ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107

Real-Time Personnel Behavior Detection in Dusty Coal Mines via Dehazing-Enhanced YOLO with Cross-Modal Guidance

Mengran Zhou[1,a], Chao Qin[2,b,*]

[1]*School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, China*
[2]*School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China*
[a]*mrzhou8521@163.com*, [b]*1920395276@qq.com*
[*]*Corresponding author*

Abstract: *To address the issues of severe video quality degradation caused by high-concentration coal dust in confined underground coal mine spaces, which leads to difficulties in behavior detection and discriminative feature learning, this study proposes an improved CRR-YOLO algorithm based on YOLOv11n. To tackle the challenge of learning discriminative features, a cross-modal scene-object matching module, CM-SOM, is designed. By introducing a Vision-Language Model (VLM), it establishes cross-modal interaction between visual and linguistic modalities, enhancing the feature space distinction between targets and backgrounds, thereby improving the semantic discrimination capability of the target detection model in scenarios lacking discriminative features. In the backbone network, a context prior-guided feature extraction network, RepVIT, is embedded. It constructs a dynamic contextual information flow through gated dynamic spatial aggregation to enhance the model, achieving dual guidance of features and weights, and strengthening the model's global semantic understanding and contextual dependency modeling of the scene. Furthermore, a feature fusion network with a recalibration mechanism, Re-FPN, is designed. Through a selective boundary aggregation module and a lightweight feature enhancement module, it enables complementary enhancement of boundary details and high-level semantic information via a bidirectional interaction mechanism, optimizing multi-scale feature fusion. Experiments on the dedicated underground coal mine behavior dataset DsLMF+ demonstrate that CRR-YOLO achieves 84.3% mAP@0.5 and 79.1% F1-score, outperforming several advanced models. With only 2.4M parameters and 6.2 GFLOPs, it achieves an inference speed of 253 FPS, striking a favorable balance among accuracy, speed, and complexity, and exhibits strong potential for practical application.*

Keywords: *Behavior Detection, Real-Time Monitoring, Cross-Modal Guidance, Yolov11n*

1. Introduction

With the expansion of coal mining areas and the deepening of mining operations, the increasing danger of the working environment poses threats to personnel life safety and health [1]. The underground coal mine environment is renowned for its extreme complexity: confined spatial structures, severely uneven lighting, and particularly the high-concentration coal dust generated by operations such as excavation and transportation, which constitutes one of the most severe challenges for underground behavior detection. This dust environment severely attenuates light propagation, resulting in low image contrast and blurred details, causing a sharp decline in the performance of traditional computer vision algorithms. Statistics indicate that the root cause of the vast majority of underground safety accidents is directly related to unsafe human behaviors (such as climbing, falling, and leaning). Therefore, achieving accurate personnel behavior detection under low imaging quality conditions has become a core technical bottleneck for improving coal mine safety levels and realizing proactive risk warning.

Regarding behavior detection, researchers have proposed various methods in recent years to improve the accuracy and robustness of underground personnel behavior detection. For instance, Wang Yu et al. [2] proposed a behavior recognition method based on multi-modal feature fusion, achieving high recognition accuracy on public datasets by fusing RGB modality and skeleton modality features, effectively utilizing the complementarity of appearance and motion information. Luo Jinjin et al. [3] addressed real-time requirements by introducing efficient multi-scale convolution and an optimized loss

ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107

function based on YOLOv8n, designing the YOLOv8-ECW model, which improved detection accuracy while maintaining speed. Chen Wei et al. [4] improved YOLOv8s by introducing an efficient multi-scale attention mechanism and lightweight modules, enhancing the ability to extract multi-pose, multi-scale features of miners and reducing model complexity.

However, when dealing with the core challenge of high-concentration, non-uniform dust and fog underground, existing methods still exhibit significant adaptability shortcomings, mainly manifested in three aspects: Firstly, regarding image dehazing, most existing behavior detection models lack built-in effective image preprocessing or dehazing modules, directly extracting features from degraded images, leading to poor model robustness against dust interference. Secondly, regarding feature fusion, whether multi-modal or cross-scale fusion, there is a lack of recalibration mechanisms for features degraded by dust and fog, failing to effectively distinguish and enhance useful information while suppressing dust interference. Finally, regarding real-time detection, while complex models achieve some accuracy improvement, they struggle to meet the stringent real-time requirements for high-frame-rate processing in underground monitoring, whereas lightweight models show insufficient ability to maintain accuracy under dust conditions.

