Physicochem. Probl. Miner. Process., 61(4), 2025, 205702

http://www.journalssystem.com/ppmp

Physicochemical Problems of Mineral Processing

ISSN 1643-1049
© Wroclaw University of Science and Technology

Research on image recognition method of coal gangue under complex working condition environment

Deyong Li[1,2,3,4,5], Yuanbo Gao[1,2,3], Yuhao Yang[1,2,3], Sihang Zhang[1,2,3], Shuang Wang[1,2,3], Yongcun Guo[1,2,3]

[1] State Key Laboratory of Digital and Intelligent Technology for Unmanned Coal Mining, Anhui University of Science and Technology, Huainan 232001, China

[2] Mining Intelligent Technology and Equipment Provinces and Ministries jointly build a Collaborative Innovation Center, Anhui University of Science and Technology, Huainan 232001, China

[3] Anhui Key Laboratory of Mine Intelligent Equipment and Technology, Anhui University of Science & Technology, Huainan 232001, China

[4] Department of Electrical Engineering, Tsinghua University, Beijing,100084, China

[5] Anhui Zhongke Optic-electronic Color Sorter Machinery Co., Ltd. Hefei, 231299, China

Corresponding author: 2052146289@qq.com (Yuanbo Gao)

Abstract: In response to the issues of low coal gangue recognition accuracy, missed detections, and false detections due to the complex working conditions of low illumination, high blur, and occlusion in underground coal mines, an ELS-YOLO coal gangue detection model is proposed. The ELS-YOLO model is based on the YOLOv10s model. Firstly, the EfficientNetV1 module is introduced into the backbone network, which expands the dimensions proportionally, balancing the scaling of the network's depth, width, and resolution, thus improving the performance of the convolutional neural network. Secondly, the LSGE attention module is introduced in the neck network, which significantly enhances the image feature extraction quality and efficiency through strategies such as local feature enhancement, multi-scale information fusion, and spatial feature aggregation. Finally, the SPPELAN module, which has an efficient local aggregation network, is selected to improve the model's detection performance when handling targets of different sizes. Experimental results show that the average detection accuracy of the ELS-YOLO model reaches 89.6%, which is 3.0% higher than the YOLOv10s model; the average detection speed is 81.30 FPS, fully meeting the real-time detection requirements of coal gangue in underground coal mines. Compared to YOLOv5s, YOLOv7-Tiny, YOLOv8s, and YOLOv9s, the ELS-YOLO model demonstrates the strongest adaptability to complex coal mine environments and the best overall detection performance, providing technical support for the intelligent and efficient sorting of coal gangue.

Keywords: YOLOv10s, complex working conditions, coal gangue recognition, attention mechanism

1. Introduction

Coal is one of the main energy sources in the current society, occupying about 66% of the total energy production and consumption (Wang et al., 2024; Yuan et al., 2018). In the process of coal mining, a large amount of gangue is often mixed, which will lead to a reduction in the overall calorific value of coal, thus affecting the combustion efficiency and economic value of coal. At the same time, gangue mixed into coal will increase the total weight and volume of coal, which increases the transportation cost. Moreover, in the combustion process, gangue will produce a significant amount of harmful gases, which significantly pollute the environment (Li et al., 2010; Wang et al., 2019). Consequently, gangue sorting (Chen et al., 2022) plays a crucial role in enhancing coal quality, lowering costs, and minimizing environmental impact.

The common methods of coal gangue identification include ray identification method, multispectral identification method (Hu et al., 2022; Rui et al., 2022) and image identification method (Alfarzaeai et

DOI: 10.37190/ppmp/205702

al., 2023; Wang et al., 2022). The ray identification method, i.e., the x-ray (Dong et al., 2016) or γ-ray (Xue et al., 2023; Kong et al., 1997) identification method is a method that utilizes the ray's penetration ability of the material and the interaction with the material to identify the coal and gangue. Material interaction characteristics of sorting technology, the technology is highly efficient and adaptable, but the ray source has a certain radiation, will pose a threat to the safety of technicians. Multi-spectral identification method is a kind of technology that utilizes the absorption, reflection and transmission characteristics of the material to different wavelengths of light for material identification and sorting. The method has high precision and is environmentally friendly, but it is not suitable for gangue sorting because it is more sensitive to environmental conditions and difficult to recognize low-contrast materials.

With the significant progress of computer vision technology, image recognition method has the advantages of high efficiency, non-contact, non-hazardous, and high accuracy, so that the image recognition method based on deep learning (Du et al., 2021; Xing et al., 2022) has a great prospect in the field of gangue recognition, and has gradually become the focus of scholars' attention. Among them, Li et al. (2022) proposed a coal gangue detection and recognition algorithm based on deformation convolution YOLOv3, which utilizes deformation convolution, multi-K-means clustering result averaging and data enhancement techniques to construct an efficient network model, which effectively improves the accuracy and efficiency of detecting small-size coal and gangue. Zhang et al. (2024) proposed an image recognition method based on YOLOv4 model. YOLOv4 model introduces a GAN-based dual-attention data enhancement strategy, which is able to produce more realistic gangue images and enhance the small-size dataset by generating rich and varied synthetic samples, thus improving the target detection performance of coal and gangue. Xue et al. (2023) proposed an improved dual-scale ResNet18-YOLO gangue detection algorithm, which utilizes the unstructured pruning theory to prune the redundant network weights, which reduces the network complexity and improves the detection speed. Yang et al. (2023) introduced a new Contextual Transformer Network (COTN) module on the basis of the YOLOv7 model, replacing the convolution in ResNet, which improves the output feature expressiveness, recognition speed, and accuracy, and at the same time minimizes the model parameter volume and complexity increase. Lai et al. (2022) proposed an improved Mask R-CNN combined with multispectral imaging for coal gangue instance segmentation method, which can accurately locate the gangue and get its relative size. Lei et al. (2021) enhanced the ability of coal gangue feature extraction by improving the network structure of deepening the YOLOv3 model, which improved the detection accuracy. Wang et al. (2020) used Tengine deep learning framework in the VGG16 deep learning model of the gangue image training recognition, to improve the network recognition accuracy. Cao et al. (2022) proposed a GoogLenet deep learning network based on the gangue recognition method, using the Inception model, and through the migration of learning to share the weights of the convolutional layer of the trained model and bias, to achieve a high classification accuracy. Nguyen and Hoang, (2025) used Lightweight-Efficient Aggregating Fusion as well as other blocks and techniques to enhance multiscale feature extraction while reducing complexity for small object detection in dense and varying backgrounds. Das et al. (2025) on the YOLOv8 architecture, proposed a domain adaptive framework YOLO-D that combines low light enhancement techniques and multi-scale domain adaptation to improve target detection under difficult lighting conditions.

The above research has made a significant contribution to improving the performance of target detection algorithms, improving detection accuracy and also making the model lightweight to improve the detection speed, but when carrying out gangue recognition, the field situation is more complex, and their research has rarely considered the detection of gangue in the complex working environment of underground coal mines. In this paper, for the complex environment of low illumination, motion blur, and occlusion, a gangue recognition method based on ELS-YOLO model is proposed, which can realize accurate real-time monitoring of gangue under complex working conditions.

2. ELS-YOLO model

2.1. Introduction to YOLOv10s

YOLOv10s (Wang et al., 2024) is the latest iteration of the YOLO (You Only Look Once) series, which inherits the core concepts of the YOLO model, and through a series of optimizations and innovations, significantly improves the speed, accuracy, and lightweight characteristics of the model, making it a

high-efficiency target detection model that performs better in complex environments and is more adaptable. While the traditional YOLO model achieves a better balance between accuracy and speed, YOLOv10s further improves computational efficiency by introducing an advanced convolutional neural network (CNN) architecture, optimized activation functions and convolutional layers. This enables YOLOv10s to run efficiently on low-computing-power devices (e.g., embedded systems, mobile devices, etc.), while still maintaining high accuracy under conditions of limited computing resources. In terms of the optimization of the feature extraction module, YOLOv10s adopts innovative techniques such as Depthwise Separable Convolution (Howard, 2024) and Sparse Convolution (Graham and Van der maaten, 2017), which aim to effectively reduce the computational complexity while retaining more valuable feature information. Through these technological innovations, YOLOv10s is able to achieve higher accuracy when dealing with small objects, complex backgrounds, and multi-scale targets, especially in high-resolution images, to better capture details and improve the accuracy and reliability of target detection. In the field of gangue recognition, there have been many improvement schemes based on YOLO, and significant results have been achieved. In this paper, based on YOLOv10s model, we further improve the model for the complex conditions in the actual working environment (such as low light, blurring, and occlusion, etc.) to enhance its detection effect in extreme environments. The structure of YOLOv10 is shown in Fig. 1.

