOLOv5s, 2.2%/1.5% over YOLOv8s, 1.4%/0.6% over YOLOv11s, and 1.0%/0.2% over YOLOv12s. This advantage is mainly due to the three core improvements proposed in this paper: the LDFB module, which is designed for directional perception of drill rods and elongated targets, effectively enhances edge contour feature extraction; DAFPN's bidirectional cross-layer fusion and attention mechanism, which improves multi-scale feature expression and small target detection; and the scene-adaptive enhancement strategy, which improves the model's adaptability to harsh underground visual environments. It is worth noting that although YOLOv12s introduces attention mechanisms and performs excellently on general datasets, it is still slightly inferior to the YOLO-Drill model, which was specifically designed for drilling tasks in complex underground scenarios. This indicates that customized design for specific tasks has irreplaceable value.

In comparison with two-stage and anchor-free models, Faster R-CNN, as a classic two-stage model, only achieves an mAP@0.5 of 75.3%, 9.9% lower than the proposed model, with a high parameter count of 41.2M and FLOPs of 180.1G, which makes it difficult to meet the deployment requirements for edge devices in underground settings. CenterNet, as an anchor-free model, achieves 77.2% mAP@0.5, still 8.0% lower than the proposed model. This shows that in detecting elongated targets such as drill rods, an end-to-end single-stage detection architecture combined with targeted module design can achieve a better precision-efficiency balance, while anchor-based designs combined with CIoU loss optimization still have irreplaceable positioning advantages.

In comparison with real-time detection models, RT-DETR, as a representative end-to-end Transformer-based model, achieves 82.1% mAP@0.5, but with a parameter count of 20.5M and an FPS of only 40.2, its efficiency is significantly lower than the proposed model (6.0M/87.9FPS). This suggests that Transformer-based architectures still have a high deployment cost for edge devices in underground environments, while the CNN-based lightweight design in this paper is more practical for engineering.

In terms of model efficiency, YOLO-Drill achieves the

highest accuracy while maintaining a parameter count of only 6.0M, which is lower than all the comparison models. Its FLOPs are only 19.9G, lower than all models except YOLOv5s, and its FPS reaches 87.9, far exceeding the 30FPS threshold required for real-time underground detection. This efficiency advantage comes from the depthwise separable convolution design of the LDFB module and the lightweight attention mechanism in DAFPN, confirming the theoretical analysis in Sections 3.2 and 3.3: by using reasonable structural designs, the model can achieve extreme lightweighting without sacrificing accuracy.

In summary, YOLO-Drill achieves the best balance between detection accuracy and computational efficiency in the coal mine underground drilling target detection task, fully verifying the effectiveness and engineering practicality of the improvement strategies proposed in this paper.