Addressing the aforementioned issues, this paper proposes the CRR-YOLO model, integrating image dehazing and feature recalibration, aiming to achieve high-precision and high-efficiency personnel behavior detection in dust and fog scenarios.

2. YOLOv11 Algorithm

YOLOv11[5] is the latest version in the YOLO series, making important improvements to the backbone network, neck network, and detection head based on YOLOv8, and has become one of the best-performing models in the current object detection field. Precisely because of its outstanding performance, we chose it as the starting point for this research.

Overall, YOLOv11 follows a three-stage architecture: the backbone network extracts image features, the neck network enhances these features, and the detection head performs the final prediction task. C3K2 adopts a structure similar to the C2f module in YOLOv8 in shallow networks, improving information flow efficiency by splitting feature maps and using multiple small convolution kernels. C2PSA introduces a multi-head attention mechanism, enabling the network to focus more intelligently on key areas of the image. Furthermore, YOLOv11 retains the SPPF module, continuing to leverage the advantages of multi-scale feature fusion. In the detection head part, YOLOv11 incorporates depthwise separable convolution (DWConv [6]), reducing both the parameter count and computational load, making the model more lightweight.

However, when the YOLOv11 model is directly applied to the specific scenario of underground coal mines, it reveals several limitations under dust interference, directly affecting the detection accuracy of personnel behaviors: Feature extraction failure: Dust and fog cause image blurring and contrast reduction, while the limited receptive field and global modeling capability of YOLOv11's C3K2 module make it difficult to extract effective features, leading to missed detections and false positives. Localization accuracy deficiency: Dust and fog blur personnel contours, and the FPN structure of YOLOv11 is insufficiently sensitive to degraded edges; its unidirectional fusion path struggles to restore fine boundaries, affecting localization accuracy. Computational inefficiency: Attention modules like C2PSA are effective on clear images but are easily distracted by background noise in low-quality dusty images, wasting computational resources and hindering real-time inference speed.

To address these problems, the CRR-YOLO model makes systematic improvements: Introducing the CM-SOM dehazing module: Through dual-frequency domain fusion and residual channel prior, it achieves global suppression of dust and local detail restoration, improving input image quality. Reconstructing the RepVIT feature extraction network: Establishes a dynamic contextual information flow, enhancing the model's global semantic understanding capability in blurred scenes. Designing the Re-FPN feature fusion network: Employs a bidirectional interaction and feature enhancement mechanism to achieve complementary fusion of cross-level features, improving boundary localization accuracy. CRR-YOLO performs systematic optimization from image preprocessing and feature extraction to feature fusion. Its structure is shown in Figure 1, effectively addressing detection challenges in dust and fog scenarios.

ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107

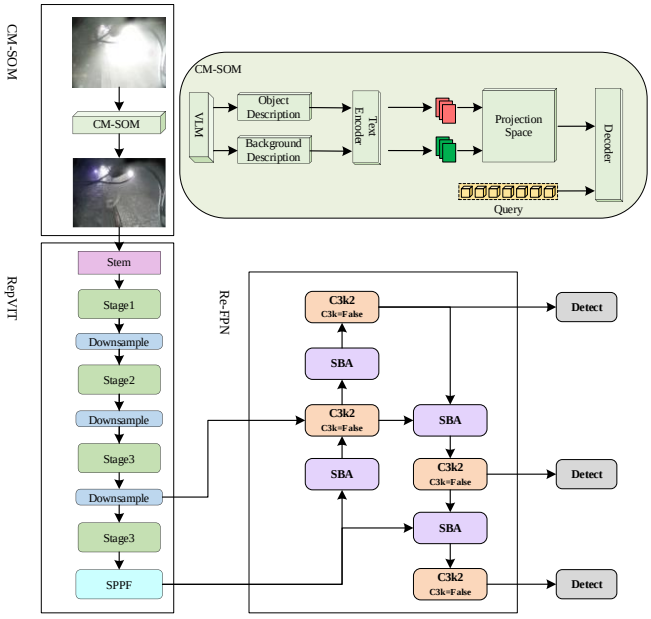

Figure 1: CRR-YOLO network structure.