2.2. EfficientNetV1 backbone network module

Convolutional Neural Networks (ConvNets) (Liu et al., 2022) are typically designed with a specific resource budget, and their accuracy can be improved by utilizing more computational resources. EfficientNetV1 (Tan and Le, 2019) is a highly efficient neural network architecture that focuses on optimizing the use of computational resources while maximizing model accuracy. It achieves this by balancing three key factors: the network's width (number of channels), depth (number of layers), and resolution (input image size). The architecture leverages a technique known as the "composite scaling method," which adapts the scaling of these three dimensions—width, depth, and resolution—based on model size and the amount of training data. This approach helps to achieve the best possible performance by tailoring the model's configuration to the available hardware, resources, and datasets. The difference between this scaling method and the conventional method is illustrated in Fig. 2, as (a) shows an example of the baseline network; (b)-(d) are conventional scaling, which increases only one dimension of the network width, depth, or resolution; and (e) is the composite scaling method proposed above, which scales all three dimensions uniformly at a fixed ratio. Intuitively, the composite scaling approach is justified because when the input image is large, the network needs to increase the number of layers to expand the receptive field and requires more channels to capture the detailed information in the large image. To enhance the model's performance further, EfficientNetV1's design includes a new baseline network created through neural architecture search. This baseline is then expanded into a series of EfficientNetV1 models, which have been shown to outperform conventional ConvNets in terms of both accuracy and computational efficiency.

2.3. LEGE attention mechanism

In complex environments, gangue usually suffers from problems such as occlusion or low visibility, making it difficult to be detected and unable to recognize features. In order to solve the above problems, this paper cites the formation of a new LSGE attention mechanism by combining SpatialGroupEnhance attention mechanism (Li et al., 2019) and LSKAttention attention mechanism (Lau et al., 2024).

The structure of LSGE is shown in Fig. 3, in the LSGE module, SpatialGroupEnhance and LSKAttention are sequentially connected through the forward method. (a) shows the structure of the SpatialGroupEnhance module and (b) shows the structure of the LSKAttention module. SpatialGroupEnhance first processes the input feature maps, and the output is used as the input for LSKAttention. LSKAttention further enhances the spatial information of the feature maps and finally returns the processed results.

The SpatialGroupEnhance Attention Mechanism Is a lightweight spatial attention enhancement module, and the core idea is to perform operations within each semantic group, thus realizing feature enhancement, which is similar to the idea of group convolution and can effectively reduce the

computational overhead. Functionally, the SGE module generates an attention mask by comparing the similarity between global statistical features and local feature descriptors. Unlike other attention mechanisms, the attention factor of SGE is completely determined by the similarity of global features and local feature descriptors in the group, so it is very simple in design and has low computational overhead.

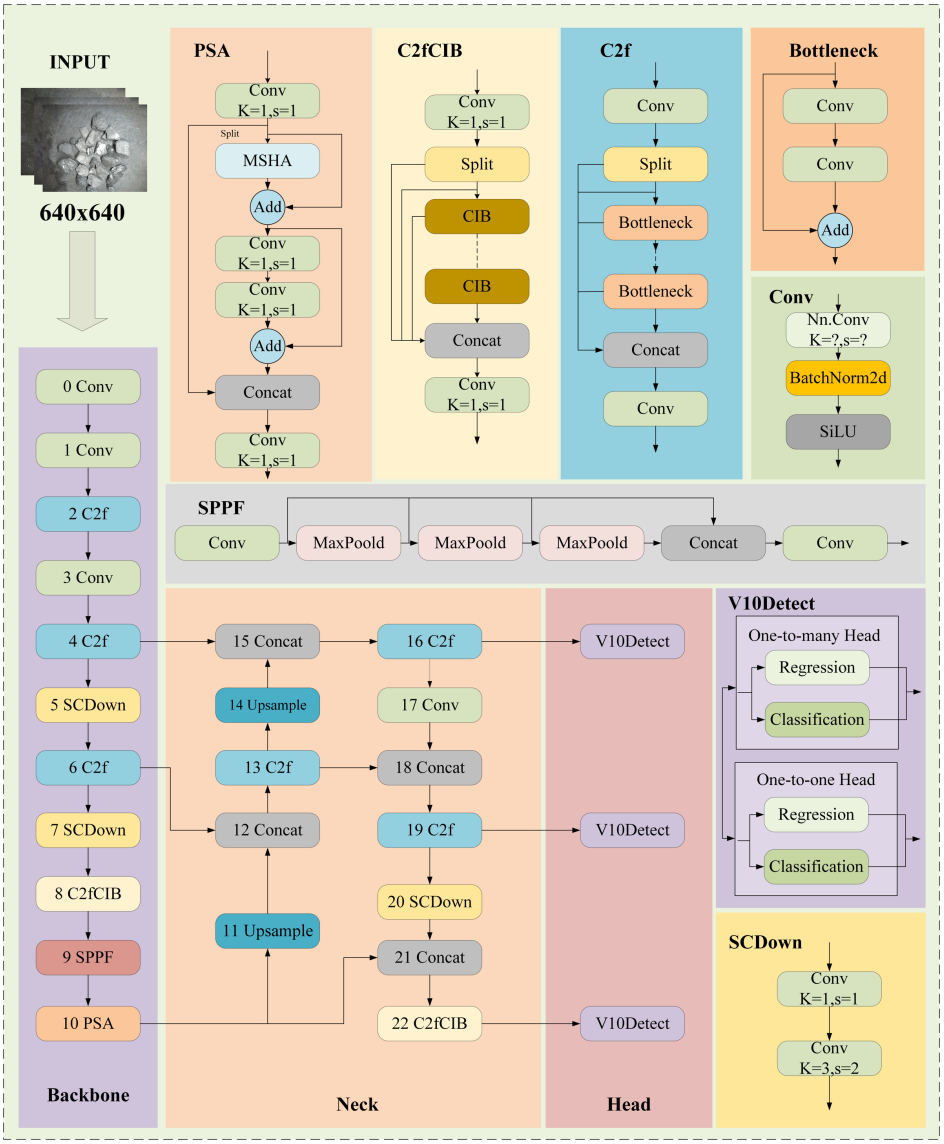


Fig.1. YOLOv10s structure diagram

LSKA module, known as Large Separable Kernel Attention (LSKA) module, is an improved version of Large Convolutional Attention (LKA) module designed for visual attention networks (VANs). The module aims to address the problem of exponentially increasing computational and memory costs associated with the use of large convolutional kernels in traditional LKA as the size of the convolutional kernel increases. The key innovation of the LSKA module is to decompose the original 2D deep convolutional layers into a combination of one-dimensional horizontal and vertical convolutions, which significantly reduces the computational complexity and memory consumption without sacrificing performance.

The LSGE module divides the feature map Into multiple groups by the SGE module and processes each group independently, extracting statistical information using global average pooling, which, after normalization and parameter tuning, enhances the local response of the feature map and helps the network to better capture image details. Next, the LSKA attention mechanism uses convolution kernels of different sizes (e.g., 5x5, 7x7, etc.) and dilation convolution operations to extract multi-scale information, which improves the recognition of targets of different sizes. Overall, the LSGE attention

mechanism significantly improves the effectiveness and efficiency of image feature extraction by enhancing local features, fusing multi-scale information, and aggregating spatial features, providing strong support for computer vision tasks.

