Research on Gangue Detection Method Based on GD-YOLO

Baoren Wang, Hu Cui, Xiaoqing Yu, Ziao Su, Yifei Zheng

***Abstract*—The screening of coal gangue is a crucial aspect of the coal mine production process. This paper proposes a new algorithm model, GD-YOLO, which improves upon the original GD feature fusion mechanism by discarding the feature fusion method based on FPN+PAN architecture. The objective is to enhance the detection of coal gangue. The bottleneck is enhanced into a PConv convolution-based FasterNet Block, which enhances the model's capacity to extract spatial features by reducing superfluous computation and memory access. The EMA attention mechanism is integrated into the enhanced C2f-Faster to achieve further enhancement in detection performance while utilizing lower GFLOPs. In order to address the issue of suboptimal model convergence speed and effectiveness, the model is guided to predict coal and gangue by improving the loss function to SIoU, which is calculated using three key aspects: angle, distance, and shape. The results of the validation experiments demonstrate the efficacy of the algorithm, with a mean accuracy (mAP) of 96.6%, which is 4.9% higher than that of the traditional YOLOv8n algorithm. Additionally, the fast detection speed reaches 61.5 FPS, indicating excellent potential for industrial applications and the capacity to meet the industrial requirements of coal gangue detection.**

***Index Terms*—Gangue detection, Feature fusion, Attention mechanism, Deep learning, YOLO**

I. Introduction

Coal gangue is the solid waste produced in the process of coal mining and treatment. However, the emission and treatment of coal gangue not only results in a considerable amount of resource waste, but also causes environmental pollution, which is contrary to the principles of developing green mines[1]. Therefore, the accurate and efficient detection and identification of coal gangue has become a pressing issue in contemporary mine management.

At this juncture, the prevailing gangue sorting methodologies are predominantly manual, X-ray gangue selection[2], and heavy media gangue selection[3]. The conventional manual sorting approach is inherently inefficient and yields a considerable expenditure of human resources, while exhibiting a lack of precision. As a physical sorting method, heavy media gangue sorting can achieve density separation when dealing with coal gangue. However, its high equipment cost, energy consumption, high quality requirements for gangue, wastewater treatment problems, and other shortcomings limit its wide application. X-ray gangue sorting can achieve high-precision sorting, but the cost of equipment and operating costs are high, and the operator's requirements and the environment have a greater impact. Consequently, the present situation requires that gangue sorting be approached on the basis of a comprehensive assessment of the technical, economic, and environmental factors, with the objective of selecting the most appropriate sorting method.

The rapid development of deep learning techniques in recent years has provided new possibilities for solving this problem. Deep learning algorithms have become the tool of choice for researchers to study image recognition problems due to their excellent performance on image recognition and classification tasks, especially CNN[4-7] and vision transformer[8-11].

The detection of coal gangue based on deep learning is confronted with three principal challenges. Initially, the shapes of coal and gangue are not fixed, and the visual differences are primarily manifested in shape, color, and texture. Consequently, the model must possess an exceptional capacity for feature information extraction and fusion. Secondly, the color between coal and gangue Furthermore, the actual background is similar, which will result in a reduction in detection accuracy. Thirdly, the model must be capable of responding rapidly to coal and gangue in the actual production process. Therefore, the model's ability to perform rapid inference is essential for effective problem-solving.

In order to address the issues of a limited sensing area and the inaccuracy of identifying small targets in gangue recognition, Li[12] put forth a deformable convolutional YOLOv3-based gangue detection and recognition algorithm. Sun[13] put forth an intelligent classification method of gangue based on multispectral imaging technology and YOLOv5 target detection. In order to address the issue of the difficulty in deployment of existing algorithms on mobile devices, Guo[14] has developed a lightweight feature extraction network by combining CSPDarknet53 and GhostNet with embedded efficient channel domain consideration. Furthermore, the use of the Meta-ACON activation function allows for the adjustment of the nonlinearity of each layer of the network. Wang[15] proposes

Manuscript received May 16, 2024; revised November 19, 2024.

