Research on Engineering Structures & Materials
www.jresm.org
Research Group

Research Article

Coal and gangue recognition and separation process optimization in coal preparation plants based on DeepLabv3+ and intelligent decision-making

Mengmeng Ma [*,1,a], Kaixiong Wang [2,b], Xuke Zhang [1,c], Jianbin Wang [2,d]

[1]*Zhongxuan Zhikong (Tianjin) Technology Co., Ltd., Tianjin, 300091, China*
[2]*Guoneng Zhuneng Group Co., Ltd., Ordos, 010300, China*

Article Info

Article History:

Received 09 June 2026
Accepted 23 July 2026

Keywords:

Coal and gangue recognition;
Intelligent separation;
DeepLabv3+;
YOLOv7;
MobileNetv4

Abstract

With the aim to solve the issues with poor accuracy of separation and delayed parameters adjustment in coal and gangue separation process, this study presents a model for coal and gangue separation that combines the DeepLabv3+ model and intelligent decision making. The model builds upon the DeepLabv3+ framework and incorporates the MobileNetv4 network to decrease parameter count and computational complexity while enhancing inference efficiency. Meanwhile, the Bottleneck Attention Module is introduced to strengthen boundary features between coal and gangue and achieve pixel-level extraction within coal and gangue regions. Moreover, the model integrates YOLOv7 to enhance multi-scale representations of coal and gangue and employs image coordinate mapping technology to enable rapid target localization and separation decision-making. Finally, this study uses a coordinated execution mechanism of a robotic arm and air blowing to achieve intelligent separation. Experimental results demonstrate that the model converges rapidly during training. Accuracy in recognition is 98.85%, while the average precision is stable at 98.64%. The model contains 30.82 M parameters and takes 17.27 ms to respond, both of which outperform the comparison models. Under dry conditions, the pixel accuracy of the three types of coal is 93.05% -94.87%, and the accuracy under high humidity is 90.74% -93.42%. Under the worst high-humidity brown coal conditions, the mAP@0.5 still reaches 91.28%, verifying the adaptability of the model. In the normal operation of the machine, coal precision is 96.83%. Even under high gangue content, coal precision is 93.76%. The above results indicate that this method has excellent coal rock recognition accuracy and sorting ability, and achieves a good balance between model complexity and accuracy, providing a feasible technical path for intelligent sorting optimization in coal preparation plants.

1. Introduction

Coal recognition and separation, as a key process in coal processing and utilization, determines the quality of coal products, resource utilization efficiency, and subsequent clean utilization performance [1-2]. Manual knowledge and fixed-parameter implementation have been the conventional approaches to recognizing and separating coal in coal processing facilities. However, this approach is characterized by high labor intensity, inefficient separation, and instability in accuracy [3-4]. As the mechanization of coal mining continues to rise, the conventional method

*Corresponding author: fxmammm@163.com
[a]orcid.org/0009-0002-6368-0675; [b]orcid.org/0009-0008-0744-7784; [c]orcid.org/0009-0005-9325-2761; [d]orcid.org/0009-0006-6633-9287
DOI: http://dx.doi.org/10.17515/resm2026-1735ce0609rs

becomes difficult to match with modern coal industry needs [5]. Therefore, an efficient and accurate coal-gangue recognition and separation technology is urgently needed.

DeepLabv3+ is a semantic segmentation network based on an encoder-decoder structure and has been applied in many fields [6-7]. For instance, Cheng L et al. employed the DeepLabv3 architecture to facilitate efficient feature extraction while saving on computation to boost the performance of segmentation of navigable zones for unmanned surface vessels. The method uses multi-scale attention atrous spatial pyramid pooling to integrate multi-level features, while subtraction spatial attention was used to emphasize edge information [8]. Chen H et al. proposed a DeepLabv3+ network with MobileNetv2 as the backbone network, employing hybrid atrous convolution in atrous spatial pyramid pooling. They also introduced strip pooling to enhance local segmentation capability and added attention modules to capture small target features, achieving accurate semantic segmentation of remote sensing images [9]. YOLOv7 has excellent feature extraction and inference capabilities with its efficient layer aggregation structure and reparametrized convolution design. The Cao Z team has improved YOLOv7 tiny to achieve lightweight detection and enhance the speed and accuracy of coal and gangue recognition under complex working conditions, addressing issues such as complex on-site models, difficult training, and poor positioning performance in coal gangue sorting. This provides a foundation for online visual sorting [10]. The above research validates the effectiveness of the DeepLabv3+ semantic segmentation network and the YOLOv7 detection algorithm in visual recognition tasks. However, most existing studies employ segmentation or detection networks separately for object recognition, lacking a collaborative fusion architecture between the two.

Over the past few years, extensive studies have investigated intelligent separation techniques for coal-gangue minerals. Shatwell D G et al. developed a method for the intelligent separation of ores through image processing and artificial intelligence techniques to minimize mining expenses. The proposed method consisted of four steps of image segmentation and partitioning, feature extraction, classification using artificial neural networks, and voting technique [11]. Jiang J et al. put forward a recognition approach that integrates negative selection edge extraction with a multilayer perceptron to enhance the accuracy and efficiency of coal-gangue separation. This method utilized negative selection edge extraction to eliminate long-term contamination and background grayscale interference, followed by multilayer perceptron classification. Results indicated that the algorithm possessed accurate recognition performance [12]. He L et al. proposed a coal-gangue image recognition method based on dual-view visible light imaging. This method leveraged the differences in surface texture and reflection characteristics between coal and gangue, evaluated feature separability through linear color dispersion and ranked them, and combined classifier models to distinguish pseudo-middlings from gangue [13]. Luo Q et al. addressed the challenges caused by fine coal particles and moisture adhesion in coal-gangue recognition by analyzing variations in feature distributions under different operating conditions. It was observed that moisture impacted the gradient features while pulverized coal impacted the gray scale texture features. An innovative approach for recognition was developed using the texture pattern distribution of coal and gangue surfaces [14]. The above studies enhanced coal-gangue recognition and separation performance from technical approaches, but relied on handcrafted features or single models with insufficient generalization capability.

To tackle the aforementioned challenges, this study develops a coal-gangue recognition and separation model that integrates an improved DeepLabv3+ with YOLOv7. First, DeepLabv3+ is used as the feature extraction framework, with MobileNetv4 replacing the original network to reduce computational redundancy. The Bottleneck Attention Module (BAM) is introduced to enhance channel and spatial feature responses at coal-gangue boundaries. Using the segmentation results as the computational basis, YOLOv7 outputs the positions of coal-gangue targets. Through image coordinate mapping technology, pixel coordinates are converted to physical coordinates. Finally, intelligent separation is achieved through robotic arm grasping and air blowing. The coal-gangue recognition and separation model is expected to achieve precise segmentation and real-time localization of coal-gangue targets, improving recognition accuracy and separation efficiency, and providing reliable technical support for the intelligent and efficient development of the coal industry. The novelty of this study is the use of MobileNetv4 to substitute DeepLabv3+ as the

backbone architecture that meets edge device implementation needs of achieving real-time speed without sacrificing segmentation precision. Simultaneously, the BAM module is incorporated into the system to ensure that interference by dust, lighting, and other variables can be successfully reduced. YOLOv7 and the coordinate mapping of the image are employed to obtain the positions of the coal-gangue targets. In addition, a robotic arm-air blowing cooperative separation strategy is introduced to achieve dynamic switching between grasping large coal-gangue blocks and air blowing small coal-gangue blocks.