3. Method

3.1. Cross-Modal Scene-Object Matching (CM-SOM)

To break the semantic similarity between camouflaged objects and the background, relying solely on image features often provides insufficient semantic information to separate them. To address this, Cross-Modal Scene-Object Matching is proposed. It leverages text features generated by a pre-trained Vision-Language Model (VLM) [7] to introduce richer, deeper semantic information, thereby overcoming visual limitations. The specific operations are as follows:

To capture the spatial dependencies of image features, the feature map $\bar{M}$ is transformed into a feature sequence $S = \{s_1, s_2, \cdots, s_N\}$, where $N = h \times w$ and each $s_i \in R^c$ represents a local image feature. Spatial information is injected using positional encodings $P = \{p_1, p_2, \cdots, p_N\}$, resulting in a position-aware feature sequence. This sequence is then fed into an encoder EN(·) for feature enhancement, yielding the enhanced feature sequence $\bar{S}$:

$$\bar{S} = \mathrm{EN}(S + P) \tag{1}$$

Simultaneously, for the camouflage image I, it is input into a VLM to obtain corresponding textual descriptions for the target and background:

$$T_{bg}, T_{ob} = \mathrm{VLM}(I) \tag{2}$$

The VLM employs BLIP-2 (Bootstrapping Language-Image Pre-Training) [8] to generate semantic descriptions for the background and object within the image. For the background description, the image is paired with the query prompt T_{qb}: "What is the environment in the picture?". For the target description, the specific prompt T_{qo}: "What is the camouflaged object in the picture?" is used to guide the model to focus on potential target regions. The VLM process is:

$$Z_{bg}, Z_{ob} = W_p \cdot Self(\mathrm{Cross}(Q_s, \mathrm{Vit}(I)) || \mathrm{BERT}(T_{qb}, T_{qo})) \tag{3}$$

The image *I* is first processed by a frozen ViT encoder to extract visual features, which interact with learnable queries Q_s via a cross-attention mechanism. The resulting visual features are then concatenated with the encoded query prompts from BERT [9], followed by further fusion through a self-attention mechanism. This fused feature is subsequently multiplied by the linear projection weight W_p to obtain the joint features Z_{bg} and Z_{ob}, respectively. These joint features are then fed into the *L*-layer decoder of the language model. For clarity, only Z_{bg} is described in detail below as the data input:

ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107

$$\begin{Bmatrix} A_{t-1} = CATT(Z_{t-1}^{l-1}, Z_{bg} || Z_{1:t-1}^{l-1}) \\ C_{t-1} = Cross(Z_{t-1}^{l-1}, Z_{bg}) \\ Z_{t-1}^{l} = FFN(A_{t-1,} C_{t-1}) \\ z_{t-1} = Z_{t-1}^{L} \end{Bmatrix} \tag{4}$$

In Eq. (4), CATT( ) denotes causal self-attention, FFN( ) is the feed-forward network, $Z_{1:t-1}^{l-1}$ represents historical hidden states, Z_{t-1}^{L} is the hidden state of the last layer, and t is the time step. The probability P over all words in the vocabulary is calculated from z_{t-1}.

$$P(\omega_t | \omega_{1:t-1}) = \text{Softmax}(W_z z_{t-1} + b_z) \in R^{|D|} \tag{5}$$

In Eq. (5), W_z is an output weight matrix that maps the hidden state to the vocabulary space, b_z is a bias term, and $|D|$ is the vocabulary size. The probabilities of the top-k words in P are normalized, and the generated word w_t is selected based on the result. This process is repeated until an end token is generated, yielding the final background textual description T_{bg}. T_{ob} is obtained through an identical process using Z_{ob}.