Baoren Wang is an associate professor of College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, SD China (Corresponding author, e-mail: kdwbr@126.com).

Hu Cui is a Postgraduate of College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, SD, China. (e-mail: 1343522438@qq.com).

Xiaoqing Yu is a Postgraduate of College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, SD, China. (e-mail: 1468915885@qq.com).

Ziao Su is a Postgraduate of College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, SD, China. (e-mail: 1571353752@qq.com).

Yifei Zheng is a Postgraduate of College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, SD, China. (e-mail: 752070671@qq.com).

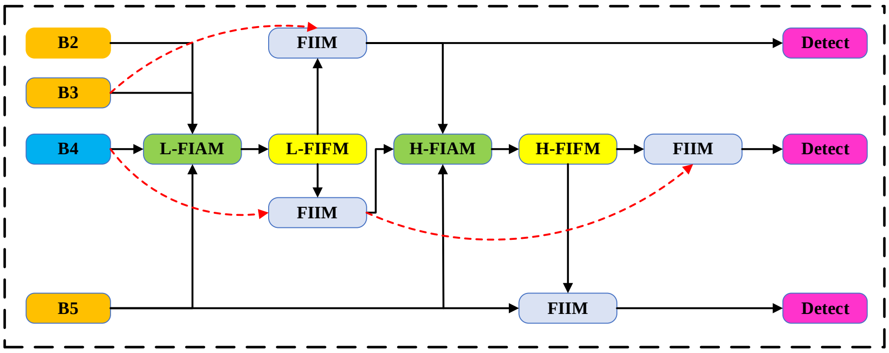

Fig. 2. Structure of the GD Network.

an inverse residual structure-based coal gangue target detection network, designs the DPsP and DsP structures, and combines them with GhostModule to construct a GDPs-YOLO network in YOLOv8s. Zhou proposed the GH-YOLOv8s model, which enhances the feature fusion capability and improves the detection ability of the model by, among other things, constructing a HAM structure in the backbone.

The location of the gangue screening site is illustrated in Fig. 1. In order to meet the industrial requirements of coal gangue detection, this paper proposes a GD-YOLO algorithm model based on YOLOv8n. The original algorithm's feature fusion method based on FPN+PAN architecture has been discarded in favor of a new GD feature fusion mechanism. The bottleneck is enhanced with the incorporation of the FasterNet Block, which is based on PConv convolution. This improvement facilitates an enhancement in the model's capacity to extract spatial features, achieved through a reduction in redundant computation and memory access. Subsequently, the EMA attention mechanism was integrated into the enhanced C2f-Faster to further enhance detection performance while reducing the GFLOPs. By refining the loss function to SIoU, the model was guided to predict coal and gangue, addressing the issue of slow model convergence and effectiveness.

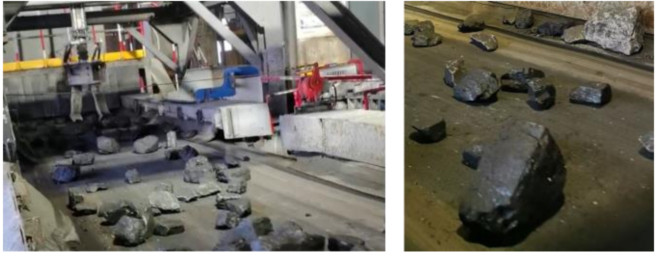

Fig. 1. On-site working environment.

II. Related Works

The YOLOv8 algorithm model has been demonstrated to be highly effective in real-time gangue detection tasks. It has also been shown to be adaptable and accurate in a variety of different scenarios, and it can be used in industrial production environments. Furthermore, it has undergone multiple versions of improvement, indicating strong stability[16-18]. Despite the YOLOv8 model's notable performance, it is not without limitations in complex scenarios. This paper proposes a novel GD-YOLO algorithm model, which, upon subsequent experimental evaluation, has demonstrated superior performance in gangue detection across multiple key indicators compared to the YOLOv8 algorithm.