2. Coal and Gangue Recognition And Separation Optimization Model Design

The recognition and segregation accuracy of coal and gangue are significant factors which influence the efficiency and quality of the product produced from the coal preparation plant. The traditional methods of coal and gangue recognition and segregation involve manual observation. However, such techniques are limited by their inefficiency, slow response, and poor accuracy [15-16]. Therefore, for enhancing recognition flexibility, this study introduces a model for optimizing recognition and separation of coal and gangue, termed the Deep-YOLOv7 model. The proposed model enhances the DeepLabv3+ architecture by improving its multi-scale feature extraction capability and pixel-wise recognition precision. On this basis, it integrates the YOLOv7 algorithm to link recognition results with separation execution parameters, achieving intelligent recognition and accurate separation of coal and gangue. The workflow of the Deep-YOLOv7 model is shown in Figure 1.

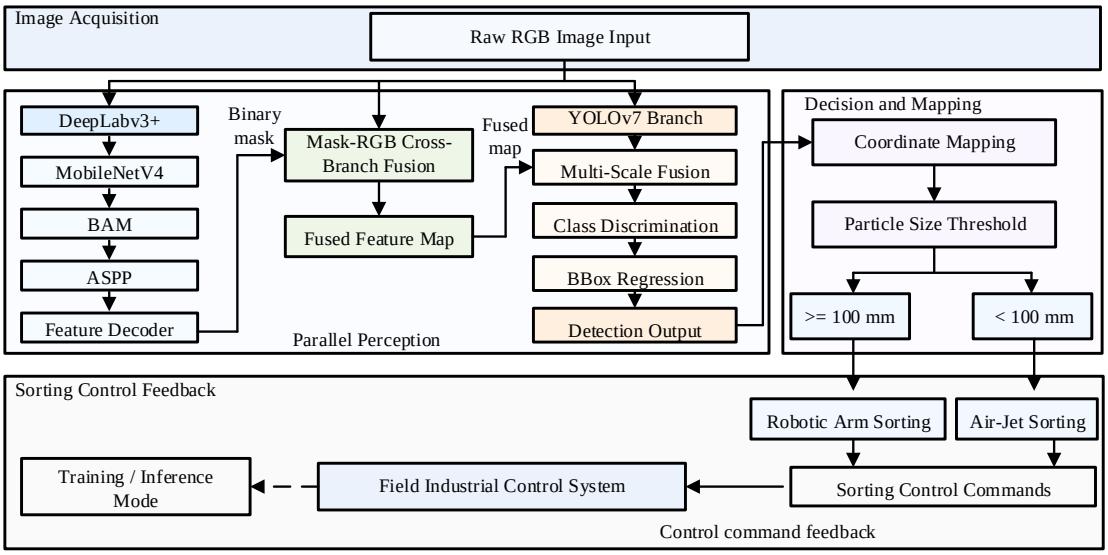


Fig. 1. Deep-YOLOv7 model overall workflow

As illustrated in Figure 1, the semantic segmentation network and object detection network in the Deep-YOLOv7 model adopt a deployment mode of offline independent training and online parallel inference. This model utilizes DeepLabv3+ to perceive coal and rock image inputs; it efficiently extracts coal and rock image features while reducing computational complexity through the MobileNetv4 backbone network. Subsequently, BAM was used to enhance feature representation, and multi-scale coal rock features were captured through Atrous Spatial Pyramid Pooling (ASPP). After the fusion and reconstruction of various levels of features by the decoder, a foreground-background binary mask image is output to achieve pixel-level division between the coal-rock material area and the conveyor belt background, which screens out invalid background areas. In order to provide spatial prior constraints to the YOLOv7 detection module, the binary mask output by the segmentation branch will be fused with the original image in the channel dimension, and the fused composite image will be used as the input data for YOLOv7. Based on the coal rock status, priority sorting is completed, and YOLOv7 is used to achieve rapid detection and classification of coal rock. After the detection is completed, the algorithm normalizes the size of the target bounding box and pixel coordinates, and converts the two-dimensional image coordinates to the actual spatial coordinates of the conveyor belt through pixel physical coordinate mapping, achieving

accurate target positioning. The model compares the material particle size with a 100 mm threshold to match the corresponding actuator. When the particle size is not lower than the threshold, the robotic arm is triggered to grab and sort. If the particle size is lower than the threshold, high-pressure airflow sorting is used. At the same time, real-time data such as material category, spatial location, and sorting execution instructions will be transmitted back to the on-site production control system. The entire model is broken down layer by layer into four core functions with clear boundaries: semantic segmentation to extract effective areas of materials, object detection to identify the categories of coal and gangue, object positioning to solve the coordinates from pixel space to physical space, and physical separation relying on mechanical arms and air blowing mechanisms to separate coal gangue entities.

2.1 Recognition Module Based on Improved Deeplabv3+

Correct detection of coal and gangue is a prerequisite for intelligent segregation, and the accuracy of such detection directly influences the precision of future decisions regarding their separation. The coal processing plant features numerous difficulties related to the similarity in grayscale and low-textured differences between coal and gangue, which leads to the inefficiency of traditional thresholding methods in detecting them [17]. The DeepLabv3+ framework effectively performs multi-scale feature extraction and contextual information modeling and has strong feature representation ability in pixel-level semantic segmentation tasks [18]. Therefore, to improve the accuracy of coal-rock recognition, DeepLabv3+ is adopted as the semantic segmentation backbone to obtain pixel-level coal-rock classification results and provide refined region priors for subsequent object detection and sorting decisions. The architecture is presented in Figure 2.

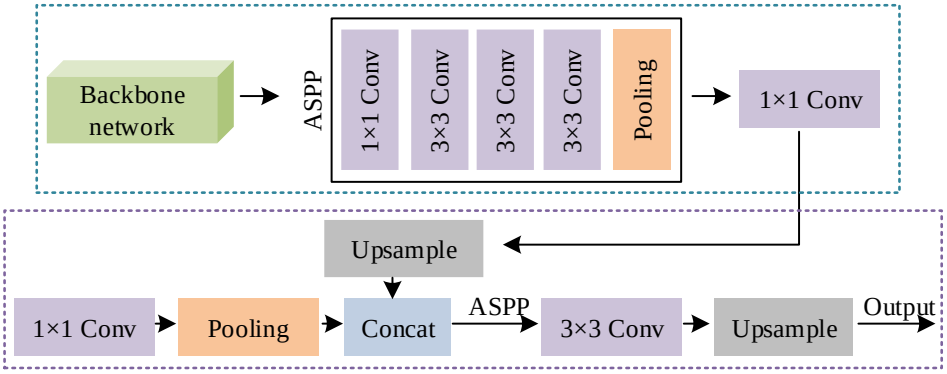


Fig. 2 DeepLabv3+ framework for coal and gangue semantic segmentation

As shown in Figure 2, DeepLabv3+ first inputs coal and gangue images. The backbone network extracts feature and sends high-level features into the ASPP module. ASPP employs parallel convolutions with varying dilation rates to extract multi-scale feature information, thereby addressing the large-scale variation and blurred boundaries of coal and gangue targets. The convolution operation of atrous convolution is shown in Equation (1).

$$y_i = \sum_{k=1}^{K} x_{i+r\cdot k} \cdot w_k \tag{1}$$

In Equation (1), x_i represents the input feature, r represents the dilation rate, and w_k represents the convolution kernel. The dilation rate and convolution stride are independent parameters; increasing the dilation rate only enlarges the receptive field and does not compress the feature resolution. Simply increasing the convolution step size can expand the receptive field, but it will down sample and reduce the resolution of the feature map, resulting in the loss of fine-grained coal gangue edge features, which is not conducive to weak feature extraction. The Equation for the relationship between receptive field size, feature map size, stride value, and kernel size is given by Equation (2).

$$r_n = (r_{n+1} - 1)s_n + k_n \tag{2}$$

In Equation (2), r_n represents the receptive field size of the n layer feature map, s_n represents the convolution stride, and k_n represents the kernel size. Then ASPP fuses each feature extraction branch, and the fusion process is shown in Equation (3).

$$F_{ASPP} = Concat(F_{1\times1}, F_{3\times3}, F_{3\times3}, F_{3\times3}, F_{pool}) \tag{3}$$

The fused features pass through a 1×1 convolution for channel reduction, then up sampling restores resolution, and the features are concatenated with low-level detail features from the backbone network. The decoder fusion process is shown in Equation (4).

$$F_{concat} = Concat(Upsample(F_{ASPP}, 4), F_{low}) \tag{4}$$

In Equation (4), $Upsample$ represents upsampling, F_{ASPP} represents ASPP output features, and F_{low} represents low-level detail features. Finally, the output image size is restored through a second upscaling process to produce the pixel-level classification results for coal and gangue. Although the general structure of the DeepLabv3+ model efficiently captures multiscale context information, the original backbone architecture is computationally expensive and requires many parameters [19]. Therefore, to balance recognition accuracy and computational efficiency, this study replaces the DeepLabv3+ backbone with MobileNetv4. The coal and gangue feature extraction mechanism of MobileNetv4 is shown in Figure 3.