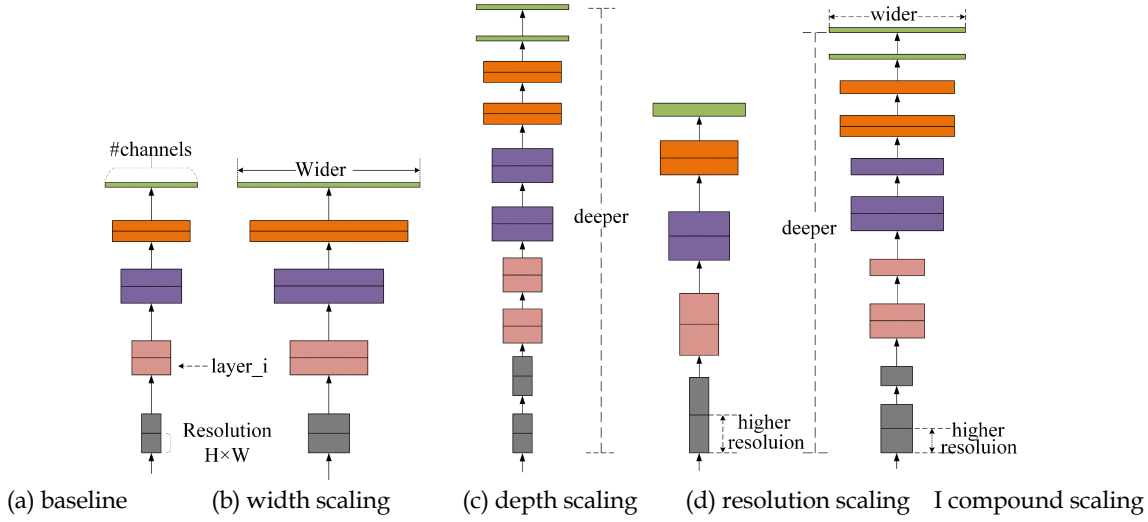


Fig. 2. Comparison of different model scaling

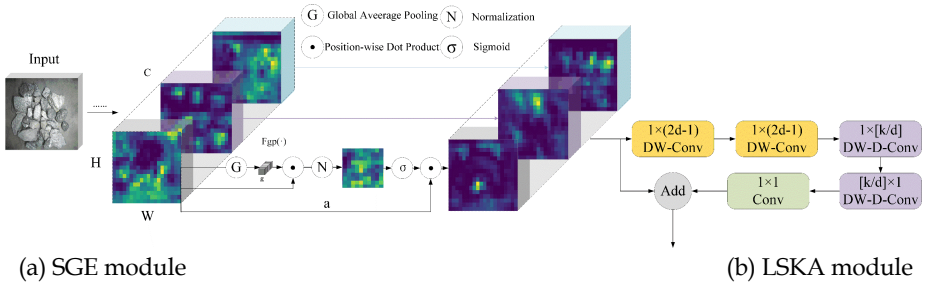


Fig. 3. LSGE structure diagram

2.4. SPPELAN module

The SPPELAN module (Wang et al., 2024) employs a combination of convolutional and maximum pooling operations as shown in Fig. 4. The core idea of the module is to reduce the number of channels of the input data by an initial convolutional layer (cv1), followed by the parallel application of three different configurations of maximal pooling layers (cv2, cv3, and cv4) on top of this, which serve to adjust the size of the sensory field by different paddings and step sizes. The information extracted from each pooling layer has a different spatial scale. Next, the outputs of all pooling layers are combined (concatenated) together to form a feature map containing multi-scale information. Finally, a convolution (cv5) is used to transform the fused feature map into the desired number of output channels.

In short, the SPPELAN module extracts different spatial contextual information of the input data through parallel multiscale pooling operations, and then fuses this information to construct a richer feature representation. This design allows the network to simultaneously focus on details at different scales when processing images, thus improving its performance.

2.5. The ELS-YOLO model

The ELS-YOLO model is mainly improved on the basis of the YOLOv10s model, which consists of three parts: replacing the backbone with the EfficientNetV1 module, adding the LSGE attention mechanism, and replacing the original SPPF with the SPPELAN module, and its structure is shown in Fig. 5.

The backbone network part uses the EfficientNetV1 module to replace the original backbone part, and the main innovation of the EfficientNetV1 module is that it adopts the composite scaling method, which scales the depth, width, and resolution of the network in a balanced way by expanding each

dimension by a fixed ratio, so as to obtain the best performance and efficiency balance. While traditional methods usually increase one of the dimensions individually, EfficientNetV1 optimizes all three dimensions at the same time to achieve better performance with fewer computational resources.