This chapter presents the GD-YOLO model, which is comprised of three primary components: the enhanced GD feature fusion network, the enhanced C2f-Faster-EMA module, and the refined SIOU loss function.

A. *Improved GD feature fusion network*

The feature fusion component of the YOLO series has been significantly advanced from the conventional FPN-based architectural framework, which addresses the multi-scale feature fusion challenge through a network of parallel branches[19,20]. However, the feature fusion structure based on the traditional FPN architecture is only capable of completely fusing the feature information of neighboring layers. It is unable to obtain the feature information of other layers directly, but can only do so indirectly through recursion. As a result, this traditional FPN-based feature fusion structure causes a significant amount of feature information loss during the model's calculation process. The unutilized feature information is then discarded by the feature fusion network. This will have an impact on the overall effect of the feature information. In YOLOv8[21,22], the GD network mechanism is employed in place of the feature fusion network of FPN+PAN, and the GD network structure is illustrated in Fig. 2.

The GD feature fusion network comprises three principal modules: the FIAM (Feature Information Alignment Module), the FIFM (Feature Information Alignment Module), and the FIIM (Feature Information Injection Module).

The FIAM feature alignment module is comprised of two distinct modules: L-FIAM (Low-stage Feature Information Alignment Module) and High-FIAM (High-stage Feature Information Alignment Module). This is illustrated in Fig. 3.

The B2, B3, and B4 features are subjected to a reduction and sampling process utilising the Average Pooling function in L-FIAM. This is employed to achieve a uniform size, whereby the input features are adjusted to the smallest feature size within the group($R_{B4} = \frac{1}{4}R$), we obtain F_{align}, the formula is as follows:

$$F_{\text{align}} = L\_FIAM([B2, B3, B4, B5]) \tag{1}$$

$$y_{kij} = \frac{1}{|\mathcal{R}_{ij}|} \sum_{(p,q)\in\mathcal{R}_{ij}} x_{kpq} \tag{2}$$

where y_{kij} represents the average pooled output value of the k-feature graph in $\mathcal{R}_{ij}$, x_{kpq} denotes the element at

(p,q) in $\mathcal{R}_{ij}$, and $|\mathcal{R}_{ij}|$ denotes the number of elements in $\mathcal{R}_{ij}$.

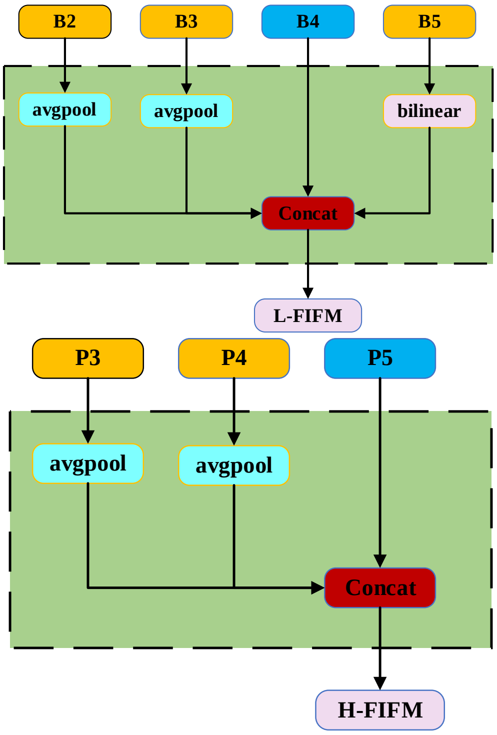

Fig. 3. Structure of the FIAM.

Concurrently, bilinear interpolation up-samples the B5 feature information output from the backbone, which is based on the linear interpolation extension of the interpolation function of the two variables.