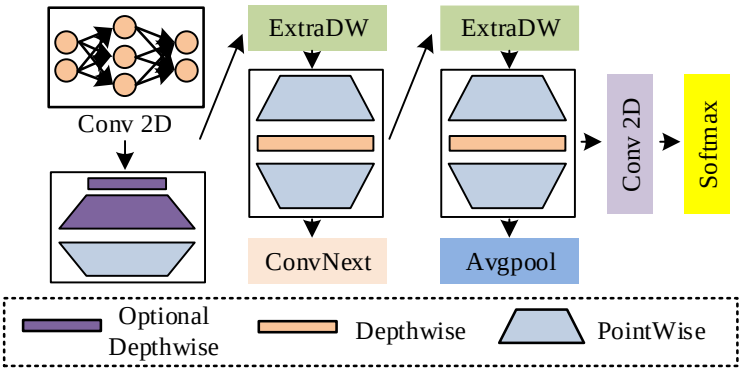


Fig. 3. MobileNetv4-based coal and gangue feature extraction mechanism

As shown in Figure 3, the MobileNetv4 network initially takes the original image of coal and gangue as input and performs feature extraction using a 2D convolutional layer, where channel dimensions are also increased. It then utilizes the inverted bottleneck layer to optimize the process of feature extraction and recognize the textures and edges of coal and gangue. The enhanced features are fed into the depth wise separable convolution module, which splits the standard convolution into a depth wise convolution and a pointwise convolution step. This reduces model parameters and floating-point operations (FLOPs) from the convolution kernel dimension, as shown in Equation (5).

$$F_a = K^2 \times N \times W \times H + M \times N \times W \times H \tag{5}$$

In Equation (5), K and N represent the height and width of the convolution kernel, and H and W represent the height and width of the feature map, while M represents the number of channels, $K^2 \times N \times W \times H$ is the total amount of floating-point calculations corresponding to deep convolution operations, and $M \times N \times W \times H$ is the total amount of floating-point calculations corresponding to pointwise convolution operations. To further enhance nonlinear representation ability and avoid gradient vanishing in the Sigmoid function, this study introduces the Swish activation function. To reduce the computational cost of Swish, this study further improves it to the Hard-Swish activation function, which is shown in Equation (6).

$$Hard-Swish(x)=x\frac{ReLU6(x+3)}{6} \tag{6}$$

In Equation (6), $ReLU6$ represents a variant of the $ReLU$ function. The network uses stacked modules to achieve deep extraction and fusion of multi-scale coal and gangue features. Finally, global average pooling compresses spatial dimensions. The two-dimensional convolution converts the features to classification channels, while the output probability distribution of the coal and gangue features is done by the Softmax function. The DeepLabv3+ model based on MobileNetv4 significantly reduces parameter count and computational complexity; however, it still exhibits limitations in detecting small objects. BAM, referring to the Bottleneck Attention Module, assigns weights to feature maps across both channel and spatial dimensions [20]. Therefore, to alleviate weak feature representation in coal and gangue recognition, this study integrates BAM to strengthen the model's capability in capturing fine textures and edge features. The workflow of BAM in coal and gangue recognition is shown in Figure 4.

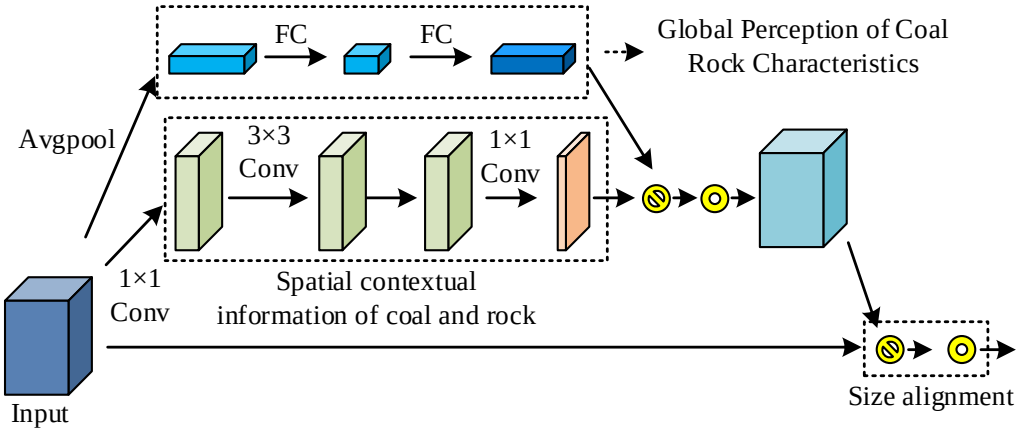


Fig. 4. BAM-based coal and gangue feature enhancement process

In Figure 4, BAM module takes input from the feature map of coal and gangue and produces attention maps using the attention mechanism to enhance features. The BAM module uses the architecture of dual-branch attention, namely channel and spatial attention. In the channel branch, to address uneven distribution of multi-channel features, global average pooling compresses spatial dimensions into a global vector, enabling global semantic understanding. A multilayer perceptron and batch normalization layer produce channel attention $M_C(F)$, as shown in Equation (7).

$$M_C(F)=BN(W_1(W_0Avgpool(F)+b_0)+b_1) \tag{7}$$

In Equation (7) BN represents batch normalization, W_1 and W_0 are learnable parameters, and b_0 and b_1 represent channel components. In the spatial branch, a 1×1 convolution first reduces channel dimensions to decrease computational redundancy. Then stacked 3×3 convolutions are used to expand the receptive field and extract spatial contextual information of coal and gangue targets. The spatial attention $M_S(F)$ calculation is shown in Equation (8).

$$M_S(F)=BN(f^{1\times1}(f^{3\times3}(f^{3\times3}(f^{1\times1}(F))))) \tag{8}$$

After completing the dual branch calculation, the channel attention and spatial attention are scaled to the same size and multiplied element by element for fusion. After activation by Sigmoid, they are mapped to the [0,1] interval to obtain the comprehensive attention weight. The complete fusion calculation formula is shown in Equation (9).

$$\begin{cases} M_{BAM}(F) = \sigma\left(M_C(F) \otimes M_S(F)\right) \\ F_{out} = F \otimes \left(1 + M_{BAM}(F)\right) \end{cases} \tag{9}$$

In Equation (9), $M_{BAM}(F)$ is the comprehensive attention weight, $\otimes$ is the Hadamard element wise product operation, σ represents the Sigmoid activation function, and F_{out} is the output feature map after attention enhancement. By relying on dual channel collaborative attention weights, the original features are weighted and corrected to enhance the effective features of coal gangue targets, suppress background interference from conveyor belts, and improve the model's ability to identify and characterize coal gangue edges and small particle materials.