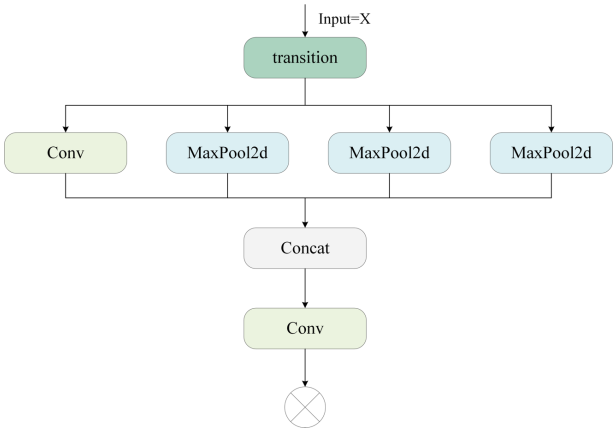


Fig. 4. Structure of SPPELAN

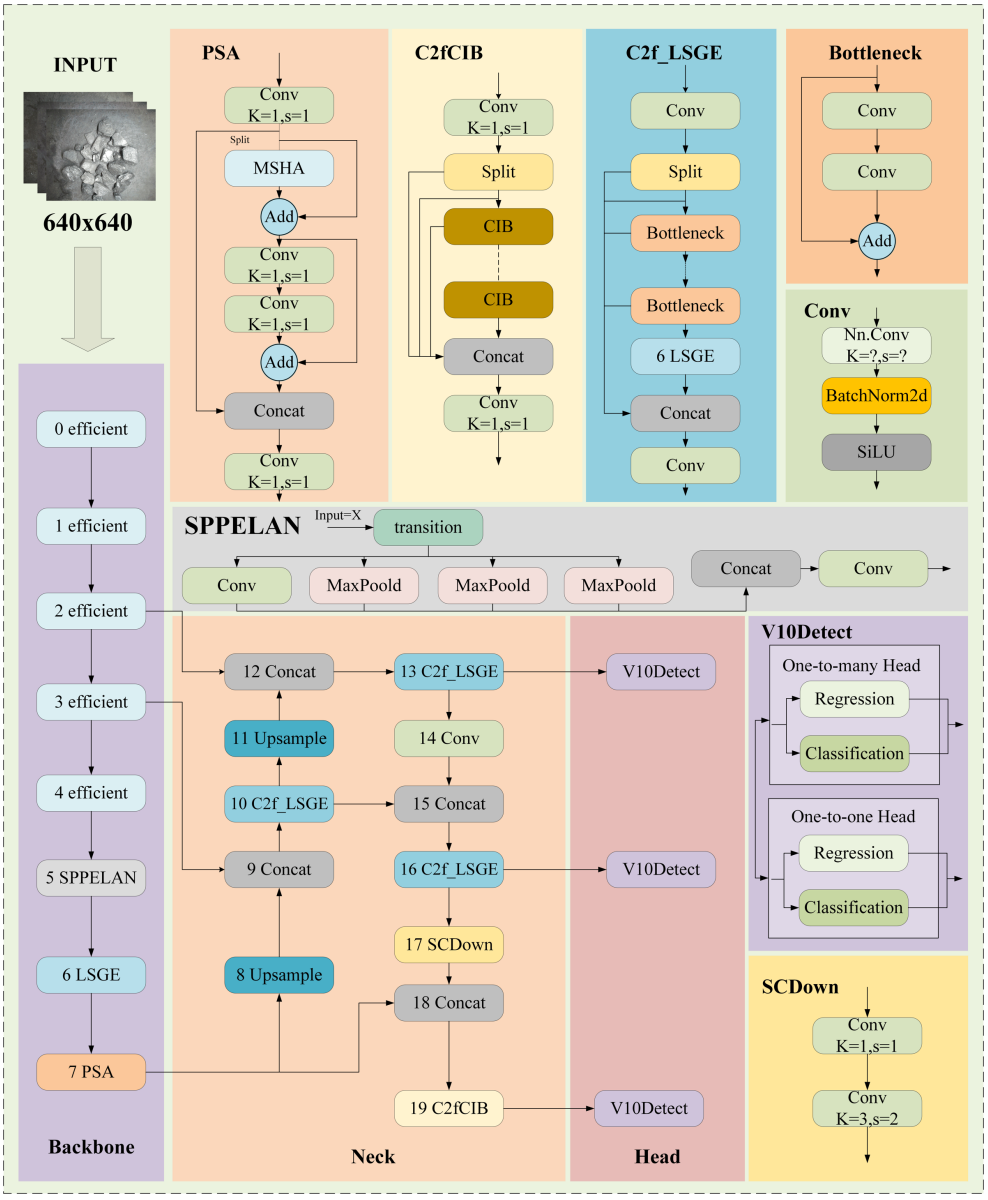


Fig. 5. ELS-YOLO structure diagram

Meanwhile, the LSGE attention mechanism is added behind the EfficientNetV1 module, and then the three C2f modules in the neck are replaced by C2f-LSGE, which is formed by fusing the SGE module and LSKA module. The SGE module splits the feature map into multiple groups, processes them separately, and uses global average pooling to gather statistical information from each group. It then enhances the feature map's local response through normalization and parametric tuning. Meanwhile, the LSKA module breaks down the 2D convolution in deep layers into cascaded horizontal and vertical 1D kernels, reducing computational complexity and memory usage while maintaining performance. The overall design significantly improves the feature extraction while keeping the computational cost low.

Ultimately，The original Spatial Pyramid Pooling Fast (SPPF) module is replaced with the SPPELAN module, which integrates the Spatial Pyramid Pooling (SPP) technique with the Efficient Local Aggregation Network (ELAN) to improve the model's capacity to recognize targets at various scales. The SPP component performs feature extraction at multiple scales, enabling the model to detect objects of different sizes. Following this, ELAN aggregates the local features to enhance the model's ability to focus on critical information within the input image, thereby improving overall performance. This combination allows for more effective and efficient processing of spatial information at different levels.

3. Data acquisition and processing

3.1. Dataset image acquisition

By building the coal gangue image acquisition system in the laboratory as shown in Fig. 6, connecting to the computer through the usb socket to display the acquisition screen in real time and save the pictures. Simulate the underground environment, the collection of photo features such as dense, scattered, obscured, and other situations. The experimental setup primarily consists of vibrating feeder, conveyor, CMOS industrial camera, adjustable light source using the device's image acquisition system to collect gangue images, CMOS industrial array camera model MV-CA050-LUC. The collected pictures are shown in Fig. 7. The coal and gangue size used is 20mm-150mm.

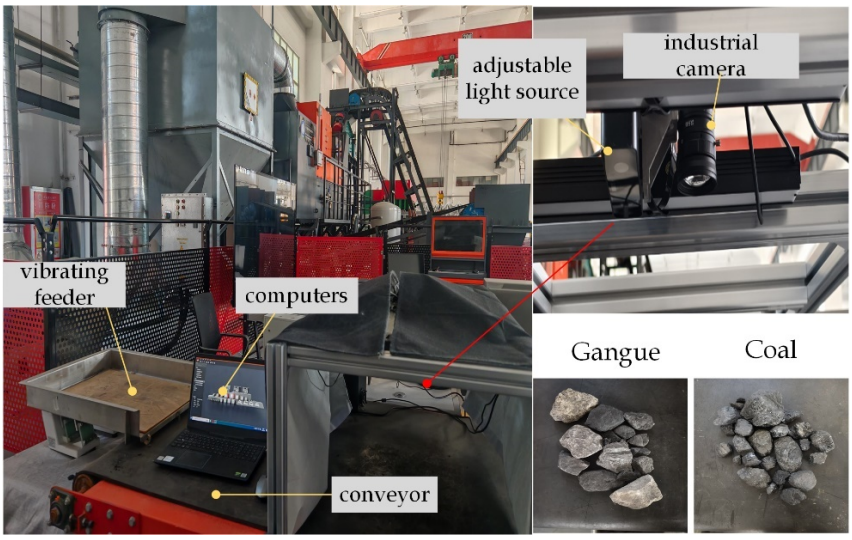


Fig. 6. Gangue recognition device

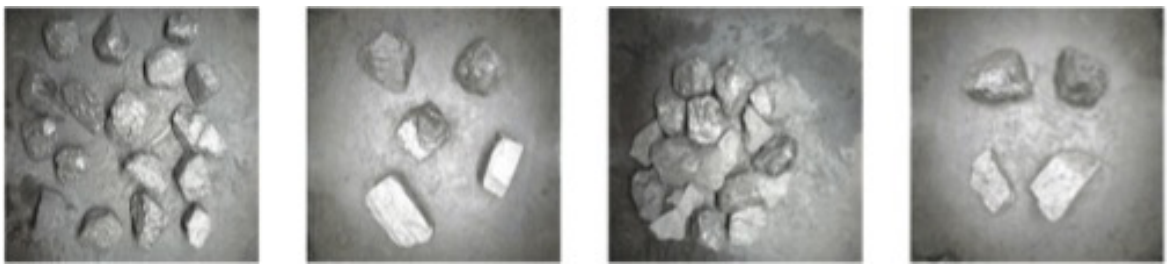


Fig. 7. Example of capturing images

3.2. Image enhancement expansion

A total of 417 images of the dataset images were collected through different placements. Due to the actual scene is more complex, so the use of image enhancement software ImgAug3.4 for data set expansion, in order to fully simulate the actual complex field environment, such as the figure were carried out in the noise fuzzy, motion blur, low illumination, occlusion fuzzy, zoom noise a variety of more extreme image processing, and ultimately the data set of pictures for the 2,486 pictures, and its image enhancement effect as shown in Fig.

(a) original figure (b) noises (c) motion blur (d) low light

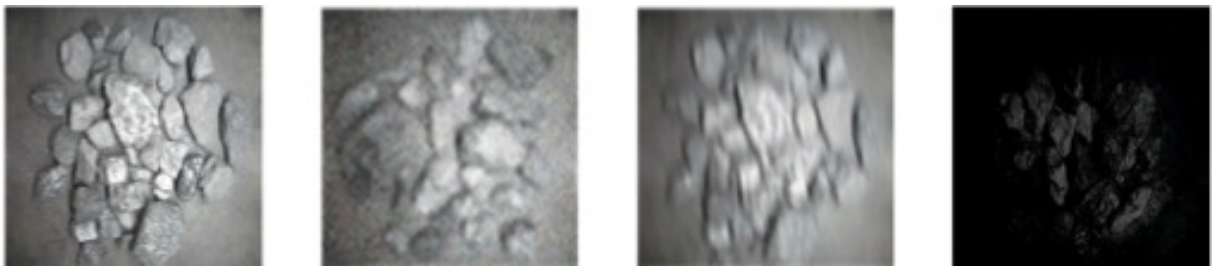


Fig. 8. Image enhancement results

4. Experiments and results

4.1. Model training

The training as well as testing of all the algorithms for the experiments were conducted based on the AutoDL cloud server platform, with the following software environment: PyTorch 2.1.0, Python 3.10 (ubuntu22.04), Cuda 12.1; hardware environment: 12 vCPU Intel(R) Xeon(R) Platinum 8375C CPU @ 2.90GHz, GPU RTX 3090(24GB) * 1. The training uses YOLOv10s model as the baseline model. 2.90GHz, GPU is RTX 3090(24GB) * 1. YOLOv10s model is used as the benchmark model for training. The number of training rounds epochs is set to 301, the batch size is set to 16, the image input size is set to 640×640, the training batch size is set to 32, the number of threads is set to 8, the initial learning rate is set to 0.01, the momentum parameter is set to 0.937, and the weight decay is set to 0.0005.