While a larger feature size may retain more of the original underlying information, it will also increase the computational burden as the feature size increases. Therefore, in order to maintain a balance between the computational speed and accuracy of the model, it is necessary to determine the optimal feature size. This model adopts R_{B4} as the target size for feature information alignment.

The High-FAM module reduces the dimensionality of the input feature information to a uniform size through an average pooling operation, thereby obtaining F_{align}. As the module extracts higher-order information, the average pooling operation promotes the aggregation of the information and improves the speed of the subsequent model computation. The formula is as follows:

$$F_{\text{align}} = High\_FAM([P2, P3, P4]) \tag{3}$$

The FIFM comprises two distinct modules: L-FIFM (Low-stage Feature Information Fusion Module) and H-FIFM (High-stage Feature Information Fusion Module). These are illustrated in Fig. 4.

The L-FIFM incorporates both RepBlock and splitting operations. The initial convolution transforms the number of feature information channels into an intermediate channel. F_{align} is employed as the input to RepBlock, and F_{fuse} is calculated. The fusion of feature information is conducted by RepConv-blocks, and subsequently, the number of feature information channels is transformed to a $C_{B3} + C_{B4}$ size through a convolution operation. The feature information computed by RepBlock is partitioned into two parts, F_{inj_P3} and F_{inj_P4}, which are then fused with the feature information of different layers. The formula for this process is as follows:

$$F_{\text{fuse}} = RepBlock\left(F_{\text{align}}\right) \tag{4}$$

$$F_{\text{inj\_}P3}, F_{\text{inj\_}P4} = Split\left(F_{\text{fuse}}\right) \tag{5}$$

H-FIFM is comprised of two primary components: the Swin Transformer Block and the Split. The former contains W-MSA (Windows Multi-head Self-Attention), SW-MSA (Shifted Windows Multi-Head Self-Attention), LE-FFN (Locally Enhanced Feed-Forward Network), and residual connectivity components. The latter is a residual connection that allows for the integration of external data.

The multi-head self-attention (MHA) mechanism captures global attention, necessitating the computation of all patches. Consequently, high-resolution image processing requires a considerable computational burden. To address this, the adoption of W-MSA with SW-MSA in lieu of global attention can significantly reduce the model's complexity. The underlying mathematical formulas for MSA and W-MSA are as follows:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2 C \tag{6}$$

$$\Omega(W - MSA) = 4hwC^2 + 2(hw)^2 C \tag{7}$$

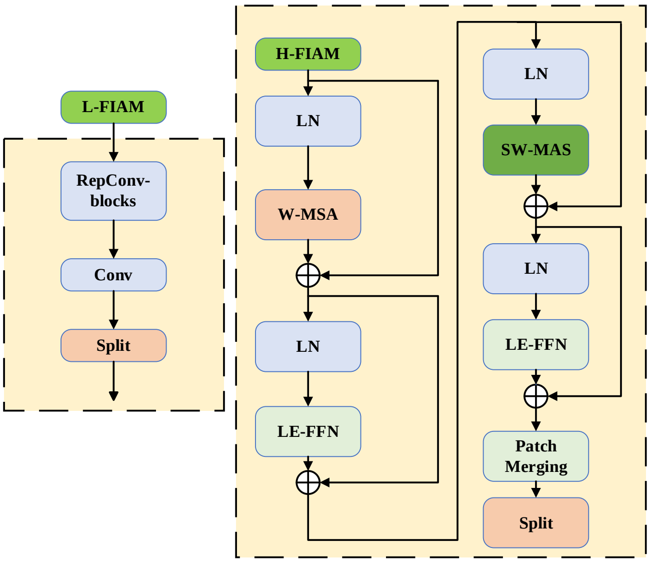

Fig. 4. Structure of the FIFM.

