Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Coal-Gangue Detection and Pose Estimation Method Based on DRD- YOLOv11n

Kaiyun Chen
Heilongjiang University of Science and Technology
Boyang Yin
Heilongjiang University of Science and Technology
Yilin Ge
gylgyl777@163.com
Heilongjiang University of Science and Technology
Minhang Zhu
Heilongjiang University of Science and Technology
Yanwei Wang
Heilongjiang University of Science and Technology

Article

Keywords: Coal Gangue Detection, YOLOv11n, DySample, DCNv3Dynamic detection head, Pose Estimation

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11055302/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Coal-Gangue Detection and Pose Estimation Method Based on DRD-YOLOv11n

Kaiyun Chen [1,3], Boyang Yin [1,3], Yilin Ge [2,3], Minhang Zhu [1,3], Yanwei Wang [1,3]

1. School of Mechanical Engineering, Heilongjiang University of Science and Technology, Harbin 150022, Heilongjiang, China;
2. School of Computer Science, Heilongjiang University of Science and Technology, Harbin 150022, Heilongjiang, China;
3. Heilongjiang Provincial Key Laboratory of Optical 3D Measurement and Detection, Harbin

Funding Project: Heilongjiang Provincial Natural Science Foundation Research Team Project (TD2025E005)

Author Biography: Kaiyun Chen (1974-), male, native of Jintan, Jiangsu Province, Han ethnicity, Professor, Doctoral Supervisor, Ph.D. E-mail: chenkaiyun@hrbeu.edu.cn.

Corresponding Author: Yilin Ge (1989-), female, native of Huaian, Jiangsu Province, Han ethnicity, Lecturer, Master's Supervisor, Ph.D. E-mail: gylgyl777@163.com.

To deal with difficult feature extraction and single-dimensional information caused by poor sorting conditions and various target shapes, this paper puts forward a coal-gangue detection and pose estimation method based on DRD-YOLOv11n. First, to avoid the checkerboard effect, we add the DySample operator for dynamic upsampling into the feature fusion network. It takes the place of common interpolation and carries out dynamic feature resampling with low computing cost. Second, to reduce noise from dust and light and strengthen detailed features, we put a residual enhancement module before the detection head. It uses depthwise separable convolution and the SE attention module to complete effective feature extraction. Besides, to realize accurate positioning of objects of different sizes, we combine multi-dimensional attention and make use of the geometric deformation of the DCNv3 operator to rebuild a dynamic detection head based on DCNv3. Finally, to offer accurate grasping positions for the mechanical claw, a post-processing module for pose estimation is set up. We use the minimum area bounding rectangle (MABR) fitting and exponential moving average (EMA) time filtering. With the separating axis theorem, we build a dynamic safety zone to calculate stable rotation angles and ensure safe grabbing. The test results show that on the coal and gangue dataset, compared with the original YOLOv11n, the improved DRD-YOLOv11n only has a model weight of 5.1 MB. Its mAP50-95 reaches 85.6%, which is 0.6% higher than the baseline. The F1 score is 95.1%, also rising by 0.6%. The number of parameters is cut down by 6%, and the detection speed reaches 114 frames per second.After adding pose estimation, the grabbing success rate of the mechanical claw for long strip objects rises from 60% to 90%, and the overall safe grabbing rate is guaranteed even under crowded working conditions.While keeping the model lightweight, this method balances detection accuracy and speed. Besides, the newly added modules solve the problem of low positioning accuracy. It provides key support and theoretical reference for the practical engineering use of sorting robots.

Keywords: Coal Gangue Detection; YOLOv11n; DySample；DCNv3；Dynamic detection head; Pose Estimation

0. Introduction

As China's primary energy source and an important industrial raw material, coal will continue to play a crucial "ballast" role for a long time to come. However, during the coal production process, a large amount of solid waste is generated, about one-third of which is coal gangue. The huge existing amount and rapid growth of coal gangue have caused serious ecological and environmental problems [1]. Under the effects of rainwater leaching and air oxidation, substances such as sulfides in coal gangue can easily produce acidic wastewater, thereby polluting the soil and groundwater systems. In addition, coal gangue dumps may spontaneously combust under specific conditions, releasing toxic and harmful gases such as sulfur dioxide, hydrogen sulfide, and carbon monoxide. Combined with the dust generated by weathering, this severely pollutes the atmospheric environment and endangers the health o

f surrounding residents [2-4]. Therefore, effectively sorting coal gangue from coal can not only avoid the above environmental and health problems, but also fully utilize the potential of coal gangue as a building material, thereby bringing significant social, ecological, and economic benefits.

Traditional coal-gangue sorting methods mainly include manual sorting, dense medium separation, and jigging separation. As a traditional operation in coal preparation, manual sorting has always had problems such as high labor intensity and poor working environments, including noise and dust [5]. Working for a long time can harm the health of workers. Dense medium separation [6] and jigging separation [7] have low sorting efficiency and high resource consumption, and they also cause significant environmental pollution.

Currently, coal-gangue sorting is evolving towards an intelligent and efficient direction. For coal-gangue recognition, there are mainly two development paths: radiation recognition and image recognition. For sorting execution, there are mainly high-pressure air jet sorting and truss robotic arm sorting. Although radiation recognition technology has certain application value in specific scenarios, it is limited by defects such as radiation safety issues and insufficient penetration for large-sized materials [8]. In contrast, image recognition technology has become the mainstream of industry development due to its advantages of high precision, low cost, and real-time performance. At the execution level, compared with the traditional air jet method, truss robotic arm sorting shows stronger adaptability in precise positioning and multi-dimensional control under complex working conditions.

The intelligent coal-gangue sorting robot prototype used in this study (as shown in Figure 1) is designed based on this technical architecture of image recognition and robotic arm sorting. This prototype integrates the image recognition system and the truss robotic arm, and has achieved basic operational capabilities in a laboratory environment. To further address the issue of insufficient grasping robustness for irregular coal and gangue, a rotation function is introduced into the prototype's gripper. On this basis, how to accurately and in real time extract the geometric center and spatial pose of the target in complex industrial environments has become the key bottleneck restricting the prototype's operational efficiency from moving from the laboratory to engineering applications.

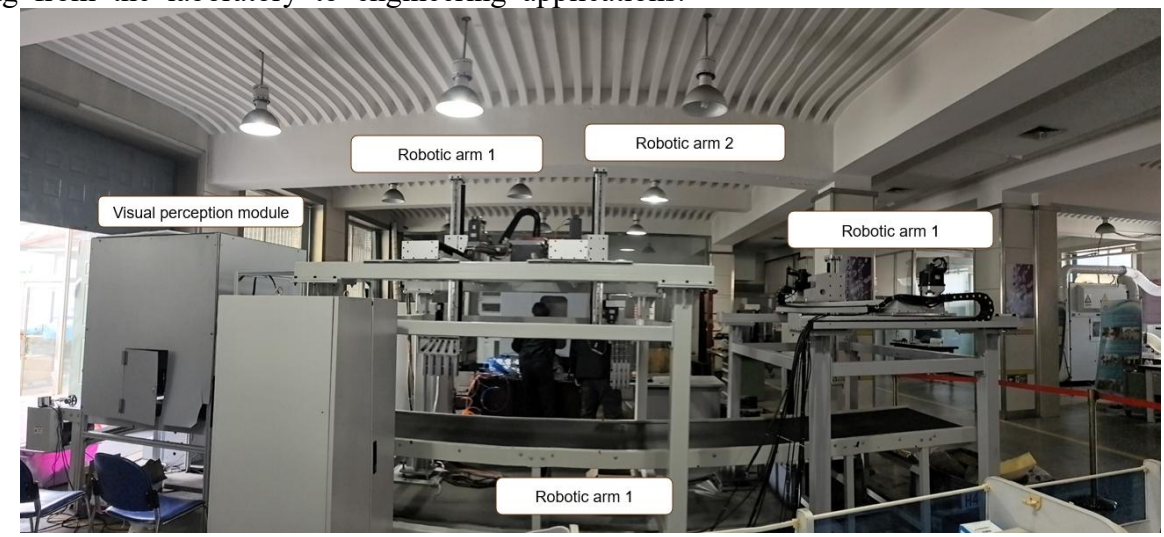


Figure 1. Prototype of intelligent coal gangue sorting robot

In terms of coal-gangue recognition algorithms, object detection networks based on deep learning have become the mainstream method. Currently, such algorithms are mainly divided into two categories: two-stage and one-stage. Although two-stage algorithms represented by Faster R-CNN have high accuracy, they are limited by the complex region proposal extraction process, making it difficult for their inference speed to meet industrial real-time requirements. In contrast, one-stage algorithms represented by SSD and the YOLO series integrat

e localization and classification into a single regression task, which significantly improves p rocessing efficiency [9]. Given the strict requirements for real-time response in coal-gangue s orting operations, this paper selects a one-stage algorithm with more balanced performance as the base architecture.