4.2. Evaluation indicators

In this experiment, in order to evaluate the effectiveness of the ELS-YOLO model for gangue detection, the main evaluation indexes used in this paper are: Precision, Recall, Average Precision, Mean Average Presicion, F1 score and FPS. score) and FPS (Frames Per Second). The evaluation metrics are formulated as in Eqs. 1-6, respectively.

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$P = \frac{TP}{TP+FN} \tag{2}$$

In Eq. 1 and Eq. 2, P denotes Precision and R denotes Recall; TP (True Positive) is the number of positive samples predicted as correct; FP (False Positive) is the number of negative samples predicted as correct; and FN (False Negative) is the number of positive samples predicted as incorrect.

$$AP = \int_0^1 P(r)dr \tag{3}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_{(i)} \tag{4}$$

In Eq. 3 and Eq. 4, AP (Average Precision) is an evaluation metric for the performance of the target detection model, which is calculated based on the integral area under the precision and recall curves. mAP (Mean Average Presicion) is the average of the AP values for each category of labels, which integrates the predictive performance of the model at target detection.

$$F_1 = \frac{2PR}{P+R} \times 100\% \tag{4}$$

$$FPS = \frac{1000}{T} \tag{6}$$

In Eq. 5 and Eq. 6, F1 combines the model's accuracy and recall ability. FPS (Frames Per Second) and T are used to measure the real-time detection speed.

4.3. Comparative experiments of different backbone modules

In order to verify the advantages of EfficientNetV1 module, GhostnetV1, ShuffleNetV2, Swin Transformer, EfficientNetV1 backbone module were added to improve YOLOv10s, and the detection performance comparison test was carried out under the same conditions, and the experimental results and shown in Table 1. It can be seen that the E-YOLO model compared to the other three backbone module modified model, the average accuracy mean value were higher than the Gho-YOLO model 3.9%, Shu-YOLO model 3.6%, Swim-YOLO model 2.7%, compared to the Gho-YOLO model in the number of parameters only increased by 1577632 based on an increase in the average speed of detection increased by 5.66FPS Compared with Shu-YOLO model, the number of parameters and the average detection speed are not enough, but the F1 is increased by 2%, which means that the balance of precision and recall is better, and the average precision mean is higher than 3.6%; Compared with Swin-YOLO model, the number of parameters is lower than 22993038, and the average detection speed is increased by 6.96FPS, and the average detection speed is increased by 2.7%, and the average detection speed is increased by 5.66FPS. The average detection speed is 6.96FPS, and the average detection mean value is 2.7%. In summary, the EfficientNetV1 module has the best overall performance.

Table 1. Comparative experiments with different backbone

Model	AP/%		F1	Parameters	FPS	mAP/%
	Coal	Gangue				
GhostnetV1	85.2	76.3	80.0	9213088	81.30	85.1
ShuffleNetV2	87.8	76.2	82.0	6412760	126.60	85.4
Swin Transformer	86.4	78.5	82.0	33783758	80.00	86.3
EfficientNetV1	87.3	80.8	84.0	10790720	86.96	89.0

In order to be able to show more intuitively the difference between different backbone modules after improvement, visualize feature maps are generated for the same batch of test set images by using visualize script, and the feature maps of the same image in the backbone layer and the last layer of C2FCIB are taken for comparison respectively.

In the feature map, these yellow dots indicate areas with high activation values on the feature map, indicating that the features in the area are highly matched with the target features learned by the model. That is, the more yellow dots and densely appear around the target object, it means that the model has captured the target features, which means that the more important parts or critical regions are detected, the better the performance of its module.

As can be seen from Fig 9, at the backbone layer, EfficientNetV1 has the most and most comprehensive feature points and extracts the most features of the image. When it comes to the last layer of C2FCIB, EfficientNetV1 extracts the most yellow points and the distribution is closer to that of the original image, so it can be seen that the performance of EfficientNetV1 is more superior.

4.4. Comparative experiments on different attention mechanisms

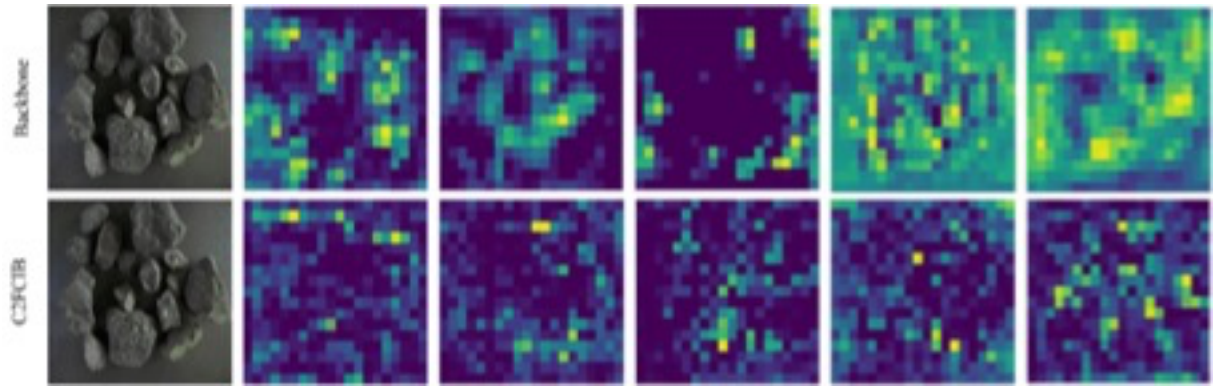
In order to verify the optimization effect of different attention mechanisms on the YOLOv10s network model, CA, EMA, GAM, and LSGE attention modules were used to improve the YOLOv10s model, and the detection performance comparison test was conducted under the same conditions, and the experimental outcomes are presented in Table 2.

From Table 2, it can be seen that the balance of precision and recall of the four models is almost no difference, and in the average precision mean, the model with LSGE module is the highest, 87.2%, which is improved by 0.3%, 0.3%, and 0.4% compared to CA, EMA, and GAM, respectively; the number of parameters is not significantly improved and much lower than the number of parameters of GAM

compared to CA and EMA, and although the FPS of LSGE is lower than that of these three models, it still meets the realization of the detection criteria. In summary, among these four attention mechanisms, LSGE has better comprehensive performance and is more helpful for accurate detection of coal gangue sorting.

Table 2. Comparative experiments of different attention mechanisms

Model	P/%	R/%	F1	Parameters	FPS	mAP/%
CA	88.4	77.5	83.0	8093548	121.95	86.9
EMA	88.7	77.8	83.0	8070524	106.38	86.9
GAM	87.1	78.5	82.0	14624060	119.05	86.8
LSGE	87.4	79.1	83.0	8632534	104.17	87.2



(a)Origin (b)YOLOv10s (b) GhostNet (c) ShuffleNet (d) Swim Transformer (e) EfficientNetV1