W-MSA is capable of extracting the feature relationship within the local window and carrying out the self-attention calculation within the window. However, W-MSA is limited in its ability to intersect neighboring windows. To address this, the SW-MSA mechanism has been developed, which employs a slice-and-block interaction approach. This enables the transfer of information between windows, facilitating interaction and enhancing the capture of local and global feature information in the image. It addresses the limitations of W-MSA and improves the detection efficacy of the model in challenging coal and gangue working conditions.

The traditional FFN does not consider the spatial relationship between significant tokens, which necessitates a substantial amount of training data to learn the generalization bias. Consequently, the traditional FFN module is substituted with LE-FFN, whose structure is illustrated in Fig. 5. This alternative is designed to enhance the correlation between diverse tokens in the spatial domain. The LE-FFN module process is as follows: first, the tokens $x_c^{\hbar} \in \mathbb{R}^{(N+1)\times C}$ computed from the MSA module are split into patch tokens and class tokens. The patch tokens are then projected to higher dimensions $x_p^{l_1} \in \mathbb{R}^{N\times(e\times C)}$ by linear projection, where e is the expansion ratio. Subsequently, the original image is reduced to the spatial dimension based on its relative position. Then, a deep convolution operation is performed on the reduced patch tokens, where the size of the convolution kernel is k, with the objective of enhancing the correlation with the neighboring $k^2 - 1$ tokens, thus obtaining $x_p^d \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times(e\times C)}$. Finally, the patch tokens are linearly projected to the initial dimension and then connected to the class token.

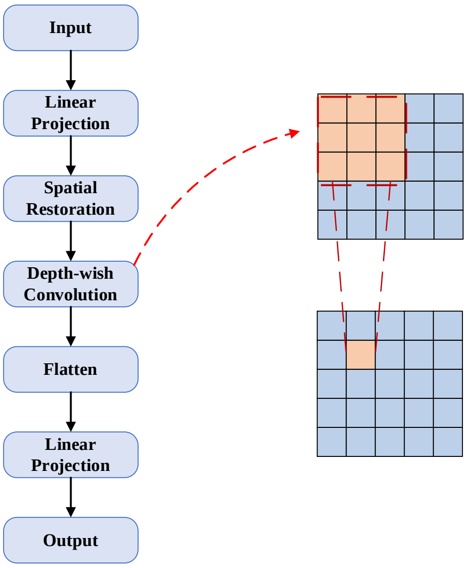

Fig. 5. Structure of the LE-FFN.

In order to enhance the efficacy of the model in integrating global information, the FIIM module leverages the expertise in segmentation and integrates the data through the application of an attention operation, as illustrated in Fig. 6. The model simultaneously inputs the local information F_{local} derived from the current hierarchical features and the global injected information F_{inj} derived from the IFM computation. Since F_{local} and F_{global} are of different sizes, they are adjusted by average pooling and bilinear interpolation. Subsequently, the RepBlock module is added to further extract and fuse the information following the attention fusion.

In particular, a lightweight neighbor layer fusion (LAF) module is incorporated at the input position of the FIIM module to facilitate cross-layer information flow, as illustrated in Fig. 7. The LAF model exhibits distinct characteristics in the shallow and deep feature processing modules. The LAF is configured using average pooling and bilinear interpolation, which enables the optimization of information flow paths between different layers and enhances the overall performance of the module without significantly impacting the computational speed.

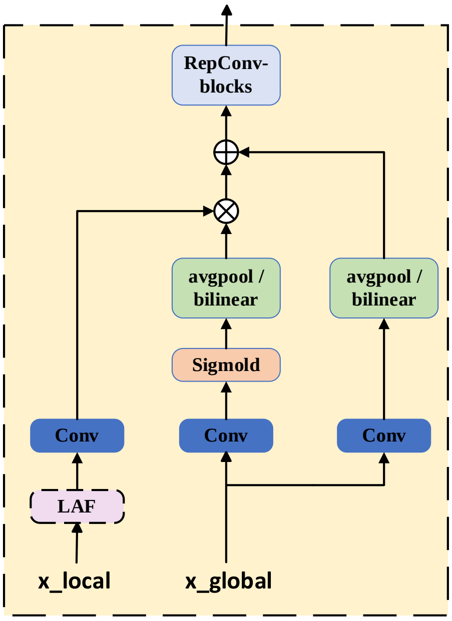

Fig. 6. Structure of the FIIM.