2.2 Intelligent Separation Optimization Module Based on YOLOv7

The improved DeepLabv3+ coal-rock segmentation module achieves pixel-level foreground-background binary classification, accurately outlining the complete area and contour boundaries of coal-gangue materials on the conveyor belt. However, this network only distinguishes materials from blank backgrounds and does not have the ability to subdivide pixels for coal and gangue materials, and cannot directly output coal gangue category labels that support sorting actions. The YOLOv7 object detection algorithm can synchronously output target bounding boxes and coal gangue subdivision categories, with recognition accuracy and real-time inference advantages [21]. For this study, YOLOv7 was introduced as an auxiliary discrimination module, relying on the effective material area extracted by the segmentation branch to complete targeted coal gangue category determination, thereby improving the discrimination accuracy. The workflow of the YOLOv7 algorithm is shown in Figure 5.

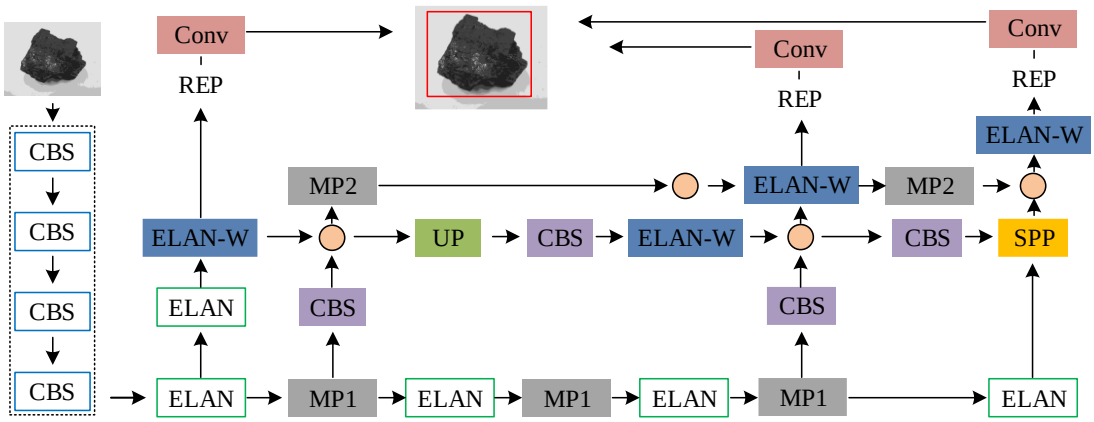


Fig. 5. YOLOv7-based coal and gangue detection and classification workflow

In Figure 5, the Conv-BatchNorm-SiLU (CBS) module in YOLOv7 is composed of a convolution layer, a batch normalization layer, and a SiLU activation function. These attributes obtained for coal and gangue from the CBS module are fed to the Extended Efficient Layer Aggregation Network (E-ELAN). The E-ELAN employs a multi-branch convolutional architecture to automatically control the width of channels and extract multi-scale feature information for coal and gangue. After that, down sampling reduces feature dimensions and extracts significant features. A matrix-weighted fusion strategy is used to optimize feature distribution, and the fusion process is shown in Equation (10).

$$F_{fusion} = \sum_{i=1}^{n} \alpha_i \cdot F_i \tag{10}$$

In Equation (10), α_i represents a learnable weight coefficient, and F_i represents the output feature. Although YOLOv7 exhibits excellent real-time object detection capabilities, there are certain limitations, including coordinate inconsistency and partial acquisition. The technology of

image coordinate mapping allows conversion from pixel coordinate systems to physical coordinate systems of the conveyor belt, ensuring correct localization data for separation. Hence, to reduce spatial inconsistencies during the separation process, image coordinate mapping technology is used to establish a transformation relationship between the image coordinate system and the physical coordinate system. Moreover, the image is partitioned into detection zones to ensure complete image acquisition. Coordinate mapping belongs to the target localization process, which receives pixel level positions provided by upstream target detection and converts them into physical coordinates recognizable by the execution end. The specific structure of image coordinate mapping technology is shown in Figure 6.

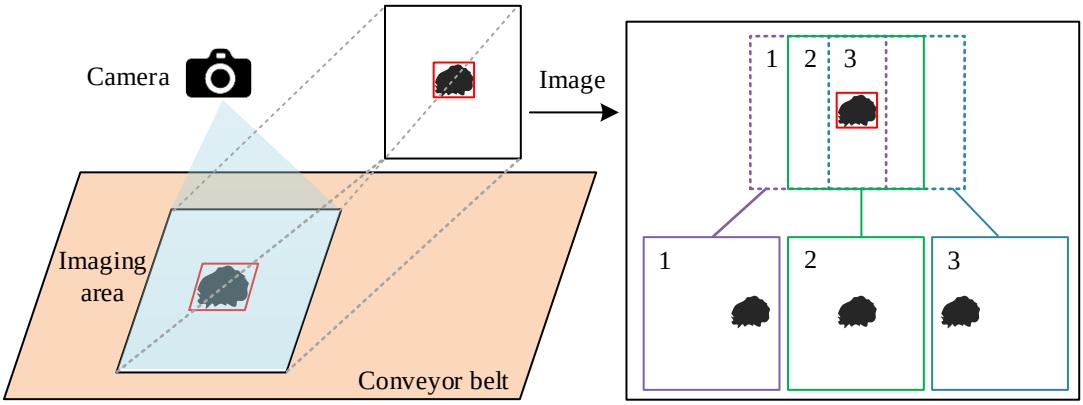


Fig. 6. Image coordinate mapping and coordinate transformation system

As shown in Figure 6, this method uses an industrial camera to capture coal and gangue targets on the conveyor belt in real time. Firstly, the chessboard calibration method is used to solve the camera intrinsic and extrinsic lens distortion coefficients, and a homography transformation model is constructed between the pixel plane and the conveyor plane. After the lens distortion correction is completed, the mapping relationship between the image pixel coordinates and the physical coordinates of the conveyor belt is established through scale translation transformation. The complete mapping process integrates homography matrix solution and pixel scale conversion, gradually achieving perspective deviation correction and accurate coordinate conversion. The scale conversion expression is shown in Equation (11).

$$s\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{11}$$

In Equation (11), x, y represents the coordinates in the physical plane coordinate system of the conveyor belt, s represents the non-zero scale factor in homogeneous coordinates, and the homography matrix $H=[h_{ij}]$ is a 3x3 matrix, with elements, h_{11}, h_{12}, h_{21}, h_{22} representing the rotation and scaling transformations between the image plane and the physical plane, h_{13}, h_{23} representing the translation transformation, representing the perspective distortion coefficient, h_{31}, h_{32} representing the positioning box coordinates. Multi-region coverage-based acquisition and detection are then performed. This method associates target features across different sub-regions through coordinate mapping to achieve complete recognition and localization. The fixed installation height of the module camera is 1200 mm, and the overhead shooting angle is 30 °. The average positioning error of the coordinate mapping of the single strain transformation is controlled within ± 1.8 mm. The conveyor belt runs at a constant speed of 0.2 m/s, and the algorithm is embedded with a motion compensation algorithm to dynamically correct the real-time physical coordinates of the target based on the conveyor speed, eliminating the positioning lag deviation caused by material movement. Through the use of spatial imaging coordinate mapping, further detection and tracking of coal and gangue targets is performed in space. However, the

conventional process of coal-gangue separation involves manual selection or a fixed-pressure air stream technique that results in low efficiency [22]. Therefore, in order to achieve automated coal rock sorting, a sorting module that integrates robotic arm grasping and high-pressure airflow injection is studied and designed. Based on the particle size information output by visual recognition, a fixed threshold rule is used to complete the switching of sorting branches, and the physical separation of coal rock is completed based on the visual recognition results. The coal rock sorting structure is shown in Figure 7.