Textual features for the background E_{bg} and the target E_{ob} are generated via a text encoder TE( ), i.e.:

$$E_{bg}, E_{ob} = \text{TE}(T_{bg}, T_{ob}) \tag{6}$$

To reduce inter-modal discrepancies, the enhanced feature sequence $\bar{S}$ and the textual features E_{bg}, E_{ob} are projected into a common space. Let the projection function be P( ), then the projected features are denoted as:

$$\bar{S}, \overline{E_{bg}}, \overline{E_{ob}} = P(\bar{S}, E_{bg}, E_{ob}) \tag{7}$$

The projected image feature sequence is then matched with the background and target textual features respectively via a cross-attention mechanism, producing background-enhanced features $\overline{S_{bg}}$ and target-enhanced features $\overline{S_{ob}}$. These subsequently interact with $\bar{S}$ to yield the final cross-modal feature sequence:

$$\bar{S} = \bar{S} - \text{Softmax}\left(\frac{(Q_E \overline{E_{bg}})(K_S \bar{S})^T}{\sqrt{d_k}}\right) V_S \bar{S} + \text{Softmax}\left(\frac{(Q_E \overline{E_{ob}})(K_S \bar{S})^T}{\sqrt{d_k}}\right) V_S \bar{S} \tag{8}$$

Finally, $\bar{S}$ is fed into a set of learnable classifiers to obtain the top$N$ elements, constructing a query sequence $\tilde{q} = \{q_1, \cdots, q_n, \cdots, q_N\}$. $\tilde{S}$ and $q\tilde{}$ then interact within a decoder DE( ) to produce the ultimate cross-modal query sequence:

$$\widetilde{S_q} = \text{DE}(\tilde{S}, \tilde{q}) \tag{9}$$

By integrating semantic descriptions of the scene and objects with visual features, this approach better captures the underlying distinctions between target and background, enhancing the separability for anomaly detection. This cross-modal method not only makes anomalous behavioral targets more salient but also redefines the dimensions of behavior detection by incorporating more precise linguistic semantics, thereby effectively overcoming detection bottlenecks caused by visual feature similarities.

3.2. Reparameterized Vision Transformer (RepVIT)

The underground behavior detection task imposes stringent demands on real-time performance. However, the substantial computational footprint of the YOLOv11 model's Backbone can hinder detection speed. The original YOLOv11 backbone is composed of convolutional layers, C3k2 modules, SPPF modules, and CBS modules. Feature extraction through these successive convolutional layers not only requires intensive computation but also inevitably introduces memory redundancy, adversely affecting real-time processing. In contrast, RepVIT employs a unique decoupled convolution design that simplifies the internal network architecture, making it more suitable for tasks with high real-time requirements. Therefore, this paper adopts the RepVIT neural network architecture to address these challenges.

ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107

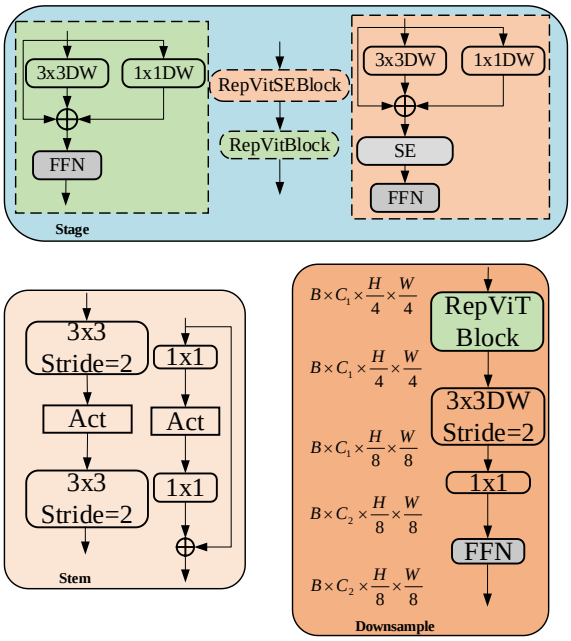

Figure 2: Structure of RepVIT.

As illustrated in Figure 2, the RepVIT structure incorporates both self-attention and multi-head attention mechanisms, which enhances its capability for image recognition tasks. It improves model efficiency and performance while simultaneously reducing the parameter count without compromising predictive accuracy. Although the original YOLOv11 backbone utilizes a series of convolutional and deconvolutional layers alongside residual connections and bottleneck structures, the kernel size of its initial feature extraction convolutional layer is halved compared to its predecessor, resulting in a further reduction of the receptive field. RepVIT, however, leverages re-parameterization techniques to optimize the model structure. This improves parameter operational efficiency and subsequently boosts detection speed. Experimental results confirm that this technique not only preserves the model's predictive power but also reduces its parameter volume. Furthermore, RepVIT utilizes the attention mechanism from Transformers [10] to process image data, enabling it to better capture long-range dependencies within images and thereby compensate for the global information loss associated with the C2f module. RepVIT also exhibits a hierarchical feature extraction characteristic, employing a multi-level structure to process information at different scales within each layer.