Many scholars have conducted in-depth research on coal-gangue recognition methods. Z hang Chao et al. [10] proposed a coal-gangue recognition method based on the Gray Level Co-occurrence Matrix (GLCM). By extracting texture features such as energy, entropy, mom ent of inertia, and correlation from coal and gangue images, they used a Support Vector M achine to build a classification model for discrimination. Li Juanli et al. [11] proposed a fast coal-gangue sorting model based on lightweight SSD. By using MobileNetV2 to replace the original VGG16 backbone feature extraction network and introducing a Feature Pyramid Net work structure for multi-scale feature fusion, they solved the problem of low recognition ac curacy for small-particle coal and gangue in the original SSD algorithm, and improved the detection speed and mAP metrics. Xue G H [12] proposed a coal-gangue detection algorithm based on ResNet18-YOLO. This algorithm used the lightweight network ResNet18 as the ba ckbone feature extraction network of YOLOv3. Through feature scale reduction and unstruct ured pruning optimization, it improved the model's lightweight nature and real-time performa nce.Li Deyong et al. [13] proposed an intelligent recognition model based on the CFS-YOLO algorithm. The model adopted the ConvNeXt V2 feature extraction module, combined with MAE and GRN layers, to solve feature collapse and maintain diversity. The Focal-EIOU los s function was introduced to optimize sample imbalance and improve convergence speed an d localization accuracy. The SimAM parameter-free attention mechanism was added to enha nce the extraction of key features, thereby improving recognition accuracy. Aiming at the lo w-illumination and high-dust environment in coal mines, Wei Xiaolong et al. [14] proposed a coal-gangue recognition model based on CSPNet-YOLOv7. The model optimized feature ex traction through CSPNet, used RFP and SAC to improve network performance, and adopted transfer learning to enhance generalization ability.Li Na et al.[15] proposed a lightweight FNF -YOLO model based on YOLOv8n. They used a multi-layer weighted fusion method, FNF, to optimize the pyramid structure to enhance feature fusion ability, and adopted the C2f-FE MA module in the neck network to reduce computational redundancy. This improved the re cognition accuracy and reduced the computation amount. Zhang Jinwang et al.[16] proposed the HDB-YOLO model based on YOLOv11n. This model achieved a lightweight design thr ough the HS-FPN neck network, used Dynamic convolution to improve generalization accur acy, and integrated the BiFormer attention mechanism to strengthen global feature modeling. Experiments showed that the mAP@50-95 of the improved model increased by 5.70%, and single-frame inference took only 3.2 ms.

Although the above algorithms have improved the accuracy of coal-gangue recognition, they still face many challenges when applied to real-world coal-gangue sorting scenarios. O n the one hand, most existing recognition algorithms are only trained and evaluated on stan dard datasets under controlled laboratory conditions. They lack adaptive optimization and ro bustness verification for the complex working conditions of on-site coal sorting. On the oth er hand, current research often focuses on improving the detection metrics of small and me dium-sized targets, while ignoring the problems of irregular shapes, large scale spans, and p oor localization accuracy of coal and gangue.More importantly, for the mainstream robotic a rm grasping sorting equipment, existing detection models can only provide target bounding box information. They cannot output the target pose parameters required for the precise pos itioning and pose adjustment of the robotic gripper. Therefore, improving the robustness and multi-scale adaptability of algorithms in complex industrial environments, and deeply integr ating pose estimation technology, has become a key problem that urgently needs to be solv ed in this field. This is essential to meet the urgent needs for efficient sorting and precise grasping of large coal and gangue on site.

To address the above challenges, this paper proposes a coal-gangue recognition and pos e estimation method based on DRD-YOLOv11n for on-site coal-gangue sorting. Using YOL Ov11n as the base framework, this method makes a series of targeted structural improveme nts. First, in the feature fusion stage, an efficient content-aware dynamic upsampling operato r DySample is introduced to replace the traditional interpolation method. It achieves dynami c resampling at a lower computational cost, reduces information loss during the feature fusi on process, and improves the model's adaptability to low-quality images. Second, a Residual Depthwise Feature Enhancement module RD-SE is embedded before the detection head. By combining residual depthwise convolution and the SE channel attention mechanism, it refin es the features a second time. This suppresses background noise while enhancing the model' s sensitivity to the subtle texture features of coal and gangue. Third, the detection head is r econstructed into a dynamic detection head based on DCNv3. This effectively solves the dif ficult problem of accurate localization for coal-gangue targets with irregular shapes and hug e scale spans. Finally, after the model completes target localization, to meet the needs of li ghtweight deployment and annotation-free pose extraction on site, this paper further designs a pose estimation method based on minimum bounding rectangle fitting. Combined with a t emporal filtering strategy, it achieves a stable and accurate output of the rotation angle of t he coal-gangue main axis.

Overall, the improved DRD-YOLOv11n reduces model complexity while maintaining hi gh recognition accuracy, and possesses the ability to detect the position and pose of coal a nd gangue at various scales. It provides a high-performance and deployable solution for coa l-gangue sorting.

1. Overview of the Coal-Gangue Detection and Pose Estimation Method Based on D RD-YOLOv11n

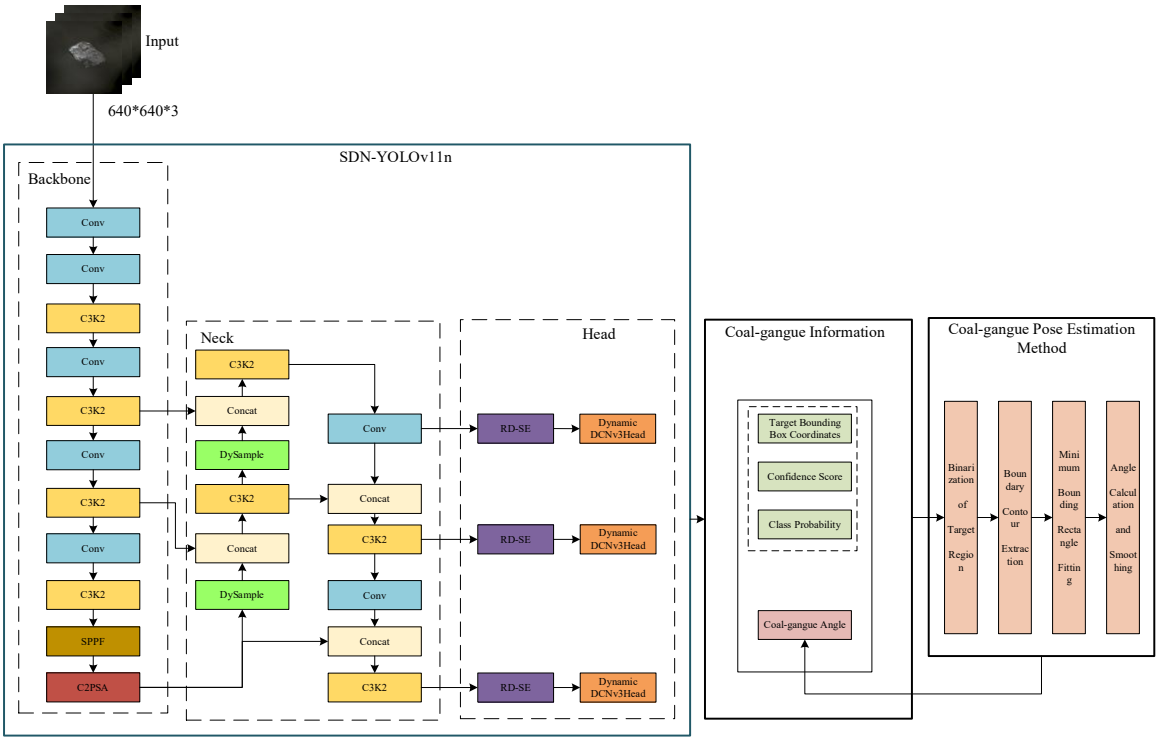


Figure 2. Coal-gangue detection and pose estimation model based on DRD-YOLOv11n

To address the environmental characteristics of coal-gangue sorting sites and the practic al requirements for efficient and precise grasping in intelligent sorting systems, this paper p roposes a coal-gangue detection and pose estimation method based on DRD-YOLO11n. Usin g YOLOv11n as the baseline, the DySample operator is introduced into the feature fusion n

etwork to improve edge recovery quality under low-illumination and motion-blurred conditio ns. By embedding the RD-SE module, the method effectively suppresses background noise a nd enhances texture sensitivity. Simultaneously, the dynamic detection head is reconstructed utilizing DCNv3, resolving the challenges of precise localization for irregular and multi-scal e targets. Building upon the detection framework, a pose estimation module based on the Minimum Area Bounding Rectangle (MABR) and Exponential Moving Average (EMA) filter ing is integrated, providing stable and reliable rotation angle parameters for the robotic man ipulator. The overall architecture of the DRD-YOLO11n model is illustrated in Figure 2.

1.1 Structural Improvement of Feature Fusion Network Based on Dynamic Upsampling

In the Feature Pyramid Network of YOLOv11, upsampling is a key step to achieve m ulti-scale feature fusion. YOLOv11 uses nearest neighbor interpolation. This method follows fixed geometric rules and ignores the semantic information in the feature maps. In the coal-gangue detection scenario, due to uneven underground illumination and motion blur, the edg es of coal and gangue are often not sharp enough. Static interpolation easily leads to the lo ss of key texture and boundary information due to the checkerboard effect or smoothing pr ocessing, which in turn affects the localization accuracy of the subsequent detection head fo r small targets. Therefore, this paper introduces an ultra-lightweight and efficient dynamic u psampling operator DySample[17], aiming to achieve content-aware feature recovery at an ext remely low computational cost. DySample treats upsampling as a resampling process from l ow-resolution feature maps. It does not rely on convolution kernels, but achieves dynamic u psampling by generating content-aware sampling point positions.

DySample abandons the heavy computation of dynamic convolution and returns to the essence of upsampling: point sampling. Its core idea is to model the upsampling process as a resampling of the input feature maps. By learning content-aware dynamic offsets, it adapti vely adjusts the sampling point positions. The sampling process is shown in Figure 3.