Fig. 9. Comparison of different backbone feature maps

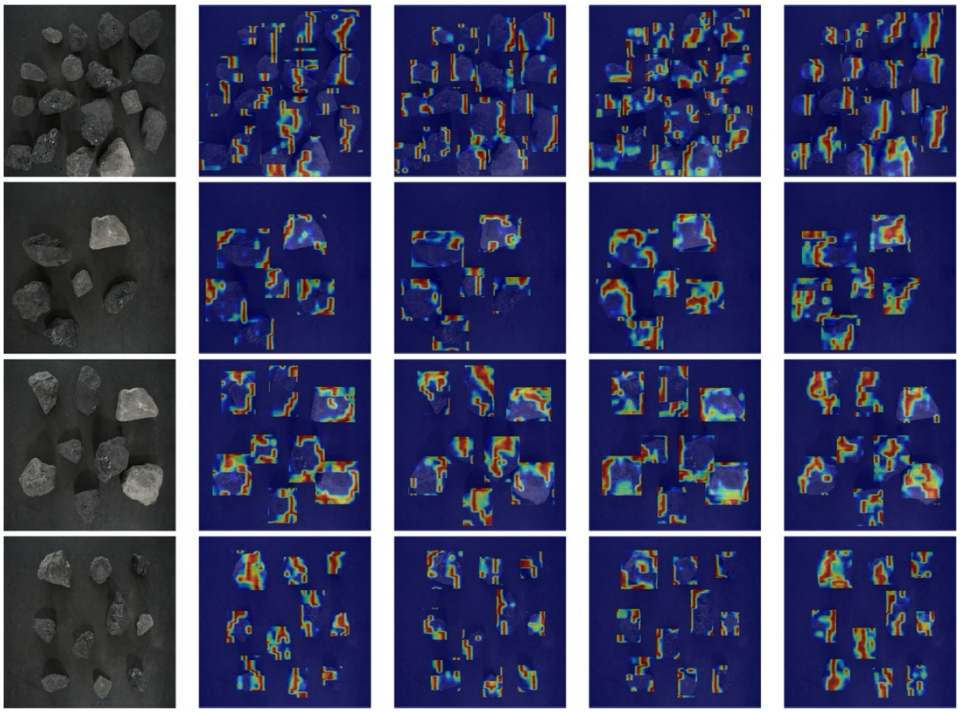
In order to show more clearly the effect of the above modules in feature extraction, the features improved by adding the attention module were extracted and superimposed to generate a network heat map using the Grad_CAM algorithm together with the weight file and yaml file obtained after the training dataset. In the heat map, the red parts indicate the strongest attention regions of the model, which usually correspond to the high confidence regions detected by the model. These regions are the parts of the network that the model considers most likely to contain the target, and thus the red regions are the main regions on which the model bases its decisions. Therefore, the more the red regions are concentrated in the main part of the target, the better the model's detection will be.

As can be seen in Fig. 10, compared with the first three modules, the red and yellow parts of the LSGE module are detected in a wider and more comprehensive range and are concentrated with the center of the gangue, while the GAM and EMA modules are detected in the edge area of the object. In summary, it can be concluded that the LSGE module is able to extract more effective features, which indicates that the LSGE module enhances the network's ability to recognize and focus on the key features of the target.

4.5. Ablation experiments

To assess the effectiveness of the enhancements in the ELS-YOLO model, ablation experiments were conducted and the results are shown in Fig. 11 and Table 3. As shown in the table, the YOLOv10s model is the model without any improvement strategy, and its average accuracy mean value is 86.6%, and the average detection speed is 128.21 FPS. The average accuracy mean value reaches 89.0% after the introduction of the EfficientNetV1 module alone, which is an improvement of 2.4% over the initial model, and the average accuracy mean value of coal and gangue are each increased, respectively, by 2.2% and 2.6%. When the LSGE attention mechanism is introduced alone, the average mean value of the average accuracy rises by 0.8%, and the average detection accuracy decreases by 24.04 FPS, which verifies the effectiveness of the attention mechanism. The accuracy of the SPPELAN module is improved

after it is introduced alone, which indicates that the module can better deal with the targets of different sizes; on the basis of adding the EfficientNetV1 module. After adding the EfficientNetV1 module, the mean average accuracy increased by 1.2% compared to LSGE alone, and the precision increased by 2.7% compared to EfficientNetV1 alone. While after adding the SPPELAN module to the EfficientNetV1 and finally, the average accuracy increased by 1.4% compared to EfficientNetV1 alone. After the introduction of all three modules, the average value of coal and gangue accuracy reaches the highest 90.7% and 88.5% in the experiment, the accuracy rate reaches 90.1, which is 1.2% higher than the initial model, and the F1 reaches the highest 85.0%, at this time, the average speed of detection is 81.30 FPS, which still meets the requirements of real-time detection (Real-time detection is generally achieved at 30FPS), and the average value of average accuracy is 3.0% higher than that of the initial model. The mean value of average accuracy is improved by 3.0% compared with the initial model. The above experimental results show that after adding EfficientNetV1 module, LSGE attention module and SPPELAN module at the same time, the model detection effect is significantly improved.



(a) Origin (b) CA (c) GAM (d) EMA (e) LSGE

Fig. 10. Comparison of heat maps for different attention mechanisms

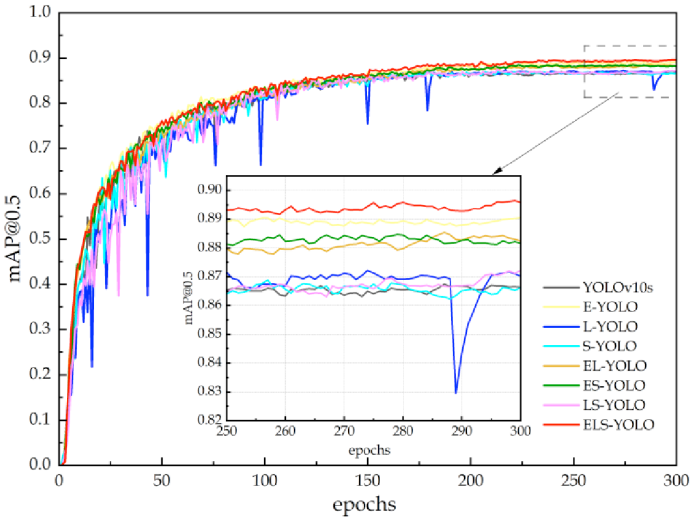


Fig. 11. Line graph of ablation experiments

Table 3. Comparison of results of ablation experiments

Model	Module			AP/%		P/%	R/%	F1	FPS	mAP/%
	EfficientNetV1	LSGE	SPPELAN	Coal	Gangue					
YOLOv10s				88.1	85.1	88.9	77.0	82.0	128.21	86.6
E-YOLO	√			90.3	87.7	86.1	82.3	84.0	86.96	89.0
L-YOLO		√		88.6	85.7	87.4	79.1	83.0	104.17	87.2
S-YOLO			√	88.3	88.1	88.2	78.3	83.0	119.05	86.7
EL-YOLO	√	√		89.3	87.6	88.8	80.2	84.0	77.52	88.4
ES-YOLO	√		√	88.3	88.5	88.4	79.8	84.0	75.19	88.5
LS-YOLO		√	√	88.7	85.7	89.3	78.8	84.0	105.26	87.2
ELS-YOLO	√	√	√	90.7	88.5	90.1	81.3	85.0	81.30	89.6

4.6. Comparative experiments

In order to verify the detection effect of ELS-YOLO model in complex environment, the four cases of normal situation, blurring, darkness, and occlusion are selected to compare and experiment with traditional YOLOv5s, YOLOv7-Tiny, YOLOv8s and YOLOv9s algorithms for the same batch of dataset, and the detection results are shown in Table 4 and as shown in Fig. 12 and Fig. 13.

From Table 4, we can see that the average detection accuracy of ELS-YOLO is the highest, and the average detection speed in the test set reaches 81.30FPS, which indicates that the ELS-YOLO model basically realizes the real-time detection function of coal gangue.