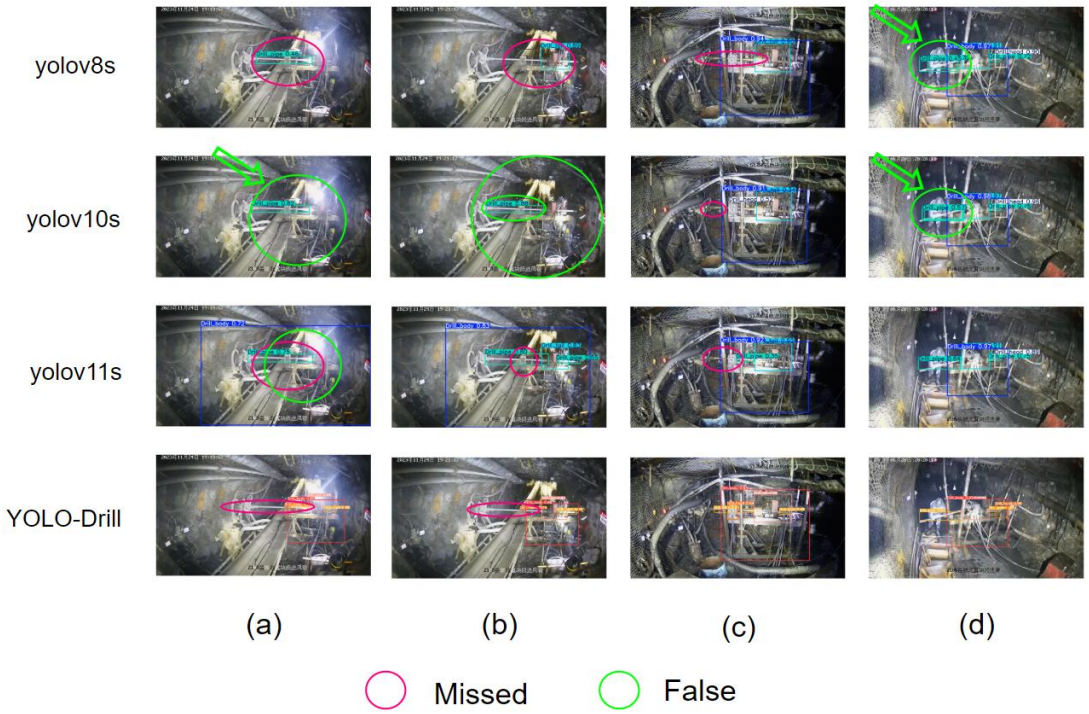

Figure 6. Performance Comparison of Different Detection Models

4.6. Visualization Analysis

Considering factors like real-time performance and detection accuracy, Fig. 6 compares the prediction results of the proposed YOLO-Drill model with representative YOLOv8s, YOLOv10s, and YOLOv11s models in four typical drilling identification tasks under scenarios (a), (b), (c), and (d). Red circles in the figure indicate missed detection targets, while green circles represent false detection targets.

From the images (a), (b), and (d), it can be seen that compared to the YOLO series, YOLO-Drill does not have false negatives. In images (a) and (b), YOLO series models can generally perform normal detection under normal conditions, but in image (a), due to strong light interference, the YOLO series models fail to detect effectively, while YOLO-Drill can still perform normal detection.

It is important to note that YOLO-Drill still experiences some false negatives when detecting small drill rods, which requires further improvement. Overall, YOLO-Drill exhibits better robustness and generalization ability compared to the original YOLO models, demonstrating higher practical value in coal mining extraction and drilling identification tasks.

5. Conclusion

This paper addresses the challenges in coal mine underground drilling target detection, such as harsh environments, large target scale differences, elongated target shapes, and limited computational resources on edge devices. A lightweight target detection model, YOLO-Drill, is proposed based on the YOLO framework, and the model's effectiveness is validated through systematic experiments. The main research conclusions and contributions are as follows:

The lightweight drill rod feature extraction module (LDFB) designed in this paper effectively solves the problem of high model complexity and inadequate feature extraction for elongated targets. The module combines depthwise separable convolutions, asymmetric convolution kernels, and residual connections, reducing model parameters and computational complexity while strengthening the edge and contour feature extraction capability for drill rods, drill heads, and other elongated targets. Ablation experiments show that after adding LDFB, model parameters decrease by 22.2%, FPS increases by 18.1%, and mAP@0.5 improves by 1.2%, fully verifying the effectiveness of the lightweight design and feature extraction capability.

The improved multi-scale feature fusion structure (DAFPN) addresses the problems of large target scale differences and poor detection performance for small and occluded targets in underground environments. This structure achieves full interaction between shallow detail features and deep semantic features through bidirectional cross-layer fusion and uses a lightweight channel attention mechanism to enhance effective target features while suppressing background interference. Experimental results show that adding DAFPN alone increases mAP@0.5 by 2.9%, significantly improving detection performance for small targets like drill heads and occluded drill rods.

The scene-adaptive optimization strategy built in this paper effectively improves the model's robustness in harsh underground environments. This strategy uses low-light adaptive enhancement, dust simulation, contrast enhancement,

and random noise perturbation to improve the model's adaptability to low-light, high-dust, and low-contrast conditions. Ablation experiments show that this strategy improves mAP@0.5 by 1.6% and, as it is applied only during training, does not increase inference load, achieving a cost-free improvement in robustness.

The CIoU-based loss function optimization effectively improves the positioning accuracy of elongated drilling targets. By replacing the traditional smooth L1 loss with CIoU loss, which considers both the overlap of bounding boxes, center point distance, and aspect ratio, the model's positioning capability for drill rods, drill heads, and other elongated targets is enhanced. This also accelerates the model's convergence speed while ensuring training stability, further improving detection accuracy.

Comparison experiments based on the self-built Drill Scene Dataset (DSD) show that YOLO-Drill achieves the optimal balance between detection accuracy and computational efficiency. Compared to mainstream models like YOLOv12s, YOLOv11s, and RT-DETR, YOLO-Drill achieves the highest mAP@0.5 (85.2%) and mAP@0.5:0.95 (51.0%), with only 6.0M parameters and an FPS of 87.9, far surpassing other models and fully meeting the deployment requirements for edge devices in underground environments. Visualization analysis also confirms that YOLO-Drill has stronger robustness in complex scenes with strong light interference and high dust, showing significant practical application value.

However, this study still has some limitations: on the one hand, the model still experiences a small number of false negatives when detecting very small drill rods, mainly due to insufficient utilization of shallow detail features; on the other hand, the scene-adaptive enhancement strategy is applied only during the training phase, and the model's real-time adaptive adjustment ability to sudden environmental changes (e.g., sudden increases in dust concentration) needs to be further improved.