Suppose the input feature map is $X \in \mathbb{R}^{C\times H\times W}$, and the upsampling scale factor is s. Th e workflow of DySample mainly includes two steps:

1. Sampling Point Generation:

Unlike standard interpolation that uses a fixed grid G, DySample dynamically generate s offsets O based on the input feature X. To ensure an orderly distribution of sampling po ints and avoid sampling overlap, DySample introduces a dynamic scope factor. Specifically, the generation formula for the offsets is as follows:

$$\mathrm{O} = s \cdot \mathrm{sigmoid}(\mathrm{linear}_1(X)) \cdot \mathrm{linear}_2(X) \tag{1}$$

Where linear_1 and linear_1 are linear projection layers. $\mathrm{sigmoid}(\cdot)$ limits the scope factor between (0,1), so that the offsets are dynamically constrained within a local range, thereby effectively avoiding boundary artifacts. The sampling point generator is shown in Figure 4.

2. Grid Sampling:

After obtaining the offset O, the final set of sampling positions O ,the final set of sa mpling positions S is the sum of the original grid and the offset: $\mathrm{S} = \mathrm{G} + \mathrm{O}$. Using PyTorc h's built-in efficient grid_sample function, the final upsampled feature X' is calculated as foll ows:

$$X' = \mathrm{grid_sample}(\mathrm{X}, \mathrm{S}) \tag{2}$$

In this way, the sampling points can deform according to the morphological features of the coal and gangue. They adaptively gather at object edges and texture-rich regions, rat her than being uniformly distributed in the background.

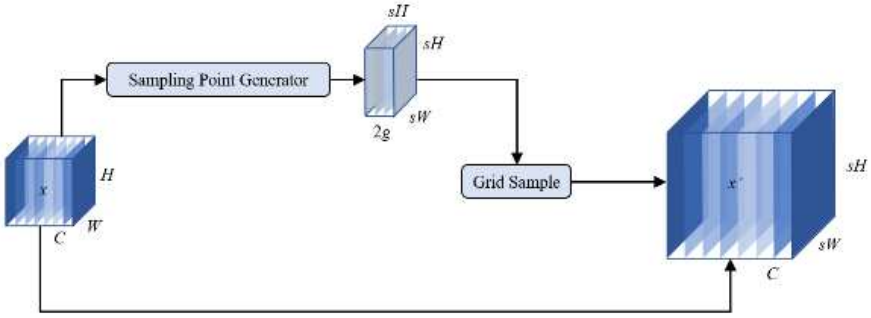


Figure 3. Sampling-based dynamic upsampling process

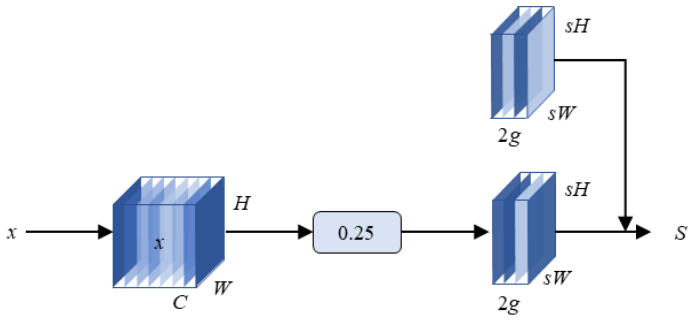


Figure 4. Sampling point generator in DySample

1.2 Residual Feature Enhancement Module

The coal-gangue sorting site is usually accompanied by dust interference and low-illumi nation conditions, resulting in a large amount of background noise in the captured images. I n addition, the visual differences between coal and gangue are mainly reflected in subtle ch anges in surface texture. Traditional convolutional feature extraction often struggles to preser ve high-frequency texture details while suppressing noise. Although the original C2PSA mod ule of YOLOv11 has a global receptive field, it incurs a high computational cost when pro cessing local texture details.

To further refine the features before they enter the detection head, this paper proposes a lightweight Residual Depthwise Feature Enhancement module. This module aims to extract local spatial features through depthwise convolution with residual connections, and integrate s the SE (Squeeze-and-Excitation) attention mechanism [18] to recalibrate feature channels. Th is thereby enhances the model's sensitivity to the key texture features of coal and gangue while suppressing the interference of environmental noise.

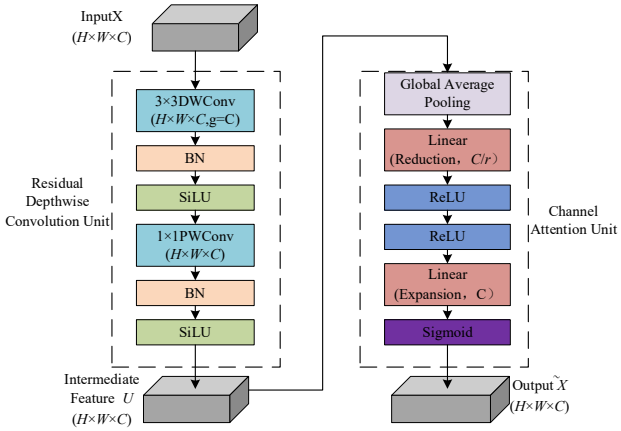


Figure 5. Structure of the Residual Depthwise Feature Enhancement (RD-SE) module

The structure of the RD-SE module is shown in Figure 5, which is mainly composed of a Residual Depthwise convolution unit (RD module) and a Channel Attention unit (SE module) connected in series:

The residual depthwise feature extraction unit adopts a depthwise separable convolution

structure. First, it utilizes a 3×3 depthwise convolution to extract spatial features independen tly on each channel, which maintains feature independence and reduces the number of para meters. Subsequently, it performs inter-channel information interaction through a 1×1 pointwi se convolution. A residual connection is introduced to add the input features to the convolu tion output. This not only promotes the effective propagation of gradients in deep networks, but also enables the network to focus on learning the residual features of coal and gangue relative to the background.

To enhance the discriminative feature channels, the SE attention mechanism is introduc ed. This unit first compresses spatial information through global average pooling, and subse quently constructs inter-channel dependencies via two fully connected layers. Finally, it gene rates channel weight coefficients to adaptively reweight the input features.

Suppose the input feature map of this module is $X \in \mathbb{R}^{C\times H\times W}$, where H,W,C represent th e height, width, and number of channels of the feature map, respectively.

First, the feature map is processed through the residual depthwise convolution unit. Let F_{pw} denote the depthwise convolution transformation and F_{pw} denote the pointwise convolutio n transformation. Then, the intermediate feature U can be expressed as:

$$U = \mathrm{F}_{pw}(\sigma(\mathrm{BN}\ (\mathrm{F}_{dw}(X)))) + X \tag{3}$$

Where σ represents the SiLU activation function, and BN represents Batch Normalizatio n. The residual connection ensures the preservation of original semantic information.

To utilize the global information between channels, global average pooling is performed on U, compressing the spatial dimensions $H\times W$ to 1×1 to obtain the channel descriptor $z \in \mathbb{R}^{C}$:

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i,j) \tag{4}$$

To capture the non-linear dependencies between channels, two fully connected layers ar e used to generate the weight vector s. The first fully connected layer reduces the channel dimension to C/r (in this paper, $r=16$), and the second fully connected layer restores it to C:

$$s = \sigma_{sig}(W_2\delta(W_1 z)) \tag{5}$$

Where is the ReLU activation function, σ_{sig} is the Sigmoid function, and $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are learnable weight matrices.

The final output feature $\tilde{X}$ is obtained by expanding the weight s and performing chann el-wise multiplication with the intermediate feature U:

$$\tilde{X}_c = s_c \cdot U_c \tag{6}$$

Through the above process, the RD-SE module effectively enhances the feature channel s containing key information of coal and gangue and suppresses the background noise chan nels. This provides higher-quality feature inputs for the subsequent DCNv3-based dynamic d etection head.

1.3 Dynamic Detection Head Based on DCNv3

In the automatic coal-gangue sorting scenario, the target objects exhibit significant unstr uctured characteristics: the shapes of coal and gangue are highly irregular, and their scale v ariation range on the conveyor belt is extensive[19]. Although the original detection head of YOLO11 is efficient, it relies on standard convolution kernels for feature extraction, which presents two main limitations:

Limited geometric perception: Standard convolution kernels possess a fixed, typically re

ctangular receptive field shape. This makes it difficult to adaptively fit the edges of irregula rly shaped coal and gangue, resulting in the introduction of a large amount of conveyor bel t background noise into the feature maps.

Insufficient multi-scale feature coupling: When handling objects with large scale variatio ns, traditional feature decoupling methods are prone to causing feature loss for small targets or semantic blurring for large targets.

To address the aforementioned problems, this paper proposes an improved dynamic dete ction head named DynamicDCNv3Head. Based on the unified attention mechanism of Dyna mic Head [20], this module upgrades the spatial perception module to the third-generation def ormable convolution, DCNv3[21]. It aims to enhance the feature representation of coal and g angue from three dimensions: scale, spatial, and task.

Treat the feature pyramid output by the backbone network as a three-dimensional tenso r $\mathrm{F} \in \square^{L\times S\times C}$, where L represents the feature level, S represents the spatial dimension ($H\times W$), and C represents the channel dimension. To avoid the high computational cost introduc ed by the full attention mechanism, DynamicDCNv3Head applies three attention mechanisms sequentially:

$$W(\mathrm{F})=\pi_C(\pi_S(\pi_L(\mathrm{F})\cdot\mathrm{F})\cdot\mathrm{F})\cdot\mathrm{F} \tag{7}$$

Where π_L, π_S, and π_C represent the scale-aware, spatial-aware, and task-aware attentio n functions, respectively.