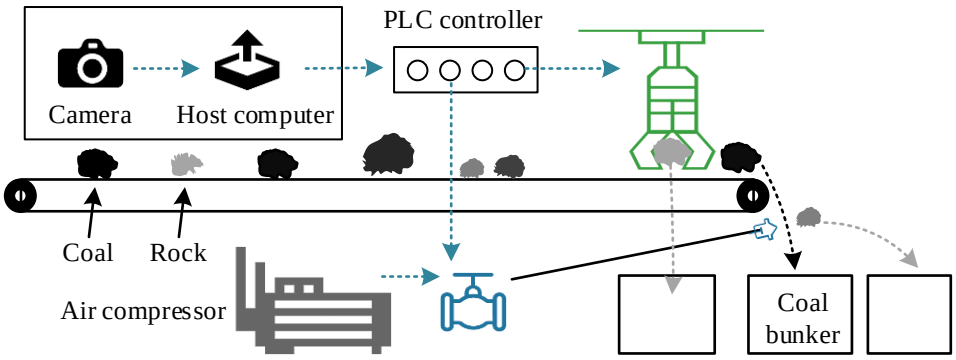


Fig. 7. Intelligent coal and gangue separation system

In Figure 7, the process first uses an industrial camera to collect real-time images of coal and rock on the conveyor belt and transmit them to the upper computer. After the algorithm completes the recognition and positioning of coal gangue targets, the pixel coordinates are converted to the physical coordinates of the conveyor belt through coordinate mapping, and sorting control instructions are generated based on the fixed particle size threshold rule. The actual size of material S_{obj} is obtained by transforming the pixel coordinates of the detection box using the camera calibration board. The sorting switching threshold S_{th} is a fixed particle size value pre-set by the process, which relies solely on size comparison rules to complete the sorting branch switching. The decision instructions are issued to the Programmable Logic Controller (PLC), and the model reserves communication and mechanical response delay compensation to match the on-site working conditions. When large-sized material reaches the execution point, the PLC drives the robotic arm to complete grasping and feeding. When small-sized materials arrive at the blowing area, the solenoid valve opens in advance, relying on the high-pressure airflow of the air compressor to complete the blowing separation. Finally, the coal material enters the coal bin, and the gangue is discharged into the gangue bin, completing the intelligent sorting of coal and rock.

3. Deep-YOLOv7 Model Performance Validation

3.1 Recognition Performance Analysis

To validate the coal-gangue recognition performance of the Deep-YOLOv7 model, this study compared it with Faster Region-based Convolutional Neural Network (Faster RCNN), Single Shot MultiBox Detector (SSD), and You Only Look Once version 5 small (YOLOv5s) models. The backbone network of Faster RCNN was ResNet50, and the backbone network of SSD was MobileNetv2. The server had Intel Xeon Gold 6248 CPU and NVIDIA Tesla V100 GPU. Edge deployment testing uses NVIDIA Jetson AGX Orin, equipped with a 12 core ARM Cortex-A78 CPU, Ampere architecture GPU (2048 CUDA cores and 64 Tensor cores), and 32GB LPDDR5 memory. The system for data acquisition involved an industrial camera Basler acA4024-29um that was positioned on top of the transfer point of the belt carrying unprocessed coal in the coal processing plant. As far as software is concerned, the system used Ubuntu 20.04/22.04, PyTorch 1.12.0, and TensorRT 8.5.

A total of 12800 original images of coal gangue were collected on site, with a single resolution of 2048 × 2048. After removing blurry and overexposed invalid samples, 10240 valid images were retained. The collected scenes covered four typical industrial working conditions: normal lighting,

low lighting, high dust, and high gangue mixing. The LabelMe tool was used to complete pixel-level labeling of coal and gangue. The dataset is strictly stratified and independently sampled according to different production days, conveyor belt operation batches, and raw coal material sources. It is divided into training set, validation set, and testing set in a 6:2:2 ratio. The three subsets have no homologous materials or overlapping duplicate images, and data leakage is prevented from the source. Scale the image uniformly to 640 × 640 pixels before inputting the model.

To analyze the recognition performance of the Deep-YOLOv7 model, this study compared the coal-gangue recognition accuracy and loss of the four models. The recognition accuracy uniformly refers to the object classification accuracy of coal and gangue targets in the image, and the statistical basis is the proportion of correct samples determined by the category of all detection boxes in the image to the total detection samples, and the results are shown in Figure 8.

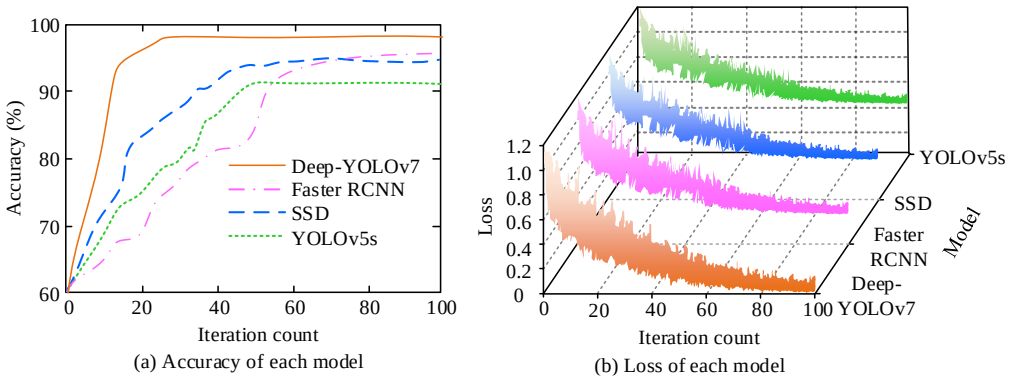


Fig. 8. Comparison of recognition accuracy and loss

In Figure 8(a), the recognition accuracy of the Deep-YOLOv7 network increased rapidly with the number of iterations, reaching 98.85% upon completion of all iterations, indicating a more favorable convergence trend and higher accuracy relative to other models. In Figure 8(b), the loss value of the Deep-YOLOv7 network decreases sharply to 0.46 within the first 20 iterations and then gradually declines, reaching 0.09 after 100 iterations. The above results indicated that the Deep-YOLOv7 model converged faster and possessed superior recognition accuracy. To analyze the performance of Deep-YOLOv7, this study compared the training, validation, and test mAP results of the four models, as shown in Figure 9.