Future research will focus on the following areas: 1) optimizing the feature extraction module to improve the feature extraction capability for extremely small targets and reduce the false negative rate for small drill rods; 2) researching real-time adaptive adjustment mechanisms to dynamically adjust model parameters according to changes in underground environments and further improve environmental adaptability; 3) expanding the dataset to include more complex underground scenarios like water mist interference and device occlusion, thereby improving the model's generalization ability; and 4) conducting real-world deployment tests on edge devices in underground environments to optimize model inference speed and deployment efficiency, promoting the practical application of the model in intelligent coal mine drilling monitoring.

Conflicts of interest

The authors declare no conflicts of interest to report regarding the present study.

Acknowledgements

The authors express their gratitude to Henan Polytechnic University, for administrative and technical support.

References

[1] CHEN Wei, REN Peng, et al. Mine object detection based on spatial attention [J]. Coal Science and Technology, 2022.

[2] Mengran Zhou, Chao Qin. Real-Time Personnel Behavior Detection in Dusty Coal Mines via Dehazing-Enhanced YOLO with Cross-Modal Guidance. Academic Journal of Computing & Information Science (2025), Vol. 8, Issue 11: 62-70.

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 779-788.

[4] Ren S., He K., Girshick R., et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[5] Duan K, Bai S, Xie L, et al. CenterNet: Keypoint Triplets for Object Detection [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 6569-6578.

[6] M. R. Jewel, M. Elmahallawy, S. K. Madria, and S. Frimpong, "DIS-Mine: Instance Segmentation for Disaster-Awareness in Poor-Light Condition in Underground Mines," in 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 1-10.

[7] He Z, Yang J, Ning H, et al. YOLO-DD: lightweight detection in complex environments [J]. EURASIP Journal on Advances in Signal Processing, 2025.

[8] Xu, B., Li, B., Xu, W., Zhu, H., Xu, Y., & Zhao, W. (2024). Research on Lightweight Open-Pit Mine Driving Obstacle Detection Algorithm Based on Improved YOLOv8s. Applied Sciences, *14*(24), 11741.

[9] CHEN Wei, REN Peng, AN Wenni, et al. Mine object detection based on space attention in coal mine edge intelligent surveillance images [J]. Coal Science and Technology, 2022.

[10] ZHOU Mengran, QIN Chao. Real-Time Personnel Behavior Detection in Dusty Coal Mines via Dehazing-Enhanced YOLO with Cross-Modal Guidance [J]. AJCIS, 2025.

[11] CHEN Wei, JIANG Zhicheng, TIAN Zijian, et al. Unsafe action detection algorithm of underground personnel in coal mine based on YOLOv8 [J]. Coal Science and Technology, 2024.

[12] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. Feature Pyramid Networks for Object Detection [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2117-2125.

[13] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. Path Aggregation Network for Instance Segmentation [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 8759-8768.

[14] Tan, M., Pang, R., & Le, Q. V. EfficientDet: Scalable and Efficient Object Detection [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 10781-10790.

[15] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. CBAM: Convolutional Block Attention Module [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[16] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. MobileNetV2: Inverted Residuals and Linear Bottlenecks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 4510-4520.

[17] Ma, N., Zhang, X., Zheng, H. T., & Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design

[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 116-131.

[18] Tan, M., & Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [C]. Proceedings of the 36th International Conference on Machine Learning (ICML), 2019: 6105-6114.

[19] LUO Bingxin, KOU Ziming, HAN Cong, et al. A Faster and Lighter Detection Method for Foreign Objects in Coal Mine Belt Conveyors [J]. Sensors, 2023.

[20] FAN Yingbo, MAO Shanjun, LI Mei, et al. CM-YOLOv8: Lightweight YOLO for Coal Mine Fully Mechanized Mining Face [J]. Sensors, 2024.

[21] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [C]. Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2016: 271-287

[22] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1150

[23] Prasad R, Hussain M. A comprehensive review of YOLOv5 towards real-time detection [J]. Materials Today: Proceedings, 2024, 85: 102-110.

[24] Zhao Y, Lv W, Xu S, et al. DETRs Beat YOLOs on Real-time Object Detection [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[25] Mupparaju S, Thotakura SR, Venkata RR. A review on YOLOv8 and its advancements [J]. Sensors, 2024, 24(11): 4532.

[26] Han J, Zhang Z, Wang X, et al. YOLOv10: Real-Time End-to-End Object Detection. arXiv preprint arXiv:2405.14458, 2024.

[27] Ultralytics. YOLOv11: An Overview of the Key Architectural Enhancements. arXiv preprint arXiv:2410.17725, 2024.

[28] Tian Y, Ye Q, Doermann D. YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv preprint arXiv:2502.12524, 2025.