1. Scale-Aware Attention（π_L）: Multi-Level Dynamic Fusion

To address the problem of large size variations in coal and gangue, the π_L module dyn amically fuses features across different levels. For the feature map at level l, the network a daptively aggregates information from the previous level $l-1$ (after upsampling) and the ne xt level $l+1$:

$$\pi_L(\mathrm{F})\cdot\mathrm{F}=\sigma\left(f\left(\frac{1}{SC}\sum_{S,C}\mathrm{F}\right)\right)\cdot\mathrm{Fuse}(\mathrm{F}_{l-1},\mathrm{F}_l,\mathrm{F}_{l+1}) \tag{8}$$

Where $f(\cdot)$ is a linear approximation function implemented by a 1×1 convolution, an d σ is the Hard-Sigmoid activation function. This mechanism enables the detection head to adaptively adjust the weights of different feature layers according to the size of the coal an d gangue.

2. Spatial-Aware Attention（π_S）: Geometric Deformation Modeling Based on DCNv3

The traditional Dynamic Head uses DCNv2 for spatial modeling, whereas this method i ntroduces DCNv3. Compared to the former, DCNv3 introduces a group mechanism and mor e stable modulation scalars, enabling it to more accurately capture the geometric deformatio n of irregular objects.

Specifically, for a spatial location p, DCNv3 divides the feature channels into G groups, and introduces a learnable offset Δp_{gk} and a modulation scalar m_{gk} at each sampling point:

$$\pi_S(\mathrm{F})\cdot\mathrm{F}=\sum_{g=1}^{G}\sum_{k=1}^{K}w_g\cdot m_{gk}\cdot\mathrm{F}(p+p_k+\Delta p_{gk}) \tag{9}$$

Where K is the total number of sampling points, and w_g is the group projection weight. By learning the offset Δp_{gk}, the sampling grid deforms, enabling it to closely fit the irregu lar boundaries of coal and gangue, thereby effectively suppressing background noise interfer ence.

3. Task-Aware Attention (π_C): Dynamic Channel Reweighting

Since coal-gangue recognition involves two tasks, namely classification for distinguishin

g coal and gangue and regression for locating bounding boxes, their requirements for featur e channels are different. The π_C module performs feature filtering along the channel dimensi on through a dynamic activation function known as Dynamic ReLU:

$$\pi_C(\mathrm{F})\cdot \mathrm{F} = \max(\alpha^k(\mathrm{F})\cdot \mathrm{F}_c + \beta^k(\mathrm{F})) \tag{10}$$

Where the hyperfunction $[\alpha, \beta]^T$ is learned through a global average pooling layer. It ca n dynamically activate or suppress specific channels based on the global context information of the input image, thereby balancing the conflicts in multi-task learning.

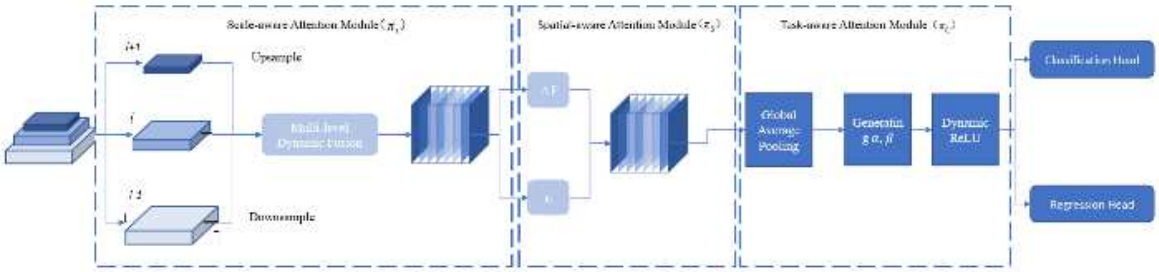


Figure 6. Architecture of the DynamicDCNv3Head

1.4 Coal-Gangue Pose Estimation Module

Oriented to the practical requirements of automatic coal-gangue sorting, to adapt to the rotational degrees of freedom of the robotic manipulator's end-effector and improve the stabi lity and accuracy of dynamic grasping operations, this paper introduces a pose estimation al gorithm based on DRD-YOLO11n object detection. This algorithm aims to calculate the pri ncipal axis rotation angle of the target coal and gangue, providing key geometric constraint information for the robotic manipulator's path planning and grasping pose alignment, thereby ensuring that the actuator engages the target at an optimal angle.

After the DRD-YOLO11n model accomplishes the rough localization of the coal-gangue targets, this section proposes a pose calculation method based on Minimum Area Bounding Rectangle (MABR) fitting. First, the Region of Interest (ROI) of the target is cropped fro m the original image according to the detected bounding box. Addressing the extremely low contrast between the coal blocks and the conveyor belt background, this study does not ad opt a global fixed threshold; instead, it utilizes Otsu's method within the ROI for local ada ptive binarization segmentation. By maximizing the inter-class variance within the local feat ure space, this strategy effectively suppresses the interference of belt dust and reflections, a ccurately extracting the edge information of the target. Subsequently, median filtering and m orphological operations are applied to eliminate contour holes, obtaining a target foreground mask with a complete topological structure, as shown in Figure 7.

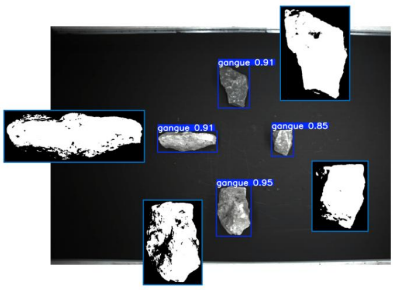


Figure 7. Schematic diagram of binarization for coal-gangue target regions

Following this, a topological structure analysis algorithm is applied to retrieve the maxi mum outer contour set of the target mask. Based on this contour, the MABR of the target is fitted using the rotating calipers principle. As illustrated in Figure 8, the MABR can effe ctively represent the principal geometric morphology of the irregular gangue. Through fittin g, the target's geometric center coordinates (x_c, y_c), the rectangle's width and height dimensio

ns (w,h), and the initial rotation angle θ can be calculated.

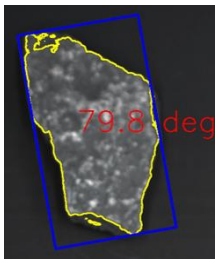


Figure 8. Schematic diagram of the bounding rectangle for coal-gangue contours

Due to potential definitional ambiguities in the initial angle θ returned by foundational libraries such as OpenCV (e.g., uncertainty regarding whether the angle is relative to the major or minor axis), it is necessary to perform a normalization correction on the initial an gle. This ensures the uniqueness of the grasping pose definition and uniformly establishes t he direction of the target's geometric major axis as the reference vector. If the width of the fitted rectangle is less than its height (i.e., (w,h)), the following correction is executed:

$$\theta = \theta_{init} + 90^{\circ}, \quad \text{if } w < h \tag{11}$$

Furthermore, in practical working conditions, for targets whose shapes approach a squar e (i.e., the aspect ratio r approaches 1), the direction of their principal axis exhibits high u ncertainty. This makes them highly prone to severe angle jumps near 0° or 90°. To elimin ate such instability, an aspect ratio threshold τ_r is introduced as a constraint. When the co ndition $r < \tau_r$ is satisfied, the target is considered to have no obvious directionality, and its pose angle is forcibly determined to be 0°, thereby preventing the mechanical gripper from performing invalid rotations.

Finally, considering that in continuous video stream detection, the single-frame pose esti mation values may experience high-frequency jitter due to the influence of image segmentati on noise and slight perspective changes. To further improve the temporal smoothness of the pose output, this paper combines the unique ID provided by the object tracking module and introduces the Exponential Moving Average (EMA) temporal filtering algorithm to smooth the angles:

$$\theta_t = \alpha \cdot \theta_{raw} + (1-\alpha) \cdot \theta_{t-1} \tag{12}$$

Where θ_t is the corrected and smoothed angle output of the current frame, θ_{raw} is the observed angle of the current frame, θ_{t-1} is the smoothed angle of the previous frame, and α is the smoothing coefficient, which adjusts the weight between the current observation an d historical information. This method effectively suppresses measurement noise and, while e nsuring the real-time response characteristics of the system, provides robust and reliable gra sping pose parameters for subsequent robotic manipulator control. The final visualization of the pose estimation is shown in Figure 9.

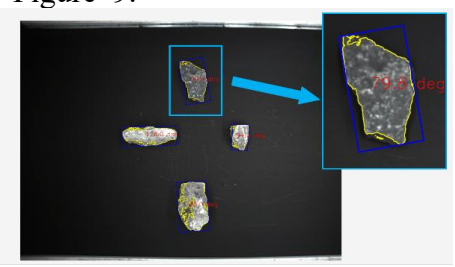


Figure 9. Visualization of the pose estimation module results

After obtaining the precise pose and center coordinates of the target coal and gangue, considering that the materials in the actual coal sorting site often present a distribution state

of high density and random stacking, forcibly executing a grasp relying solely on pose information makes the mechanical gripper highly susceptible to physical interference (collision) with adjacent materials during the descending and closing process. This can lead to grasp failure, target dropping, or even damage to the end-effector. To address this, based on pose estimation, this paper further proposes a dynamic geometric constraint-based anti-interference grasping mechanism to provide safe grasping targets for the robotic manipulator.

First, a dynamic safe workspace model for the mechanical gripper is established. In the physical world, the mechanical gripper requires a specific spatial range to execute a grasp. After extracting the target's minimum bounding rectangle, let the actual physical dimensions of the target gangue be w and h, the maximum physical opening of the gripper be O_{max}, the physical width of the gripper be W_g, and a safety margin Δs be introduced to absorb systemic mechanical errors. Combined with the rotation angle θ output by the target pose estimation, a dynamic adaptive "safe zone" model corresponding to the target pose is constructed. This restricted zone is also represented as a rotated bounding box, whose safe width W_{safe} and safe height H_{safe} are calculated as follows:

$$W_{safe} = \max(W_{real}, G_{open}) + M_{safe} \tag{13}$$

$$H_{safe} = \max(H_{real}, G_{width}) + M_{safe} \tag{14}$$

Subsequently, spatial geometric interference discrimination is performed. Intersection detection is conducted between the safe zone of the target coal or gangue and the true physical bounding boxes of all other objects within the current field of view. Based on the Separating Axis Theorem (SAT), the intersection status between different rotated rectangles is calculated in real time. If the true physical boundary of any other object intrudes into the "safe zone" of the target gangue, the target is deemed to have a high collision risk, and its safety status is marked as ungraspable. Conversely, if there is no interference within the restricted zone, it is marked as safely graspable. Figure 10 illustrates the visualization effect of the real yellow bounding boxes output by the vision system and the dynamic red safe zones.