3.3. Recalibrate Feature Pyramid Network(Re-FPN)

During the feature fusion stage of behavior detection tasks, the semantic differences between features at different CNN[11] layers are exacerbated by the dust and fog environment, often making effective fusion between deep and shallow features difficult.

The Re-FPN proposed in this paper addresses this issue through its internal Selective Boundary Aggregation (SBA) module.

The SBA module abandons the traditional direct fusion of deep and shallow features, adopting two independent RC modules to enhance shallow and deep features separately before fusion. The structure of the SBA module is shown in Figure 3. This design has dual advantages: on one hand, shallow features can supplement the boundary detail information lost in deep features; on the other hand, high-level semantic information in deep features can enhance the abstract representation capability of shallow features. Then, the outputs of the two RC modules are finally fused via a 3×3 convolution.

The SBA module employs a bidirectional fusion strategy: in the shallow-to-deep fusion path, boundary information $F_b \in \mathrm{R}^{\frac{H}{4}\times\frac{W}{4}\times 32}$ from shallow features is injected into the deep features $F_s \in \mathrm{R}^{\frac{H}{8}\times\frac{W}{8}\times 32}$ to enhance their spatial details; in the other path, high-level semantic information F_s from deep features is fused into the shallow features F_b, serving to suppress noise. Finally, the enhanced features from the two paths are concatenated along the channel dimension and further fused through a 3×3 convolution to improve semantic consistency among features at different levels. This process is formulated as follows:

ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107

$$Z = C_{3\times3}(Concat(PAU(F^s, F^b), PAU(F^b, F^s))) \tag{10}$$

Where $C_{3\times3}(\cdot)$ represents the 3×3 convolution including batch normalization and ReLU activation, $Concat(\cdot)$ represents the concatenation operation along the channel dimension, and the final output is $Z \in R^{\frac{H}{4}\times\frac{W}{4}\times32}$.

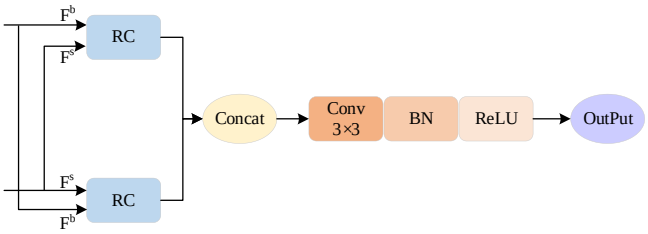

Figure 3: SBA module.

Each RC module internally contains a Pooling Aggregation Unit (PAU) for further reconstruction and refinement of features. The RC module structure is shown in Figure 4.

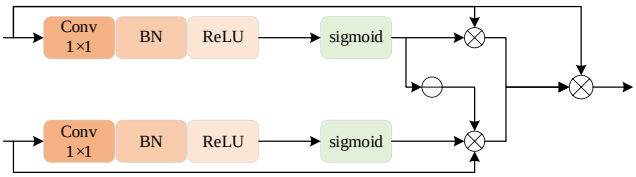

Figure 4: RC module

The calculation process of the PAU module is as follows:

The PAU operation in the SBA module uses learnable weight maps $W_\theta(T_1)$ and $W_\varphi(T_2)$ to adaptively select which semantic information and spatial details from the two paths are most effective for behavior detection in the current dust scene, thereby achieving fusion:

$$T_1' = W_\theta(T_1), T_2' = W_\varphi(T_2) \tag{11}$$

$$PAU(T_1, T_2) = T_1' \odot T_1 + T_2' \odot T_2 \odot \left(\ominus (T_1')\right) + T_1 \tag{12}$$

where T_1 and T_2 are the input features. T_1 and T_2 have their channel numbers compressed to 32 via W_1 and W_2, respectively, obtaining mapped features T_1' and T_2'. The symbol $\odot$ denotes element-wise multiplication, and the operation of subtracting T_1' from T_2' refines the rough initial estimate into a more accurate and complete prediction map. Throughout the process, 1×1 convolution is used as the linear mapping operation.