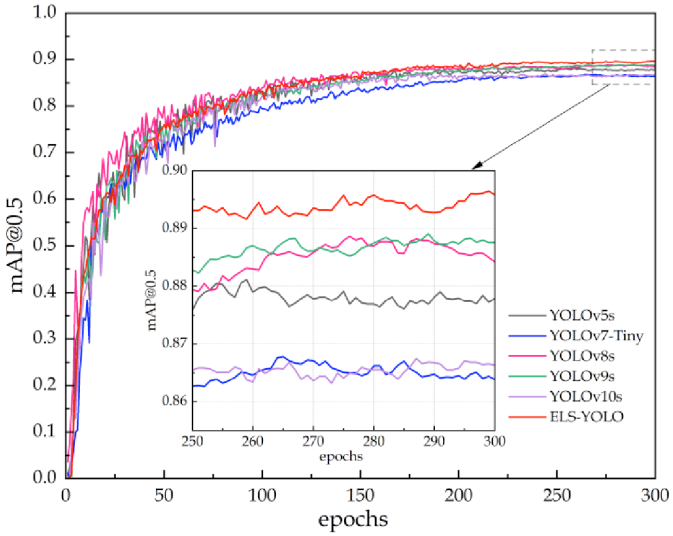


Fig. 12. Comparison experiment line graph

Table 4. Comparison of results of comparative experiment

Model	P/%	R/%	F1	Layers	FPS	mAP/%
YOLOv5s	87.2	79.6	83.0	214	24.8	87.8
YOLOv7-Tiny	85.6	78.6	82.0	263	161.3	86.5
YOLOv8s	88.8	80.1	84.0	225	84.7	88.8
YOLOv9s	87.3	80.8	84.0	1215	35.2	88.8
YOLOv10s	88.9	77.0	82.0	402	128.21	86.6
ELS-YOLO	90.1	81.3	85.0	582	81.30	89.6

From Fig. 13, it can be seen that the green box represents coal, the red box represents gangue, the purple circle represents missed detection, the blue circle represents re-detection, and the orange circle represents false detection. In the YOLOv5s algorithm, (a) shows re-detection, (b) shows missed

detection, and (c) shows both re-detection and false detection. In the YOLOv7-Tiny algorithm, (a) (b) (c) show re-detection, and (c) (d) show false detection. In the YOLOv8s algorithm, (a) shows missed detection, and (c) (d) show false detection. In the YOLOv9s algorithm, (a) (b) (c) show re-detection, and (c) (d) show false detection. In the YOLOv10s algorithm, (b) (c) (d) show missed detection, and (a) (c) show false detection. The ELS-YOLO algorithm correctly detects coal and gangue in complex situations without false or missed detections, demonstrating the superior performance of the model in complex environments.

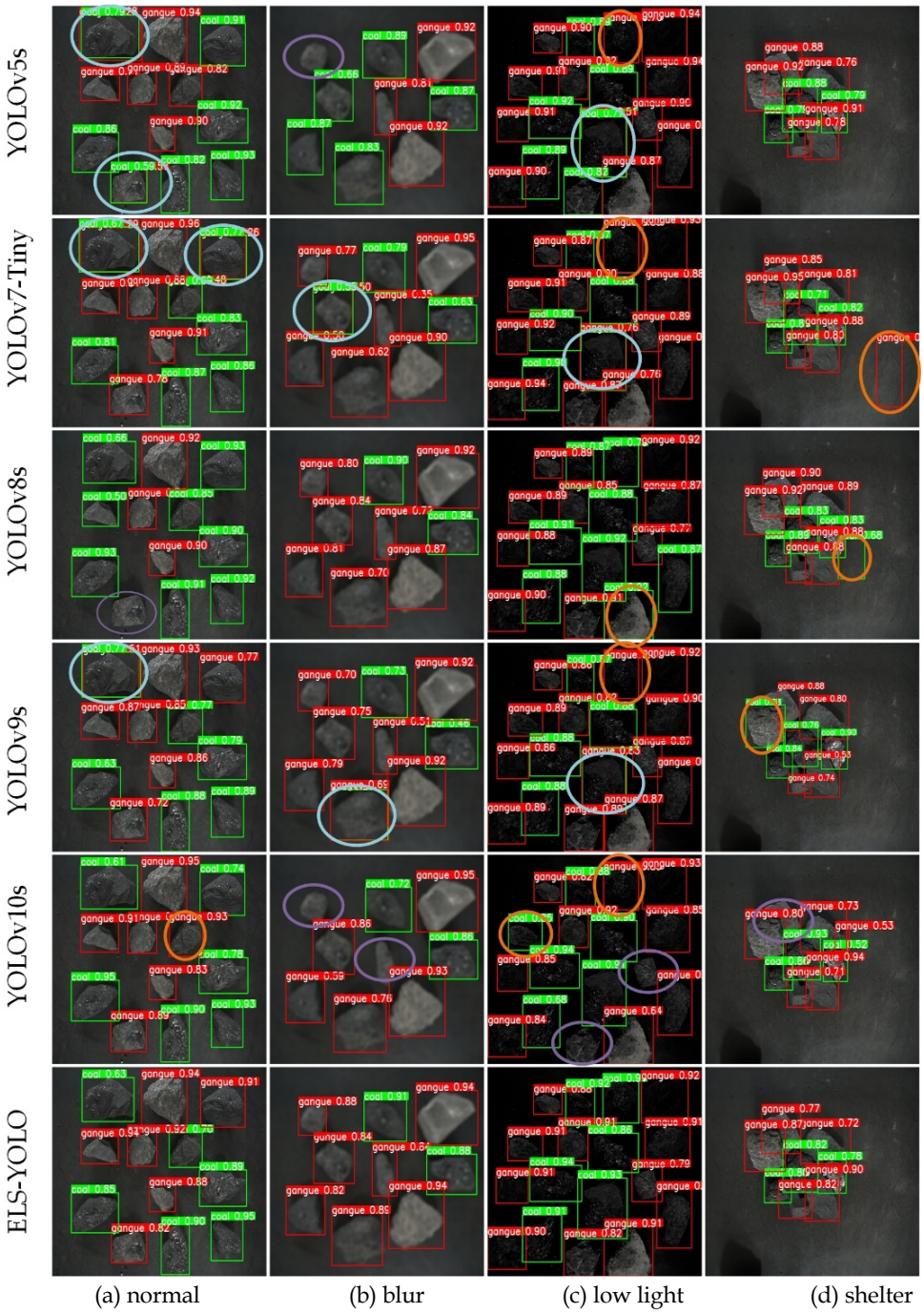


Fig. 13. Comparison of model recognition effect

5. Conclusions

(1) Based on the YOLOv10s network algorithm, the EfficientNetV1 module is introduced in the backbone part, the LSGE attention module is introduced in the neck network part, and finally

SPPELAN is used to replace the original SPPF. the map of the improved ELS-YOLO model is 89.6%, which is 3.0 percentage points higher than that of the original YOLOv10s model. The improved ELS-YOLO model map is 89.6%, which is 3.0% higher than the original YOLOv10s model. The map of the algorithm in this paper is better than 88.7% of the Ghost-SSD model proposed by Du et al. and 86.7% of the CAP-YOLO model proposed by Xu et al. (2022).

(2) The effectiveness of the improved strategy is verified by ablation experiments. The final improved ELS-YOLO network model has a substantial improvement in recognition accuracy relative to the YOLOv10s model. As a final result, its average detection accuracy reaches 89.6% and the detection speed is 81.30FPS, which can realize real-time detection under complex working conditions. It provides faster detection of the model than the 45.50 FPS of the improved ResNet-YOLO of Xue et al. and the 80.3 FPS of the CSR-YOLOv5s proposed by Sun et al. (2023).

(3) The experimental results show that the ELS-YOLO model can accurately and efficiently recognize the coal gangue under complex working conditions, which provides an effective technical means for the accurate recognition of underground coal gangue. At present, the model still has problems such as the difficulty of detecting small targets and the possible influence of the environment in practical application. In the future, the model will be continuously optimized and lightweight, so that it can better cope with the various changes in the underground environment, thus providing a more efficient solution for the automatic identification of coal gangue in the underground environment.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant (No. 52404160, No. 52274152), in part by the Natural Science Outstanding Youth Research Foundation of Anhui Province under Grant (No. 2308085Y37), in part by the Open Fund of Collaborative Innovation Center for Mine Intelligent Technology and Equipment (No.XTZX202404), in part by the Open Fund of Anhui Key Laboratory of Mine Intelligent Equipment and Technology (No.ZKSYS202401), in part by the Open Fund of State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines under Grant (No. SKLMRDPC23KF21)，in part by the Scientific Research Foundation for High-level Talents of Anhui University of Science and Technology under Grant (No. 2023yjrc56).