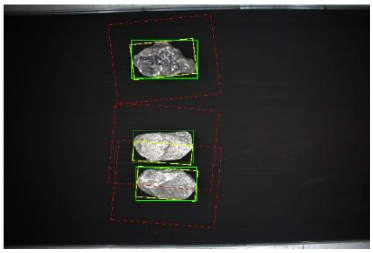


Figure 10. Visualization of the target's real boundary and dynamic safety zone

In summary, oriented towards the engineering application requirements of intelligent coal-gangue sorting, this paper proposes a visual perception-based coal-gangue detection and localization method for pose estimation and safety decision-making. This method not only achieves a synergistic optimization of detection accuracy, inference speed, and grasping pose accuracy by constructing a lightweight feature extraction network and introducing a robust planar pose estimation algorithm, but it also integrates a dynamic constraint-based anti-interference strategy. This effectively overcomes the engineering bottleneck wherein the mechanical gripper is highly susceptible to physical collisions under dense and disordered material streams. The proposed approach substantially enhances the operational robustness of the sorting system under complex working conditions, providing a comprehensive theoretical foundation and core technical support for the transition of high-performance intelligent coal-gangue sorting robots from pure vision algorithms to industrialized physical deployment.

2 Coal-Gangue Dataset Construction

To acquire authentic image data at the actual coal-gangue sorting site, a portable on-sit

e image acquisition platform was constructed. Due to sample imbalance, supplementary imag e collection was conducted in a laboratory setting designed to simulate the on-site environm ent. Through manual annotation and quality evaluation, a comprehensive coal-gangue dataset was ultimately established. Utilizing this dataset enables the objective evaluation of performa nce metrics such as the algorithm's detection accuracy and false detection rate, thereby guid ing the further improvement and optimization of the algorithm.

2.1 Image Data Acquisition

To capture the background interference under actual working conditions and obtain high -quality raw features, a portable data acquisition darkbox was independently designed. The a cquisition equipment is shown in Figure 11. In this study, approximately 14,135 raw images were collected on-site at the Dongrong No. 3 Coal Mine in Shuangyashan, Heilongjiang. L imited by factors such as the imbalanced distribution ratio of coal and gangue on-site, it is difficult to comprehensively represent the appearance variations of the targets relying solely on field data. Therefore, a standardized image acquisition platform was further established i n the laboratory to supplementarily collect 2,415 images. In the laboratory environment, by adjusting the light source direction and intensity, varying the conveyor belt speed, and alteri ng the placement of the coal and gangue, clearer and more balanced image samples were o btained using industrial cameras.

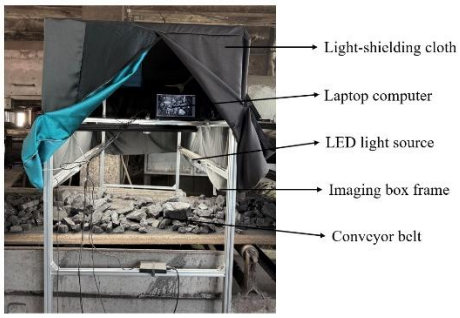


Figure 11. Data acquisition box

2.2 Data Annotation

All images were manually annotated pixel-by-pixel utilizing the Labelme tool. As illustr ated in Figure 12, "coal" and "gangue" were designated as the segmentation category labels. The annotation files were subsequently structured according to the YOLO dataset format to accommodate the training requirements of the subsequent detection and segmentation model s. To ensure sufficient model training and reliable performance evaluation, the dataset was p artitioned into a training set, a validation set, and a test set at a ratio of 8:1:1.

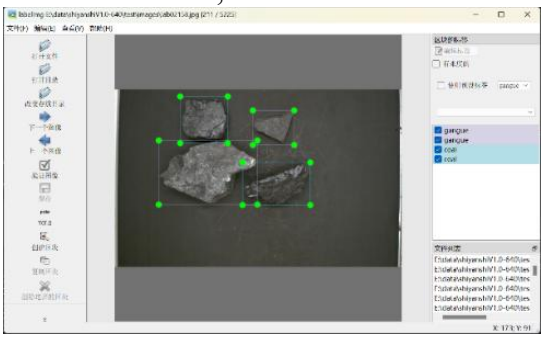


Figure 12. LabelImg labeling software interface

2.3 Data Augmentation

To enhance the model's robustness in complex gangue sorting environments, various dat a augmentation strategies were applied to the laboratory image samples, including 180° ima ge rotation, the addition of Gaussian noise, and Gaussian blur processing. This simulates th e on-site dust occlusion and illumination variations to a certain extent. The scale of the fin ally constructed dataset reached 25,818 images, encompassing coal and gangue images under multi-environment, multi-pose, and multi-interference conditions.

3 Experimental Design and Implementation

3.1 Experimental Environment and Parameter Settings

The deep learning method proposed in this paper is implemented based on the PyTorch 2.6 framework. The model is trained using an NVIDIA L20 GPU with 48 GB of video memory (VRAM). Anaconda is utilized to manage the experimental environment, with Python 3.10 as the programming language and PyCharm 2022.3 as the code editor. The hardware system operates on a 64-bit Windows 10 operating system, powered by an Intel(R) Xeon(R) Gold 6226R processor, and the CUDA version is 12.6.

During the training phase, the input image size is set to 640×640, and the total number of training epochs is set to 100. The Stochastic Gradient Descent (SGD) optimizer is adopted with an initial learning rate of 0.01, and Cross-Entropy Loss is selected as the loss function.

3.2 Evaluation Metrics

To evaluate the model performance, this paper adopts mean Average Precision (mAP), F1-score, number of parameters (Params), computational complexity (GFLOPs), and Frames Per Second (FPS) as the evaluation metrics.

Mean Average Precision (mAP) is adopted to measure detection accuracy, and its calculation involves Precision (P), Recall (R), and Average Precision (AP). Precision refers to the proportion of actual positive samples among those predicted as positive, while Recall refers to the proportion of correctly predicted positive samples among all actual positive samples. The F1 score is the harmonic mean of the two, used to comprehensively evaluate the model's balancing ability between precision and recall. The relevant calculation formulas are as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{15}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{16}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{17}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} \int_0^1 P(R) dR \tag{18}$$

Where TP, FP, and FN represent the number of correctly detected targets (True Positives), falsely detected targets (False Positives), and missed targets (False Negatives), respectively; N is the total number of categories; and $P(R)$ is the function of the Precision-Recall $(P-R)$ curve. The experimental results in this paper primarily refer to mAP@0.5 and mAP@0.5:0.95.

In addition, this paper adopts Params and GFLOPs to measure the spatial and computational complexity of the model, where Params represents the total number of model parameters (M), and GFLOPs stands for Giga Floating-point Operations. Meanwhile, FPS (Frames Per Second) is adopted to measure the actual inference speed of the model. Its calculation formula is as follows:

$$FPS = \frac{1000}{T_{pre} + T_{inf} + T_{post}} \tag{19}$$

Where the unit of 1000 is milliseconds (ms); and T_{pre}, T_{inf}, and T_{post} represent the pre-processing, inference, and post-processing times of the model, respectively.

3.3 Comparative Experiments

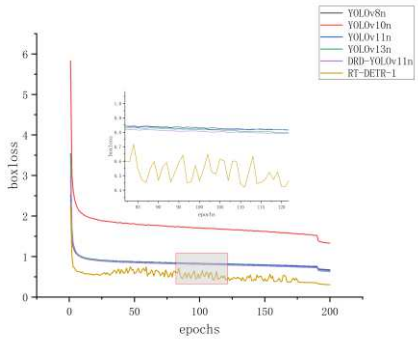


Figure 13. Box loss training curves of the YOLO series

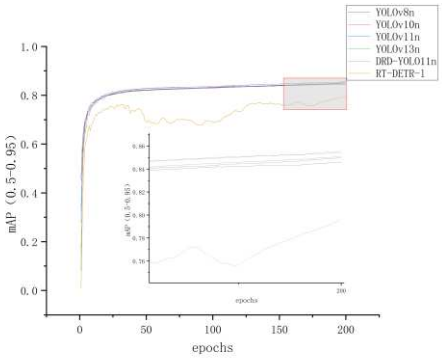


Figure 14. mAP (0.5-0.95) training curves of the YOLO series

To verify the performance of the DRD-YOLOv11n model, it is compared with object detection models including Faster R-CNN, SSD, YOLOv8n, RT-DETR-l, YOLOv11n, and YOLOv13n. The detection results based on the aforementioned constructed coal-gangue dataset are presented in Table 1. The training variations of box_loss and mAP@0.5:0.95 are illustrated in Figure 13 and Figure 14, respectively.