4. Experimental Analysis

4.1. Dataset and Evaluation Metrics

We evaluated the proposed model on two datasets, including the publicly available DsLMF+ dataset 24 and a self-built Worker Action dataset in Coal Dust Scenes (WAICDS). The DsLMF+ dataset includes a total of 24,709 miner behavior images from 67 different scenes, with 19,767 for training and 4,942 for validation. There are 8 image categories: 5 safe behaviors: operate, sit, stand, stoop, and walk; and 3 unsafe behaviors: climb, fall, lean_against, with a total of 28,289 annotations.

To verify the dust removal effect of the improved model, we propose the WAICDS dataset, using coal mine video surveillance for image collection. We captured 500 images of mine personnel behavior during periods of high coal dust concentration in coal mine operation scenes and selected 700 high dust/fog images from the DsLMF+ dataset, totaling 1,200 images. The dataset contains 6 categories: walk, climb, fall, operate, lean_against, and data augmentation was performed via random rotation and brightness adjustment. The dataset was divided into training and validation sets in an 8:2 ratio. The effectiveness of the model is verified by monitoring mine video surveillance. Training set annotations include class labels, bounding box center coordinates (x, y), and width and height measurements. Examples from the WAICDS dataset are shown in Figure 5.

ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107


Figure 5: Example of WAICDS dataset

To accurately evaluate object detection performance, four fundamental metrics were used: Precision (P), Recall (R), F1-score (F1), and mean Average Precision (mAP) [12, 13].

4.2. Experimental Environment Configuration

Several crucial training environments are involved in this study, including the need for CUDA and PyTorch support. Due to the large number of images in the dataset and the need to improve model training accuracy, YOLOv11 utilizes GPU and CUDA to accelerate computation, while PyTorch is used as the tool and library to build and train the model. The experiment used a Nikon D850 camera to capture some images, with lighting conditions ranging from 5-100 Lux. Dust/fog concentration was monitored via a particle counter, ranging from 50-500 μg/m³. Labeling was completed using the LabelImg tool, with cross-validation ensuring consistency.

4.3. Ablation Studies

To verify the effectiveness of the three proposed improvement strategies, ablation experiments were performed on the baseline model. For consistency and convenience in performance evaluation, these ablation experiments were conducted on the DsLMF+ dataset. Experiment 1 introduced CM-SOM (C) alone. Experiment 2 introduced RepVIT (R) alone. Experiment 3 introduced Re-FPN (F) alone. Experiment 4 added RepVIT to Experiment 1. Experiment 5 added Re-FPN to Experiment 1. Experiment 6 added Re-FPN to Experiment 2. The results are shown in Table 1.

Table 1: Ablation experiment.

Model	*C*	*R*	*F*	*P%*	*R%*	*mAP@0.5/%*	*FPS/ ($f\,s^{-1}$)*	*Parameters/M*	*GFLOPs*
YOLOv11n	-	-	-	78.1	75.7	79.9	188	2.6	6.3
1	√	-	-	77.5	75.2	81.6	180	3.2	6.8
2	-	√	-	81.5	76.0	81.9	195	2.3	5.9
3	-	-	√	81.1	78.7	82.0	200	2.8	6.3
4	√	√	-	83.5	72.6	82.5	231	2.5	6.1
5	√	-	√	73.9	80.8	82.5	228	2.7	6.5
6	-	√	√	81.4	76.2	82.6	244	2.7	6.4
7	√	√	√	85.4	74.6	84.3	253	2.4	6.2

Based on the ablation results, the following conclusions can be drawn: Model 1, introducing the CM-SOM structure alone, improves average precision by 1.7%, but precision, recall, and real-time performance all decrease, indicating that the CM-SOM structure can effectively improve model detection accuracy. However, due to its multi-layer convolutional structure increasing parameters, real-time performance is reduced. Model 2, introducing the RepVIT structure alone, improves precision, recall, and average precision, indicating that RepVIT can effectively bridge the semantic gap between different layers during feature extraction and shallow-deep feature fusion. Model 3, introducing Re-FPN alone, also improves precision, recall, and average precision, especially recall increased by 3%, representing enhanced model adaptability and alleviated missed/false negative rates. In Model 4, the introduction of CM-SOM and RepVIT increased mAP50 from 79.9% to 82.5%, showing their key role in improving model accuracy, especially in complex scenes. Model 5, with the further integration of Re-FPN on top of Model 1, further improved mAP50 to 82.5%. Model 6, based on Model 2, further optimizes detection performance by introducing Re-FPN, maintaining mAP50 at 82.6%, enhancing focus on key features, and improving small target detection. Model 7 is the final model. The fusion of the three innovative schemes yields significant performance gains, with precision, recall, average precision, and real-time performance reaching 85.4%, 74.6%, 84.3%, and 253 FPS respectively. Moreover, parameter volume and complexity are reduced, ultimately achieving a good balance between accuracy and efficiency.

ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107

4.4. Comparative Experiments

To verify the behavior detection capability of the proposed model for coal dust scenes, it was compared with currently popular object detection algorithm models. Under the same experimental environment, using the same configuration and parameters, detection experiments were conducted on the DsLMF+ dataset. Comparative models included YOLOv5s, YOLOv7-tiny, Yolov8n, YOLOv9-tiny, YOLOv10n, RTDETR-18, and the latest coal mine underground miner behavior detection methods proposed by Liu [14] and Ni [15] (represented by Model-1 and Model-2, respectively). Comparative experimental results are shown in Table 2.

Table 2: Performance comparison of different detection algorithms on the DsLMF+ dataset.

Models	*F1-Score%*	*mAP@0.5/%*	*FPS(f s^{-1})*	*Parameters/M*	*FLOPS/G*
YOLOv5s	74.3	77.1	187	7.8	18.8
YOLOv7-tiny	73.4	77.5	190	6.0	13.3
YOLOv8n	73.1	77.2	202	3.1	6.8
YOLOv9s	76.8	82.5	194	61.9	22.1
YOLOv10n	73.6	78.1	200	2.7	6.5
YOLOv11n	76.5	79.9	188	2.6	6.3
RTDETR-18	78.6	82.6	249	20.0	57
Model-1	75.7	79.3	221	3.4	9.8
Model-2	75.0	78.4	211	3.3	8.6
CRR-YOLO	79.1	84.3	253	2.4	6.2

According to the comparison results shown in Table 2, the CRR-YOLO algorithm outperforms other YOLO variants in terms of detection accuracy and computational efficiency. Specifically, CRR-YOLO achieves 84.3% mAP@0.5 and 79.1% F1-Score, surpassing most lightweight models and even some large models. For example, CRR-YOLO's mAP@0.5 significantly exceeds that of YOLOv5s (77.1%), YOLOv7-tiny (77.5%), YOLOv8n (77.2%), and is even higher than YOLOv9s (82.5%). Compared to newer models like YOLOv10n (78.1%), YOLOv11n (79.9%), and RTDETR-18 (82.6%), CRR-YOLO still maintains higher detection accuracy. In terms of F1-Score, it also outperforms the second-highest RTDETR-18 (78.6%) and the third-highest YOLOv9s (76.8%), indicating that CRR-YOLO achieves a good balance between precision and recall and possesses robustness in handling more challenging detection thresholds. In terms of lightweight design, CRR-YOLO also has advantages, requiring only 2.4M parameters and 6.2 GFLOPs, which is substantially lower than the recent model RTDETR-18's 20M parameters and 57 GFLOPs, and also lower than the lightest version in YOLOv11, YOLOv11n (2.6M parameters, 6.3 GFLOPs). Compared with the latest behavior detection models, CRR-YOLO's mAP@0.5 reaches 84.3%, significantly higher than Model-1 (79.3%) and Model-2 (78.4%), representing an improvement of 5-6 percentage points. This highlights the importance of the synergistic design of front-end image enhancement (dehazing) and back-end feature optimization (recalibration). CRR-YOLO leads with an inference speed of 253 FPS, approximately 15-20% faster than Model-1 (221 FPS) and Model-2 (211 FPS), fully meeting the real-time monitoring requirements underground. The customized network structure for specific scenarios shows clear advantages over general models. CRR-YOLO requires only 2.4M parameters and 6.2 GFLOPs, representing approximately 30% fewer parameters and 35% lower computational complexity compared to the comparative models, demonstrating excellent lightweight characteristics. These results indicate the effectiveness of CRR-YOLO in achieving the optimal trade-off between accuracy and efficiency. Its lightweight design, combined with its strong detection capability, especially for small and difficult targets, makes it a promising candidate for real-time deployment in scenarios such as underground coal mines with dust and fog.