References

ALFARZAEAI, M. S., HU, E., PENG, W., QIANG, N., AlKAINAEAI, M. M. 2023. *Coal gangue classification based on the feature extraction of the volume visual perception ExM-SVM*. Energies, 16(4), 2064.

CAO, X., LIU, S., WANG, P., XU, G., WU, X. 2022. *Research on coal gangue identification and positioning system based on coal-gangue sorting robot.* Coal Science and Technology, 50(1), 237-246.

CHEN, Z., HUANG, L., JIANG, H., ZHAO, Y., LIU, C., DUAN, C., ZHANG, B., YANG, G., CHAI, J., BAN, H., YU, S. 2022. *Application of screening using a flip-flow screen and shallow groove dense-medium separation in a steam coal preparation plant.* International Journal of Coal Preparation and Utilization, 42(8), 2438-2451.

DAS, P. P., GANGULY, T., CHAUDHURI, R., DEB, S. 2025. YOLO-D: *A Domain Adaptive approach towards low light object detection*. Procedia Computer Science, 258, 3042-3051.

DONG, Z. C., XIA, J. W., DUAN, X. M., CAO, J. C. 2016. *Based on curing age of calcined coal gangue fine aggregate mortar of X-ray diffraction and scanning electron microscopy analysis.* Guang pu xue yu Guang pu fen xi= Guang pu, 36(3), 842-847.

DU, J., SHI, Z., HAO, L., CHEN, R. 2021. *Research on lightweight coal and gangue target detection method.* Ind Mine Autom, 47(11), 119-125.

GRAHAM, B., VAN DER MAATEN, L. 2017. *Submanifold sparse convolutional networks.* arxiv preprint arxiv:1706.01307.

HOWARD, A. G. 2017. *Mobilenets: Efficient convolutional neural networks for mobile vision applications.* arxiv preprint arxiv:1704.04861.

HU, F., HU, Y., CUI, E., GUAN Y., GAO, B., WANG, X., et al. 2022. *Recognition method of coal and gangue combined with structural similarity index measure and principal component analysis network under multispectral imaging*. Microchemical Journal.

KONG, L., LI, H., XU, S., XU, Q. 1997. *The On-Line Identification and Separation System for Coal Gangues Based on Double Energy gamma-Ray Transmission.* JOURNAL-HUAZHONG UNIVERSITY OF SCIENCE AND

TECHNOLOGY CHINESE EDITION, 25, 107-108.

LAI, W., HU, F., KONG, X., YAN, P., BIAN, K., DAI, X. 2022. *The study of coal gangue segmentation for location and shape predicts based on multispectral and improved Mask R-CNN.* Powder Technology, 407, 117655.

LAU, K. W., PO, L. M., REHMAN, Y. A. U. 2024. *Large separable kernel attention: Rethinking the large kernel attention design in cnn*. Expert Systems with Applications, 236, 121352.

LEI, S., XIAO, X., ZHANG, M. 2021. *Research on coal and gangue identification method based on improved YOLOv3.* Mining Safety & Environmental Protection, 48(3), 50-55.

LI, D. Y., WANG, G. F., ZHANG, Y., WANG, S. 2022. *Coal gangue detection and recognition algorithm based on deformable convolution YOLOv3.* IET Image Processing, 16(1), 134-144.

LI, J., DU, C, L., BAO, W. 2010. *Direct-impact of sieving coal and gangue.* Mining Science and Technology (China), 20(4), 611-614.

LI, X., HU, X., YANG, J. 2019. *Spatial group-wise enhance: Improving semantic feature learning in convolutional networks.* arxiv preprint arxiv:1905.09646.

LIU, Z., MAO, H., WU, C. Y., FEICHTENHOFER, C., DARRELL, T., XIE, S. 2022. *A convnet for the 2020s.* In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11976-11986).

NGUYEN, H. H., HOANG, M. S. 2025. LEAF-YOLO: *Lightweight Edge-Real-Time Small Object Detection on Aerial Imagery*. Intelligent Systems with Applications, 25, 200484.

RUI, L. I., BO, L. I., XUE-WEN, W., TAO, L., LIAN-JIE, L., SHU-XIANG, F. (2022). *A classification method of coal and gangue based on XGBoost and visible-near infrared spectroscopy.* Spectroscopy and Spectral Analysis, 42(9), 2947-2955.

SUN, Z. P., TAO, H. J., LI, X. W. 2023. *Coal and gangue target detection algorithm based on CSR - YOLOv5s*. Coal. 32(7): 31−34, 69.

TAN, M., LE, Q. (2019). *Efficientnet: Rethinking model scaling for convolutional neural networks.* In International conference on machine learning (pp. 6105-6114). PMLR.

WANG, A., CHEN, H., LIU, L., CHEN, K., LIN, Z., HAN, J., DING, G. 2024. *Yolov10: Real-time end-to-end object detection.* arxiv preprint arxiv:2405.14458.

WANG, C. Y., YEH, I. H., MARK LIAO, H. Y. 2024. *Yolov9: Learning what you want to learn using programmable gradient information.* In European conference on computer vision (pp. 1-21). Cham: Springer Nature Switzerland.

WANG, G., LIU, F., MENG, X., FAN, J., WU, Q., REN, H., PANG, Y., XU, Y., ZHAO, G., ZHANG, D., CAO, X., DU, Y., ZHANG, J., CHEN, H., MA, Y., ZHANG, K. 2019. *Research and practice on intelligent coal mine construction (primary stage).* Coal Science and Technology, (8).

WANG, G. J., SU, T. T., LIU, W. B., QIAN, Z. P., LI, J. 2020. *Design of intelligent coal and gangue sorting system based on EAIDK.* Ind. Mine Autom, 46(1), 105-108.

WANG, G., PANG, Y., Ren, H., ZHAN, K., DU, M., ZHANG, Y., CHENG, J., DU, Y., ZHANG, J., GONG, S., WANG, D., MENG, L., MENG, J. 2024. *System engineering and key technologies research and practice of smart mine.* Journal of China Coal Society, *49*(1), 181-202.

WANG, J., PAN, W., ZHANG, G., YANG, S., YANG, K., LI, L. 2022. *Principles and applications of image based recognition of withdrawn coal and intelligent control of draw opening in longwall top coal caving face.* Journal of China Coal Society, 47(1), 87-101.

XING, H. Y., YI, M., YANG, J. P., ZHU, K. Y., LIU, C. 2022. *Metal magnetic memory quantitative inversion model based on IPSO-GRU algorithm for detecting submarine pipeline defect.* Chinese Journal of Engineering, 44(5), 911-919.

XUE, B., ZHANG, Y., LI, J., WANG, Y. (2023). *A review of coal gangue identification research – application to China's top coal release process.* Environmental Science and Pollution Research, 30(6), 14091-14103.

XUE, G., LI, S., HOU, P., GAO, S., TAN, R. 2023. *Research on lightweight Yolo coal gangue detection algorithm based on resnet18 backbone feature network.* Internet of Things, 22, 100762.

XU, Z., LI, J., MENG, Y., ZHANG, X. 2022. *CAP-YOLO: Channel attention based pruning YOLO for coal mine real-time intelligent monitoring*. Sensors, 22(12), 4331.

YANG, D., MIAO, C., LI, X., LIU, Y., WANG, Y., ZHENG, Y. 2023. *Improved YOLOv7 Network Model for Gangue Selection Robot for Gangue and Foreign Matter Detection in Coal.* Sensors, 23(11), 5140.

YUAN, L., ZHANG, N., KAN, J. G., WANG, Y. 2018. *The concept, model and reserve forecast of green coal resources in China.* J. China Univ. Min. Technol, 47, 1-8.

ZHANG, K., YANG, X., XU, L., THÉ, J., TAN, Z., YU, H. 2024. *Enhancing coal-gangue object detection using GAN-based data augmentation strategy with dual attention mechanism.* Energy, 287, 129654.