Table 1. Performance comparison of different models

Model	quantity of parameter	GFLOPs	mAP(0.5-0.95)/%	F1	FPS
FasterRCNN	137.1×106	370.2	51.4	90	23
SSD-VGG16	26.3×106	62.8	65.6	89	90
RT-DETR-l	32.8×106	108	76.9	92.4	24
YOLOv8n	2.69×106	6.9	84.6	94.5	138
YOLOv10n	2.70×106	8.4	85.1	94.9	86
YOLOv11n	2.59×106	6.4	84.9	94.5	105
YOLOv13n	2.46×106	6.4	85.4	94.7	53
YOLOv11n-DRD	2.43×106	7.3	85.6	95.1	114

The comparative analysis of Figures 14 and 15 reveals that among the evaluated models, DRD-YOLO11n exhibits the best performance. Its box_loss curve descends relatively rapidly and converges to a lower final value, while simultaneously maintaining the fastest growth trend and the highest detection accuracy on the mAP@0.5:0.95 metric.

The comparative experimental data recorded in Table 1 indicates that the DRD-YOLOv11n model proposed in this paper demonstrates strong performance in the coal-gangue detection task. It achieves the highest detection accuracy among all compared models, with an mAP reaching 85.6% and an F1-score of 95.1, surpassing the latest YOLOv13n and traditional algorithms such as Faster R-CNN in terms of overall performance. Regarding the metrics for lightweight design, although the computational complexity of DRD-YOLOv11n is 7.3 GFLOPs—a slight increase compared to the baseline model—its parameter count is further compressed, representing the lowest value among all evaluated models. Furthermore, while maintaining this extremely low parameter count, it achieves an inference speed of 114 FPS, which is faster than the real-time processing speeds of YOLOv10n and YOLOv13n. Integrating modules such as DySample and RD-SE into the model only marginally increases the comp

utational overhead, yet it substantially strengthens the feature extraction capabilities, fulfilling the dual requirements of real-time performance and precision for coal-gangue recognition at the sorting site.

3.4 Ablation Study

To verify the effectiveness of the various improvement strategies, an ablation study was conducted to comparatively analyze the impact of different modifications on the model's de tection performance. The checkmarks in the table indicate the application of the correspondi ng improvement strategy, and the results are presented in Table 2.

Table 2. Detection performance of different improvement strategies in coal gangue dataset

model	DySampl e	RD-SE	DynamicD CNv3Head	quantity of paramet er	GFLOPs	mAP(0.5-0.95)/%	FPS（GPU）
YOLOv11n	-	-	-	2,590,230	6.4	85	105
YOLOv11n -A	√	-	-	2594894	6.5	85.4	131
YOLOv11n -B	-	√	-	2686030	6.6	85,3	98
YOLOv11n -C	-	-	√	2417126	7.2	85,3	87
YOLOv11n -D	√	√	-	2698382	6.6	85,2	124
YOLOv11n -E	√	-	√	2427430	7.2	85,3	109
YOLOv11n -F		√	√	2428670	7.3	85,2	93
YOLOv11n-DRD	√	√	√	2438974	7.3	85,6	114

As indicated by the ablation study results in Table 2, the various improvement strategie s proposed in this paper optimize the model's performance to varying degrees. Selecting YO LOv11n as the baseline model, Experiment A, which incorporates the DySample operator, n ot only improves detection accuracy but also enhances the model's inference speed, raising i ts FPS from 105 to 131. Although Experiment B, which introduces the RD-SE module, res ults in a slight increase in computational complexity, it successfully elevates the mean Aver age Precision (mAP). Experiment C, adopting the DynamicDCNv3Head, serves as the core f or realizing a lightweight model, effectively reducing the parameter count from 2.59 M to 2.41 M.The finally proposed DRD-YOLOv11n model integrates the three aforementioned str ategies, achieving an optimal balance in performance. Compared with the baseline model, th e mAP of the proposed model increases by 0.6%, reaching 85.6%, while its parameter coun t drops by approximately 5.8% to merely 2.438 M. Furthermore, alongside a marginal incre ase in computational complexity, the inference speed is synchronously improved to 114 FPS. These experimental results validate the synergistic effect among the modification modules, allowing the model to acquire stronger feature capturing capabilities and a faster real-time r esponse speed under the premise of accomplishing parameter compression.

3.5 Comparison of Detection Results

To more intuitively demonstrate the accuracy of the DRD-YOLOv11n model in recogni zing moving coal-gangue streams on a conveyor belt, it is compared with YOLOv8n, YOL Ov11n, YOLOv13n, and RTDETR-l on the dataset. The detection results are illustrated in F igure 15.The three columns of images, from left to right, are from the Jixian Coal Mine sit e, the Jixian Coal Mine site, and the laboratory, respectively.

As illustrated in Figure 15, Faster R-CNN (ResNet50) exhibits significant classification confusion, tending to simultaneously predict the same target as both coal and gangue. Regar ding localization accuracy, both Faster R-CNN and SSD-VGG16 underperform compared to the YOLO series, primarily manifested by loose bounding box regression. In contrast, DRD-YOLOv11n not only eliminates category conflicts but also significantly enhances localization accuracy, demonstrating the optimal overall performance among all evaluated models.

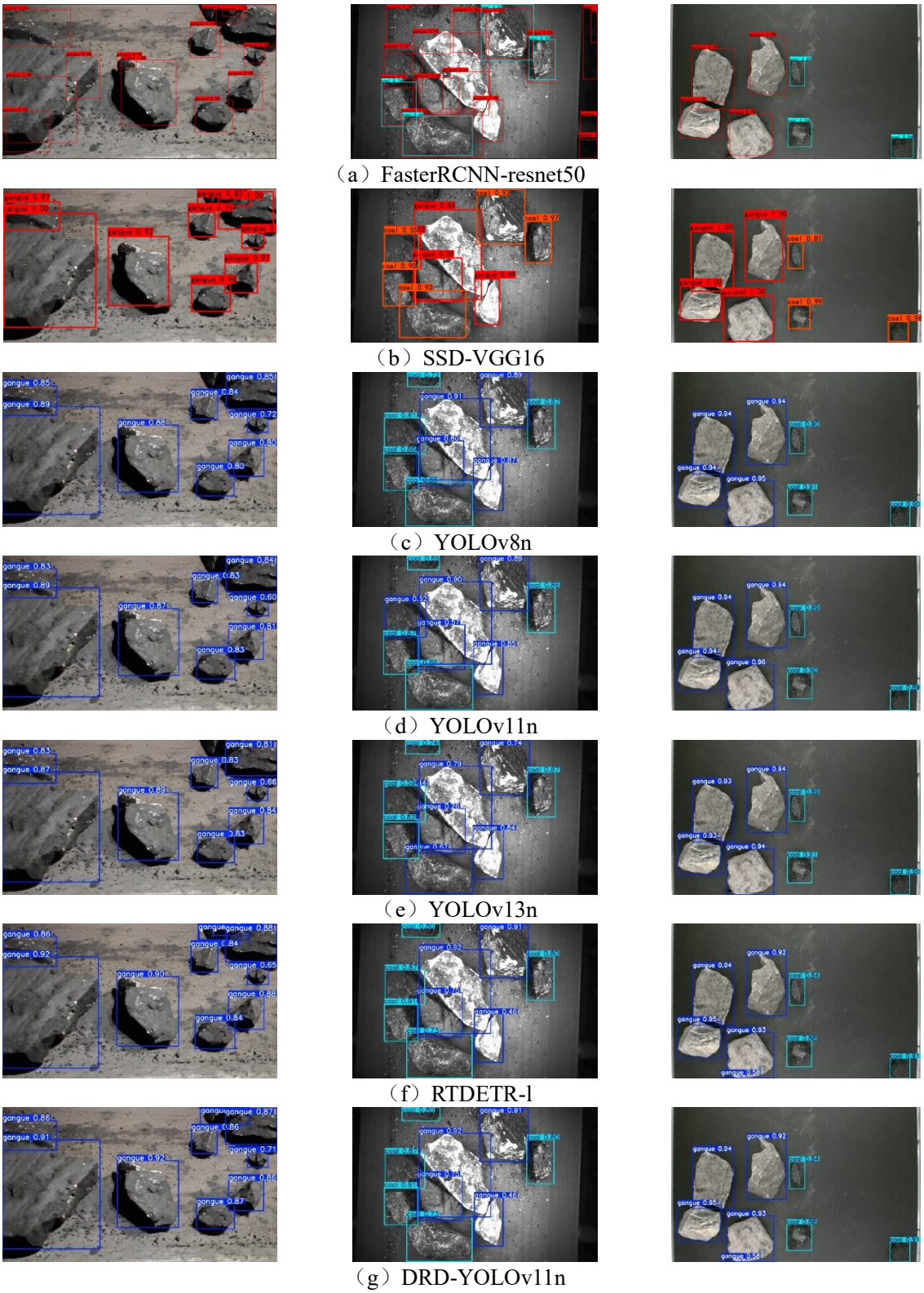
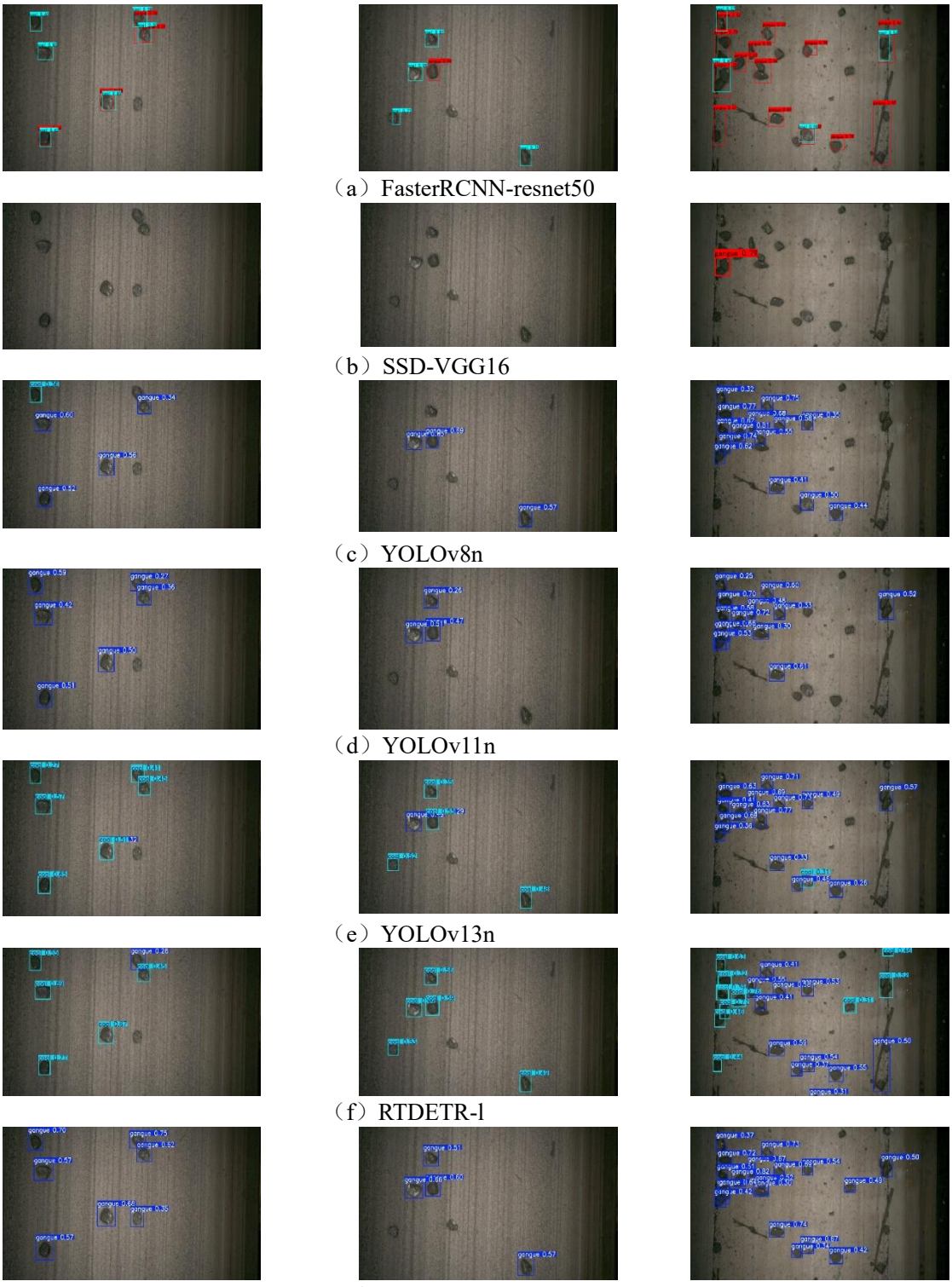