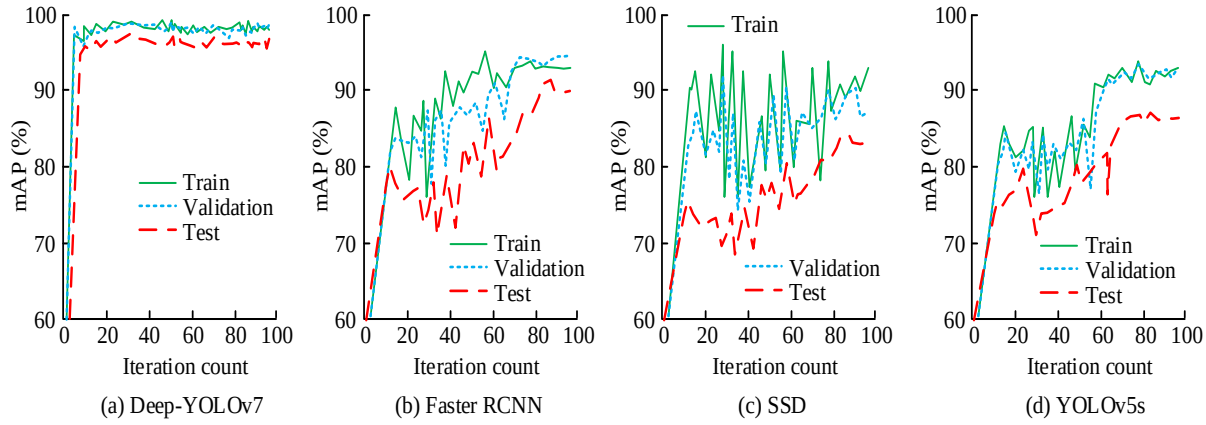
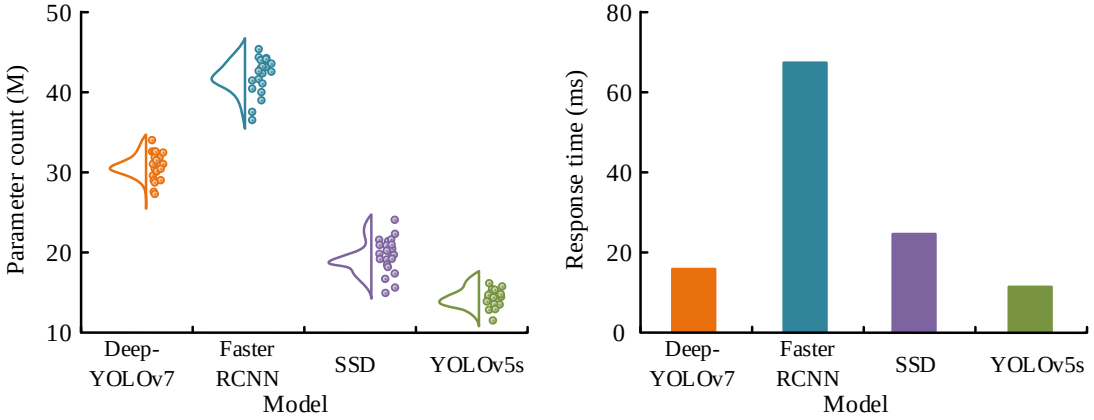


Fig. 9. Comparison of mAP results among different models

In Figure 9(a), the mAP of the Deep-YOLOv7 model increased rapidly during 0-20 iterations, and the mAP of the training, validation, and test sets quickly converged to 98.64%, with the three curves highly overlapping and no significant differences. This phenomenon was not due to the lack of independent dataset partitioning or data leakage. The core reason is that stratified sampling ensures a highly balanced distribution of working conditions, particle size, and category samples in each subset. In addition, the significant differences in visual characteristics between coal and gangue resulted in outstanding model generalization performance. In Figures 9(b), (c), and (d), the

convergence speed of Faster R-CNN, SSD, and YOLOv5s models was significantly slower than Deep-YOLOv7, and the mAP fluctuated considerably. The final test set mAP values were only 90.12%, 83.73%, and 87.27%, respectively. The training set and test set mAP of the comparison models were both lower than the proposed model. Overall, Deep-YOLOv7 had higher precision and learning ability, implying that the DeepLabv3+ semantic segmentation method proposed in this study conducted feature extraction with high accuracy for the region of coal-gangue, which contributed to better features representation in the training process without affecting its accuracy. In order to further analyze the computational efficiency of the Deep-YOLOv7 model, the response time and parameter quantity of coal rock recognition for four models were studied on NVIDIA Jetson AGX Orin edge devices. The test uses TensorRT FP16 precision, with an input image resolution of 640 × 640 and a batch size of 1. Preprocessing and post-processing are included in the total time consumption, and each model is tested 30 times, and the specific results are shown in Figure 10.



(a) Parameter count of each model

(b) Response time of each model

Fig. 10. Comparison of parameter quantity and response time

According to Figure 10(a), the model with the largest number of parameters is Faster R-CNN, with a total of 43.20 M. The model with the second-largest number of parameters is SSD, with 20.11 M. The third model, YOLOv5s, has 14.75 M. The fourth model, Deep-YOLOv7, possessed 30.82 M parameters, which was greater than the parameters of the second and third models but less than those of the first model. As shown in Figure 10(b), the response time of the Deep-YOLOv7 model was 17.27 ms, only 25.22% of Faster R-CNN. The above results verified that the model effectively balanced accuracy and computational efficiency. These trade-off advantages were mainly attributed to the MobileNetV4 backbone network controlling a significant increase in the number of parameters in the feature extraction stage, while YOLOv7 improved feature expression ability without significantly increasing latency through efficient convolution and hierarchical feature fusion. In order to analyze the integration effectiveness of the model, ablation experiments were designed, and the specific results are shown in Table 1.

In Table 1, the ablation experiment results indicate that each module contributes positively to the model performance. After replacing the backbone network from Xception to MobileNetV4, although mIoU slightly decreased by 0.41 percentage points, the inference speed increased from 18.6 FPS to 42.3 FPS, meeting real-time processing requirements. After introducing the BAM attention module, mIoU increased to 88.04% and FPS only slightly decreased to 39.7, verifying the enhancing effect of BAM on coal gangue boundary features. After integrating YOLOv7 and the coordinate mapping strategy into the complete model, the mAP@0.5 reached 96.38%, representing an increase of 1.17 percentage points compared to YOLOv7 alone. Compared to the standard YOLOv7, the research model not only improves detection accuracy, but also overcomes the positioning deviation of rectangular boxes in stacked occlusion scenes by guiding detection through segmentation masks, providing pixel level contour information. The final model achieved a good balance between accuracy and efficiency, proving the collaborative effectiveness of each module.

Table 1. Results of ablation experiment

Model	MobileNetV4	BAM	YOLOv7	Coordinate Mapping + Hybrid Sorting	mIoU (%)	mAP@0.5 (%)	FPS
DeepLabv3+ (Baseline)	×	×	×	×	86.32	/	18.6
MobileNetV4-DeepLabv3+	√	×	×	×	85.91	/	42.3
DeepLabv3+ + BAM	×	√	×	×	87.58	/	17.2
MobileNetV4-DeepLabv3+ + BAM	√	√	×	×	88.04	/	39.7
Standard YOLOv7 (Baseline)	×	×	√	×	/	95.21	27.5
MobileNetV4-DeepLabv3+ + YOLOv7	√	×	√	×	85.91	96.02	25.1
MobileNetV4-DeepLabv3+ + BAM + YOLOv7	√	√	√	×	88.04	96.15	25.0
Full Model	√	√	√	√	88.04	96.38	24.8

3.2 Separation Performance Analysis

The experimental server was based on the Intel Xeon Gold 6248 processor and NVIDIA Tesla V100 graphics card, whereas edge computing deployment experiments involved NVIDIA Jetson AGX Orin. The image capturing equipment was a Basler industrial camera. The development platform included Ubuntu 20.04/22.04, PyTorch 1.12.0, and TensorRT 8.5. After eliminating poor-quality samples, the coal-gangue image database contained 10240 images under various scenarios, including normal light, weak light, high dust, and high gangue concentration. The experimental equipment consisted of a six-axis collaborative robotic arm paired with an elastic adaptive gripper. The working pressure of the jet system is 0.6 MPa, and the nozzles are evenly arranged along the width of the conveyor belt. The conveyor belt runs at a constant speed of 0.2m/s, with a particle size range of 10-160mm for the test material. The model has an hourly processing capacity of 12t and has completed a total of 120 independent sorting tests.