5. Conclusion

(1) This paper proposes CRR-YOLO, a personnel behavior detection model for underground dust and fog scenarios. The model achieves an F1-Score of 79.1% and a mean Average Precision of 84.3%, representing a 4.4% improvement over the baseline model. The model has 2.4M parameters and a detection speed of 253 FPS. This model can rapidly and accurately detect the behaviors of underground workers in high coal dust scenes.

(2) An image dehazing network named CM-SOM for underground dust environments is proposed. It achieves efficient global-local synergistic modeling and detail recovery by fusing Fourier convolution and contextual priors. A feature extraction network named RepVIT is designed to enhance the global

ISSN 2616-5775 Vol. 8, Issue 11: 62-70, DOI: 10.25236/AJCIS.2025.081107

semantic perception and contextual understanding capability of the backbone network under blurred detail conditions. Re-FPN is constructed. Through its core Selective Boundary Aggregation and Feature Enhancement modules, the cross-layer feature fusion process is optimized, effectively alleviating the semantic gap and detail loss problems.

(3) Extensive experiments on self-built and public datasets demonstrate that the proposed CRR-YOLO model outperforms current mainstream methods in terms of detection accuracy, robustness, and computational efficiency, providing a reliable technical pathway for the practical deployment of underground intelligent surveillance systems.

Acknowledgements

This research was support by the National Natural Science Foundation of China (Grant No. 52374177), the Anhui Provincial Major Science and Technology Project (Grant No. 201903a07020013) and the University-level general projects of Anhui University of science and technology (Grant No. 2024cx2119).

References

[1] Wang H, Mou L. An Improved YOLOv8 Based Unsafe Behavior Detection Algorithm for Coal Mine Underground Personnel[C]//2025 6th International Conference on Computer Engineering and Application (ICCEA). IEEE, 2025: 01-05.

[2] Yu W, Chunhua Y U, Xiaoqing C, et al. Recognition of unsafe behaviors of underground personnel based on multi modal feature fusion[J]. Journal of Mine Automation, 2023, 49(11): 138-144.

[3] Jinjin L U O, Wei C, Zijian T, et al. Real-time detection algorithm of underground personnel behavior based on YOLOv8-ECW[J]. Journal of Mining Science and Technology, 2025, 10(2): 316-327.

[4] Chen W, Mu H X, Guan Y Y, et al. Improving YOLOv8s for behavior detection of underground miners in coal mine[J].Journal of Liaoning Technical University (Natural Science),2025,44(3):257-264. (in Chinese).

[5] Redmon J, Farhadi A. YOLOv11: A New Generation of Real-Time Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 1-9.

[6] Wang Z, He X, Li Y, et al. EmbedFormer: embedded depth-wise convolution layer for token mixing[J]. Sensors, 2022,22(24):9854.

[7] Xu H, Ghosh G, Huang P Y, et al. VLM: Task-agnostic Video-Language Model Pre-training for Video Understanding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2021.

[8] Ding X, Zhang Y, Ge Y, et al. Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 5513-5524.

[9] Koroteev M V. BERT: A Review of Applications in Natural Language Processing and Understanding[J]. Computation and Language,2021

[10] Berger C, Premaraj N, Ravelli R B G, et al. Cryo-electron tomography on focused ion beam lamellae transforms structural cell biology[J]. Nature Methods, 2023, 20(4): 499-511.

[11] Alzubaidi L, Zhang J, Humaidi A J, Al-Dujaili, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. Journal of big Data,2021, 8(1):53.

[12] Liu W, Quijano K, Crawford M M. YOLOv5-Tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 8085-8094.

[13] Wang C, Sun W, Wu H, et al. A low-altitude remote sensing inspection method on rural living environments based on a modified YOLOv5s-ViT[J]. Remote Sensing, 2022, 14(19): 4784.

[14] Liu N, Huang G, Xu D, et al. Research on Illegal Behavior Detection Algorithm of Underground Mine Workers Based on Improved YOLOv8[C]//2024 17th International Conference on Advanced Computer Theory and Engineering (ICACTE). IEEE, 2024: 164-168.

[15] Ni Y, Huo J, Hou Y, et al. Detection of underground dangerous area based on improving YOLOV8[J]. Electronics, 2024, 13(3): 623.