Figure 15. Comparison of detection results among different models

3.6 Generalization Validation

To verify the generalization ability of the DRD-YOLOv11n model across different datas ets, it is compared with YOLOv8n, YOLOv11n, YOLOv13n, and RTDETR-l on the public DsCGF dataset. The detection results are illustrated in Figure 16. As illustrated in Figure 1 6, Faster R-CNN (ResNet50) suffers from classification ambiguity, tending to simultaneously predict the same target as both coal and gangue, which leads to a high false detection rat e. SSD (VGG16) exhibits severely limited detection capabilities, managing to detect only is

olated targets across multiple sample groups, thereby demonstrating the poorest generalizatio n performance. In contrast, DRD-YOLOv11n exhibits the optimal overall performance amon g the YOLO series, effectively circumventing the aforementioned issues while maintaining h igh detection accuracy.3.7 Validation of the Impact of Pose Estimation on Grasping Success Rate.



（a）FasterRCNN-resnet50

（b）SSD-VGG16

（c）YOLOv8n

（d）YOLOv11n

（e）YOLOv13n

（f）RTDETR-l

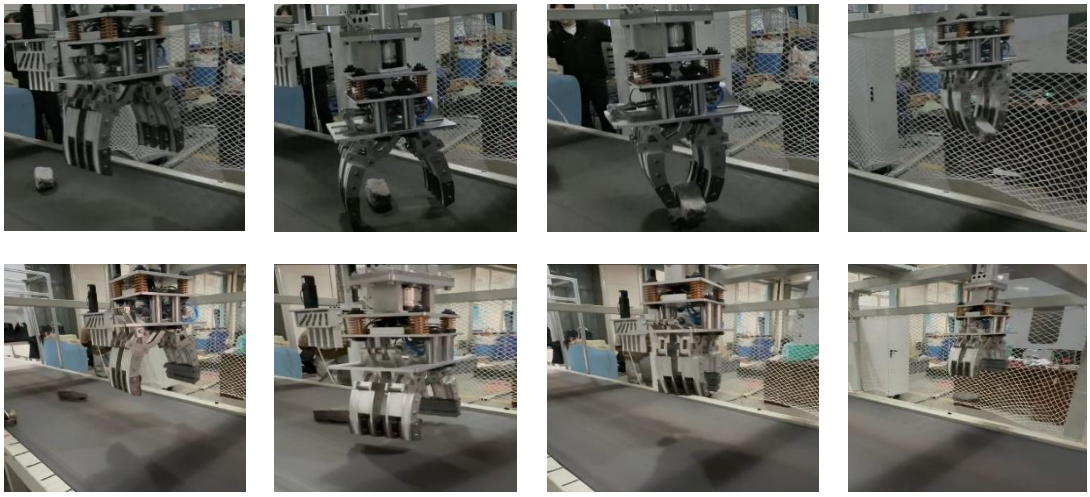
（g）DRD-YOLOv11n

Figure 16. Comparison of generalization detection results among different models

To verify the enhancement effect of the pose estimation module integrated into the DR D-YOLOv11n model on the actual grasping success rate of the sorting robot, 10 elongated coal-gangue samples were randomly selected for comparative testing on a sorting prototype

experimental platform. The experiment evaluated two strategies: a control group relying sole ly on bounding box center localization, and an experimental group incorporating pose estim ation for end-effector pose compensation. The grasping test results are presented in Table 3 and Figure 17.

Observation of experimental phenomena and data analysis reveal that the experimental group incorporating the angle estimation method exhibits a significant advantage when handl ing elongated and irregularly shaped coal-gangue targets, effectively overcoming grasping fai lures caused by the irregular shapes of the materials. Furthermore, comparative data demons trates that after introducing pose estimation, the grasping success rate of the mechanical gri pper increases from 60% to 90%, and the overall sorting reliability is significantly enhance d.



(a) Preparation phase (b) Descending phase (c) Grasping phase （d）Gangue placement phase

Figure 17. Schematic diagram of rotational gripping

Table 3. Comparison of grasping success rates of sorting robots under different strategies

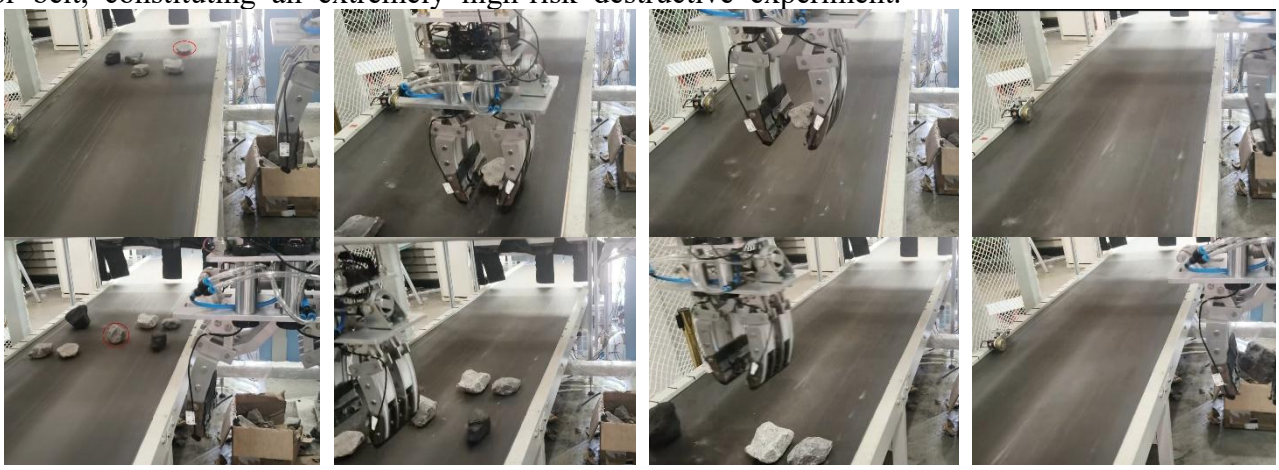
Group	Number of failed grasps		Number of successful grasps
	Number of missed grasps	Number of dropped items	
Experimental group	0	1	9
Control group	3	1	6

3.8 Validation of Pose Estimation and Anti-Collision Grasping under Simulated On-Site Conditions

To verify the enhancement effect of the dynamic geometric anti-collision strategy integrated into the DRD-YOLOv11n model on the actual grasping safety of the sorting robot, 10 grou ps of coal-gangue samples, simulating the dense stacking and minimal spacing of the on-sit e environment, were randomly arranged on the sorting prototype experimental platform for t esting. Considering that in an actual physical environment, forcibly executing a descending action with the anti-interference mechanism disabled is highly likely to cause collisions bet ween the rigid mechanical gripper and adjacent gangue or coal blocks, this could lead to ir reversible damage to the gripper equipment or overturn the conveyor belt. Therefore, based on hardware protection principles, this experiment did not establish a physical control group without the anti-collision strategy. Instead, it focused on evaluating the system's "proactive a voidance decision accuracy" and "safe target grasping success rate" under extreme dense co nditions after introducing the anti-collision mechanism.