To comprehensively evaluate the coal rock recognition and sorting performance of the Deep-YOLOv7 model under different working conditions, three typical coal types, anthracite, bituminous coal, and lignite, were selected for experiments and tested under three humidity conditions: dry, medium humidity, and high humidity. The semantic segmentation branch used Mean Intersection over Union (mIoU) and pixel accuracy as evaluation metrics to reflect the segmentation accuracy and overall pixel level classification accuracy of the model for coal gangue and other types of pixels, respectively. The object detection branch adopts precision, recall mAP@0.5 and mAP@0.5 0.95 is used as the evaluation index. The results are shown in Table 2. According to the results in Table 2, all indicators show a decreasing trend with increasing humidity. Under dry conditions, the mIoU of anthracite, bituminous coal, and lignite are 88.39%, 87.73%, and 86.17%, respectively, with pixel accuracies of 94.87%, 94.21%, and 93.05%, respectively, mAP@0.5 97.45%, 97.02%, and 96.08% respectively. Under high-humidity conditions, the performance of lignite decreased most significantly, with mIoU decreasing from 86.17% to 80.36%, mAP@0.5 0.95 decreased from 77.36% to 69.87%, due to the loose and porous surface of lignite, which has a stronger diffuse reflection effect on light after water attachment, exacerbating the convergence of visual

characteristics of coal gangue. Under high humidity conditions, the mIoU of smokeless coal is 83.55%, mAP@0.5 93.76%, and 82.80% and 93.12% respectively for bituminous coal. Under the worst high humidity brown coal conditions, the mIoU is 80.36%, mAP@0.5 At 91.28%, the adaptability of the model to complex humidity conditions was verified, and the enhancement of boundary features by the BAM module effectively suppressed the halo interference caused by moisture.

Table 2. Comprehensive performance of coal rock identification and detection under different humidity conditions

Coal Type	Moisture	mIoU (%)	Pixel Accuracy (%)	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5 :0.95 (%)
Anthracite	Dry	88.39	94.87	97.21	96.35	97.45	78.92
	Medium Moisture	86.70	93.12	95.84	95.12	96.18	76.43
	High Moisture	83.55	90.45	93.42	91.87	93.76	72.85
Bituminous Coal	Dry	87.73	94.21	96.78	95.96	97.02	78.15
	Medium Moisture	85.84	92.67	95.23	94.48	95.67	75.82
	High Moisture	82.80	89.82	92.85	90.96	93.12	71.98
Lignite	Dry	86.17	93.05	95.62	94.87	96.08	77.36
	Medium Moisture	84.02	91.38	94.08	93.23	94.53	74.69
	High Moisture	80.36	88.14	90.74	88.36	91.28	69.87

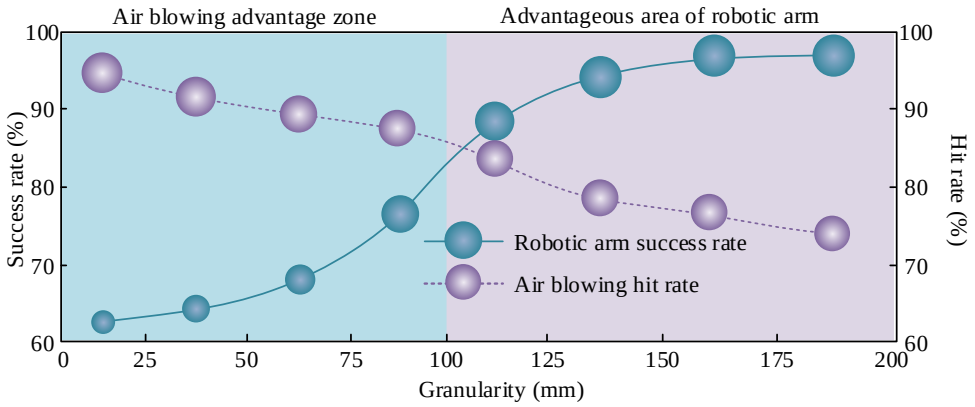


Fig. 11. Performance comparison of robotic arm grasping and air blowing separation

To analyze the adaptability differences of separation under different coal-gangue particle sizes, this study tested the robotic arm grasping success rate and air blowing separation hit rate of the Deep-YOLOv7 model in separation, as shown in Figure 11. In Figure 11, when the coal-gangue particle size was less than 100 mm, the air blowing separation hit rate remained between 87.18% and 95.92%, while the robotic arm grasping success rate was only between 62.37% and 88.45%. This was because small-particle coal-gangue had light weight and small inertia, and high-pressure airflow could quickly blow it off the conveyor belt. The clamping stability of the robotic arm for small coal-gangue blocks was poor, and grasping slippage easily occurred. As the particle size went beyond 100 mm, the rate of successful pickup by the robot arm greatly improved and reached 97.08%, while the hit rate of air blowing decreased to 74.34%. The large particle size of coal-gangue was large, and the air pressure during blowing increased considerably, resulting in a great reduction in the hit rate. The robot arm performed the active pick-up operation and had relatively higher pickup efficiency due to its high capability. The results mentioned above show that dynamic

conversion between robot arm pickup and air blowing segregation could efficiently exploit the advantages of both approaches. To compare the separation performance of Deep-YOLOv7, Faster RCNN, SSD, and YOLOv5s, this study calculated four indicators: Coal Precision (CP), Gangue Precision (GP), Gangue in Coal (GIC), and Coal in Gangue (CIG). All indicators are calculated based on the target pixel area as the basic statistical measure, where CP is the ratio of the pixel area correctly identified as coal by the model to the total pixel area determined as coal. GP is the ratio of the area of pixels correctly identified as gangue by the model to the total area of pixels identified as gangue. GIC is the ratio of the total pixel area of gangue mistakenly identified as coal to the total pixel area of coal. CIG is the ratio of the total pixel area of coal mistakenly identified as gangue to the total pixel area of gangue. The results are shown in Table 3.

Table 3. Comparison of separation performance among different models

Working Condition	Model	CP (%)	GP (%)	GIC (%)	CIG (%)
Normal Condition	Faster R-CNN	93.47	91.82	4.23	3.89
	SSD	89.16	86.73	6.58	5.94
	YOLOv5s	94.28	92.54	3.91	3.42
	Deep-YOLOv7	96.83	95.27	2.34	2.15
High Gangue Content (>30%)	Faster R-CNN	89.23	87.56	6.17	5.68
	SSD	84.37	81.22	9.14	8.47
	YOLOv5s	90.61	88.15	5.38	4.96
	Deep-YOLOv7	93.76	91.89	3.96	3.67

In Table 3, under normal working conditions, the CP and GP of the Deep-YOLOv7 model reached 96.83% and 95.27%, respectively, superior to the comparison models. Meanwhile, the GIC and CIG of the proposed model were only 2.34% and 2.15%, respectively, reduced by 1.57 percentage points and 1.27 percentage points compared to YOLOv5s, demonstrating optimal coal-gangue recognition accuracy in scenarios without complex interference. With respect to high gangue percentage situations, the results for all models have reduced considerably, but still, the Deep-YOLOv7 model showed superiority. Under dry conditions, the mIoU values of anthracite, bituminous coal, and lignite in the model were 88.39%, 87.73%, and 86.17%, respectively. Under high humidity conditions, the corresponding values are 83.55%, 82.80%, and 80.36%, with the worst case being high humidity brown coal mAP@0.5 Still reaching 91.28%, verifying the adaptability of the model. For the model, the CP and GP scores were 93.76% and 91.89%, respectively. With regards to GIC and CIG values, the model achieved 3.96% and 3.67%, respectively, solving the issue of coal-gangue confusion under high gangue percentage conditions. Overall, the Deep-YOLOv7 model retained its superiority in separation efficiency under both high and normal gangue percentage conditions.

4. Conclusion and Future Work

The conventional approaches for coal-gangue recognition and separation suffer from some drawbacks, such as low recognition accuracy and weak anti-interference ability. In view of this, this study put forward an innovative approach based on Deep-YOLOv7 towards improving the online recognition accuracy and intelligent separation of coal-gangue. The model utilized the DeepLabv3+ semantic segmentation framework to perform pixel-level extraction of coal-gangue images on conveyor belts in coal preparation plants, and enhanced coal-gangue boundary features through the MobileNetv4 backbone and BAM attention mechanism. YOLOv7 object detection and image coordinate mapping technology were integrated to optimize coal-gangue localization and separation decision-making, and dynamic separation was achieved through the coordination of robotic arm and air blowing. The results indicated that the Deep-YOLOv7 model achieved rapid training convergence, with recognition accuracy reaching 98.85% and mAP stabilizing at 98.64%. Meanwhile, the parameter quantity of the Deep-YOLOv7 model was 30.82 M, and the recognition response time was 17.27 ms, achieving optimal balance between parameter quantity and computational overhead. In addition, the proposed model achieved efficient separation across all particle sizes and maintained stable detection performance under both normal and high gangue content working conditions, with CP of 96.83% and 93.76%, respectively. In summary, this model

demonstrated significant advantages in recognition accuracy, real-time performance, and robustness under diverse operating conditions, thereby providing a practical technical solution for intelligent sorting in coal preparation facilities. In this study, performance evaluation was only conducted based on the normal working conditions and sizes of coal-gangue particles in coal preparation facilities, without considering any performance analysis under other working conditions. The long-term operational stability, multi-type coal-gangue adaptability, and generalization capability of the system still need further exploration.

Reference

[1] Zhang Y, Tong L, Lai X, Cao S, Yan B, Liu Y, et al. Coal-rock interface perception and accurate recognition in heading face under coal dust environment based on machine vision. J China Coal Soc. 2024; 49(7): 3276-3290.

[2] Waqas M, Naseem A. Artificial intelligence in sustainable industrial transformation: A comparative study of industry 4.0 and industry 5.0. FinTech Sustain Innov. 2025; 1: A2-A2. https://doi.org/10.47852/bonviewFSI52025321

[3] Guoxin L, Zhang S, Haiqing H, Xinxing H, Zhe Z, Xiaobing N, et al. Coal-rock gas: Concept, connotation and classification criteria. Pet Explor Dev. 2024; 51(4): 897-911. https://doi.org/10.1016/S1876-3804(24)60514-8

[4] Li G, Wang G, Feng N, Chen F, Feng Y, Lu C, et al. Deep coal-rock gas in China: A review of distribution, geological characteristics, and its enrichment conditions. ACS Omega. 2025; 10(22): 23472-23491. https://doi.org/10.1021/acsomega.5c02056

[5] Khoshmagham A, Hosseini Alaee N, Shirinabadi R, Bangian Tabrizi AH, Gholinejad M, Kianoush P. Investigating the time-dependent behavior of intact rocks and fractured rocks using unconfined relaxation testing in underground coal mines. Geotech Geol Eng. 2024; 42(8): 6889-6922. https://doi.org/10.1007/s10706-024-02902-5

[6] Zhao X, Li S, Sun Z, Ma H, Kong X, Xu Y. Detection of building shadows in high-resolution remote sensing images by using improved DeepLabV3+. Remote Sens Lett. 2025; 16(3): 290-301. https://doi.org/10.1080/2150704X.2025.2452310

[7] Hu C, Tan L, Wang W, Song M. Lightweight tea shoot picking point recognition model based on improved DeepLabV3+. Smart Agric. 2024; 6(5): 119-127.

[8] Cheng L, Xiong R, Wu J, Yan X, Yang C, Zhang Y, et al. Fast segmentation algorithm of USV accessible area based on attention fast DeepLabV3. IEEE Sensors J. 2024; 24(15): 24168-24177. https://doi.org/10.1109/JSEN.2024.3410403

[9] Chen H, Qin Y, Liu X, Wang H, Zhao J. An improved DeepLabv3+ lightweight network for remote-sensing image semantic segmentation. Complex Intell Syst. 2024; 10(2): 2839-2849. https://doi.org/10.1007/s40747-023-01304-z

[10] Cao Z, Li Z, Fang L, Li J. Lightweight coal and gangue detection algorithm based on improved Yolov7-tiny. Int J Coal Prep Util. 2024; 44(11): 1773-1792. https://doi.org/10.1080/19392699.2023.2301304

[11] Shatwell DG, Murray V, Barton A. Real-time ore sorting using color and texture analysis. Int J Min Sci Technol. 2023; 33(6): 659-674. https://doi.org/10.1016/j.ijmst.2023.03.004

[12] Jiang J, Han Y, Zhao H, Suo J, Cao Q. Recognition and sorting of coal and gangue based on image process and multilayer perceptron. Int J Coal Prep Util. 2023; 43(1): 54-72. https://doi.org/10.1080/19392699.2021.2002852

[13] He L, Wang S, Guo Y, Hu K, Cheng G, Wang X. Study of raw coal identification method by dual-energy X-ray and dual-view visible light imaging. Int J Coal Prep Util. 2023; 43(2): 361-376. https://doi.org/10.1080/19392699.2022.2051013

[14] Luo Q, Wang S, Guo Y, Li D, He L, Cheng G. Research on identification of coal and gangue under complex working conditions based on relief feature. Int J Coal Prep Util. 2024; 44(9): 1431-1446. https://doi.org/10.1080/19392699.2023.2282633

[15] Liu H, Liu Y, Yao S, Yu T, Gao K, Hao P, et al. ISTFormer: lightweight transformer for enhanced super-resolution of coal rock images via iterative feature extraction. Visual Comput. 2025; 41(10): 7079-7092. https://doi.org/10.1007/s00371-024-03793-6

[16] Khan M, Xueqiu H, Dazhao S, Xianghui T, Li Z, Yarong X, et al. Extracting and predicting rock mechanical behavior based on microseismic spatio-temporal response in an ultra-thick coal seam mine. Rock Mech Rock Eng. 2023; 56(5): 3725-3754. https://doi.org/10.1007/s00603-023-03247-w

[17] Greve J, Busch B, Quandt D, Knaak M, Hilgers C. Understanding the interplay of depositional rock types, mineralogy, and diagenesis on reservoir properties of the coal-bearing Langsettian and Duckmantian strata (Bashkirian, Pennsylvanian) of the Ruhr Area, NW Germany. Int J Earth Sci. 2024; 113(8): 2275-2304. https://doi.org/10.1007/s00531-024-02454-2

[18] Sundaresan AA, Solomon AA. Post-disaster flooded region segmentation using DeepLabv3+ and unmanned aerial system imagery. Nat Hazards Res. 2025; 5(2): 363-371. https://doi.org/10.1016/j.nhres.2024.12.003
[19] Anugrah MN, Rachmat N. Klasifikasi Penyakit Daun Tanaman Jagung Menggunakan Pendekatan Transfer Learning Arsitektur MobileNetv4. J Technol Manag Ind Appl. 2025; 4(4): 2174-2182. https://doi.org/10.55826/jtmit.v4i4.1393
[20] Amrani M. Deep convolutional multi-informative metric correlation analysis with bottleneck attention module for face recognition in the wild. Multimed Tools Appl. 2024; 83(22): 62459-62487. https://doi.org/10.1007/s11042-023-15962-1
[21] Mammeri S, Amroune M, Haouam MY, Bendib I, Corrêa Silva A. Early detection and diagnosis of lung cancer using YOLO v7, and transfer learning. Multimed Tools Appl. 2024; 83(10): 30965-30980. https://doi.org/10.1007/s11042-023-16864-y
[22] Wang P, Ma H, Zhang Y, Cao X, Wu X, Qiao H, et al. A cooperative strategy of multi-arm coal gangue sorting robot based on immune dynamic workspace. Int J Coal Prep Util. 2023; 43(5): 794-814. https://doi.org/10.1080/19392699.2022.2078808