Consolidating the experimental observation records and data calculation results, it is evi dent that when targeting randomly stacked coal-gangue objects simulating the on-site enviro nment, the dynamic geometric anti-collision strategy proposed in this paper exhibits excellen

t self-protection capability and decision reliability. Facing dense arrays with an extremely hi gh risk of interference, the system calculates the dynamic safe zone in real-time based on t he Separating Axis Theorem (SAT). It accurately identifies and proactively avoids high-risk targets lacking sufficient space for the gripper to descend, allowing them to pass safely. Si multaneously, after eliminating interference risks, the system ensures precise pose alignment and interference-free grasping for safe targets with ample surrounding space (as shown in Fi gure 18).Experimental data indicates that across the 10 groups of dense tests, the system ac hieved 100% collision avoidance and a 0% physical accident rate. The continuous operation al robustness and hardware safety of the entire sorting system under complex working scena rios have been substantially strengthened. It should be noted that, for equipment safety cons iderations, a conventional control group without the anti-collision strategy was not establishe d in this experiment. This is because, in the absence of this protective mechanism, the mec hanical gripper is highly prone to severe rigid interference with irregular stones during the descending phase. Such forced rigid descent not only risks burning out the end-effector mot or due to instantaneous stall overload and causing irreversible deformation to the mechanical structure of the gripper, but could also directly puncture or damage the underlying convey or belt, constituting an extremely high-risk destructive experiment.



(a) Preparation phase (b) Descending phase (c) Grasping phase （d）Gangue placement phase

Figure 18. Illustration of safe rotation and grasp after collision-avoidance decision

4 Conclusion

Oriented towards the environmental characteristics of coal-gangue sorting operation sites and the practical requirements for efficient and precise grasping, this paper designs a coal-gangue detection and pose estimation scheme based on DRD-YOLOv11n. By conducting ad aptive lightweight adjustments and corresponding functional expansions to the model architec ture, this study optimizes the model's detection performance and overcomes the related issue s of grasping pose alignment for the robotic manipulator. The core research findings and rel evant conclusions are summarized as follows:

(1) Following the optimization of the model structure from multiple dimensions, both t he detection accuracy and inference speed of the model have been improved. First, the cont ent-aware dynamic upsampling operator DySample is introduced into the feature fusion netw ork, replacing the original interpolation method. DySample can accomplish dynamic resampli ng based on feature content at an extremely low computational cost, improving the recovery quality of coal and gangue edge features in low-illumination and motion-blurred scenarios, and circumventing the rigid limitations of traditional methods. Second, the Residual Depthwi se Squeeze-and-Excitation (RD-SE) feature enhancement module is embedded at the front en d of the detection head. Relying on residual depthwise convolution and the SE channel atte

ntion mechanism, it enhances the model's sensitivity to the fine texture features of coal and gangue while suppressing underground background noise. Third, the detection head is reconstructed into a DCNv3-based dynamic detection head. By integrating a unified attention mechanism and relying on the geometric deformation capabilities of the DCNv3 operator, it resolves the precise localization problem for coal and gangue targets with large morphological differences and significant scale fluctuations. The aforementioned improvements not only compress the computational overhead but also enhance the model's adaptability under complex working conditions.

(2) Relying on the integrated pose estimation module, precise geometric constraints can be provided for the automated mechanical grasping of coal and gangue. This module employs minimum bounding rectangle fitting technology to accurately calculate the principal axis rotation angle of the target coal or gangue. Coupled with angle normalization and Exponential Moving Average (EMA) temporal filtering methods, this module can stably output reliable pose parameters, providing a core reference basis for the path planning and pose adjustment of the mechanical gripper. Simultaneously, addressing the engineering problem of mechanical interference prone to occur in coal sorting under complex working conditions, this module introduces a dynamic safe zone interference discrimination mechanism based on the Separating Axis Theorem (SAT). It conducts proactive avoidance of high collision risk targets to achieve self-protection of the end-effector. This function improves the grasping stability and precision of the robotic manipulator when dealing with coal and gangue in various poses, ensuring the efficient and safe execution of sorting operations.

(3) Comprehensive experiments and comparative analysis reveal the technical advantages of the dynamic lightweight framework in the intelligent coal-gangue sorting system. Experimental results demonstrate that, under the premise of a 6% reduction in parameter count, the constructed DRD-YOLOv11n model maintains an mAP@0.5:0.95 of 85.6% and achieves a real-time processing speed of 114 FPS, realizing a synergistic optimum of accuracy, speed, and model complexity. This confirms that integrating modules such as DySample dynamic feature resampling, RD-SE residual depthwise feature enhancement, and the DCNv3-based dynamic detection head can effectively overcome difficulties such as incomplete target feature extraction and ambiguous boundary localization under harsh on-site working conditions in coal mines. Compared to traditional YOLO series models and two-stage detection algorithms, this architecture not only lowers the deployment requirements for edge computing platforms but also provides reliable technical architecture support for the efficient and intelligent sorting of coal and gangue through the parallel calculation of pose and detection information, possessing significant practical engineering application value.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions Statement

Author contributions Kaiyun Chen was responsible for project management, Minhang Zhu wrote the methodology and drafted the initial manuscript, Boyang Yin supplemented the methodology, conducted the experiments, and wrote the manuscript, and Yilin Ge and Yanwei Wang provided guidance on the methodology. All authors contributed to the writing and review of the manuscript.

References

[1] WANG Guofa, Ren Shihua, Pang Yihui, et al. Development achievements of China’s coal industry during the 13th Five-Year Plan period and implementation path of "dual carbon" target[J]. Coal Science and Technology, 2021,49(09):1-8.

[2] DONG Xingling, Dong Shuning, Wang Bao, et al. Hydraulic conductivity of geosynthetic clay liners to loess leachate[J]. Journal of China Coal Society, 2018, 43(1): 228-235.

[3] GUO Kang, Liu Zhicheng, Yi Weixun, et al. Current situation and prospects of coal gangue comprehens ive utilization[J]. Coal, 2025, 34(10): 96-104.
[4] KANG Wang, Tian Xiangqin, Zhao Kang, et al. Characteristics and risk assessment of heavy metal poll ution in a coal gangue stockpile[J]. China Environmental Science, 2026, 46(5): 2610-2617.
[5] WANG Aimin, Wang Tao. Application of intelligent separation and grinding system[J]. Mechanical Engin eer, 2022(6): 65-68.
[6] ZHAO Yiding,HE Xiaoming. Recognition of coal and gangue based on X-ray[C]//Applied mechanics and materials. Trans Tech Publications Ltd, 2013, 275: 2350-2353.
[7] KONG Li, Li Hong, Xu Shuhong, et al. The on-line identification and separation system for coal gangu es based on double energy γ-ray transmission[J]. Journal of Huazhong University of Science and Techno logy, 1997(10): 108-109.
[8] WANG Yanwei, Tao Wenbin, Chen Kaiyun, et al. Coal gangue recognition technology based on SegFor mer-CG[J]. Journal of China Coal Society, 2026, 51(4): 2771-2782.
[9] CAO Shuai, Dong Lihong, Deng Fan, et al. A small object detection method for coal mine underground scene based on YOLOv7-SE[J]. Industry and Mine Automation, 2024, 50(3): 35-41.
[10] ZHANG Chao, Xiao Shuaiyi, Li Lianpo, et al. Recognition method of coal and gangue based on image gray level co-occurrence matrix[J]. Coal Technology, 2025, 44(10): 239-243.
[11] LI Juanli, Wei Dailiang, Li Bo, et al. Improved SSD coal-gangue rapid sorting model based on lightwei ght deep learning[J]. Journal of Northeastern University (Natural Science), 2023, 44(10): 1474-1480.
[12] XUE G H,LI S X,HOU P,et al. Research on lightweight Yolo coal gangue detection algorithm based o n resnet18 backbone feature network[J]. Internet of Things,2023,22:100762.
[13] LI Deyong, Wang Guofa, Guo Yongcun, et al. Image recognition method of coal and gangue in comple x working conditions based on CFS-YOLO algorithm[J]. Coal Science and Technology, 2024, 52(6): 226 -237.
[14] WEI Xiaolong, Wang Fangtian, He Dongsheng, et al. Coal gangue image recognition model based on C SPNet-YOLOv7 target detection algorithm[J]. Coal Science and Technology, 2024, 52(S1): 238-248.
[15] LI Na, Wu Sibo, Feng Weihao, et al. Lightweight coal-gangue identification method based on FNF-YOL O[J]. Computer Engineering and Design, 2025, 46(12): 3413-3421.
[16] ZANG Jinwang, Wang Jingzheng, Yang Shengli, et al. Infrared image recognition of coal and gangue wi th difficult identification under liquid intervention based on HDB-YOLO algorithm[J/OL]. Journal of Min ing & Safety Engineering: 1-13 [2026-05-06].
[17] LIU W, LU H, FU H, et al. Learning to upsample by sample: DySample[C]//Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV). 2023: 16512-16522.
[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR). 2018: 7132-7141.
[19] YU Dawei, Shao Ming, Cui Meng, et al. Research on multi-scale coal-gangue identification based on S T-YOLOX-S[J]. Nonferrous Metals (Mineral Processing Section), 2025(10): 63-72.
[20] DAI X, CHEN Y, XIAO B, et al. Dynamic Head: Unifying object detection heads with attentions[C]//Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 737 3-7382.
[21] WANG W, DAI J, CHEN Z, et al. InternImage: Exploring large-scale vision foundation models with def ormable convolutions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR). 2023: 14408-14419.