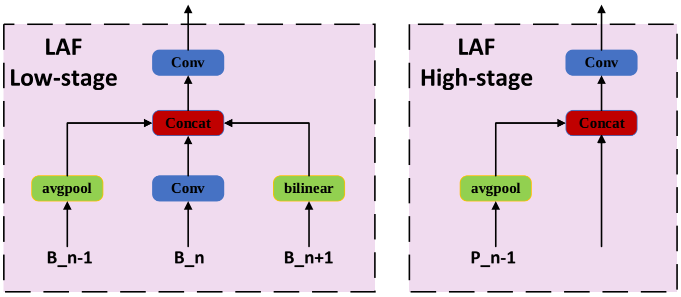

Fig. 7. Structure of the LAF.

B. Improved C2f-Faster-EMA module

The FasterNet Block is designed based on PConv, which can further reduce the computational burden of the model while ensuring its accuracy. The FasterNet Block consists of one PConv convolutional layer and two 1×1 Conv ordinary convolutional layers. The first 1×1 convolutional layer in the FasterNet Block is followed by the batch normalization (BN) and ReLU activation layer. However, excessive use of these two elements may result in suboptimal model performance. Therefore, they are placed in the first 1×1 convolutional layer. Subsequently, the batch normalization (BN) and ReLU activation layer are employed. However, excessive utilization of these two elements may result in a decline in model performance. Therefore, they are situated after the initial 1×1 Conv ordinary convolutional layer. The configuration of the FasterNet Block and PConv is illustrated in Fig. 8.

For the memory sequential access case, PConv keeps most of the channels unchanged, and only the input channel C_p is used for spatial feature extraction, and when the number of input and output feature map channels are the same, the FLOP is $\hbar \times w \times k^2 \times c_p^2$, and the FLOP of ordinary convolution is sixteen times higher than that of PConv if $r = \frac{c_p}{c} = \frac{1}{4}$, and the memory access of PConv is reduced drastically in comparison with ordinary convolution, and it is about 1/4 that of the original convolution, in which the memory access of PConv is $h \times w \times 2c_p + k^2 \times c_p^2$.

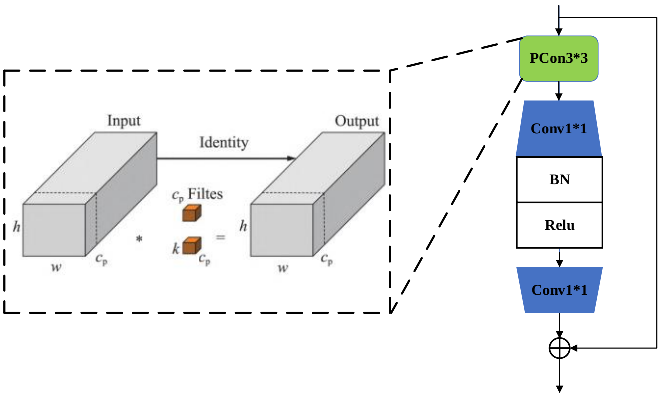

Fig. 8. Structure of the FasterNet Block and PConv.

The EMA is an efficient multi-scale attention mechanism that enhances the correlation between different information, improves the integrity of channel information retention, increases the uniformity of spatial feature distribution within each feature group, and realizes richer feature fusion effects by employing feature grouping, parallel sub-networks, and cross-spatial learning. The structure of the EMA module is illustrated in Fig. 9.

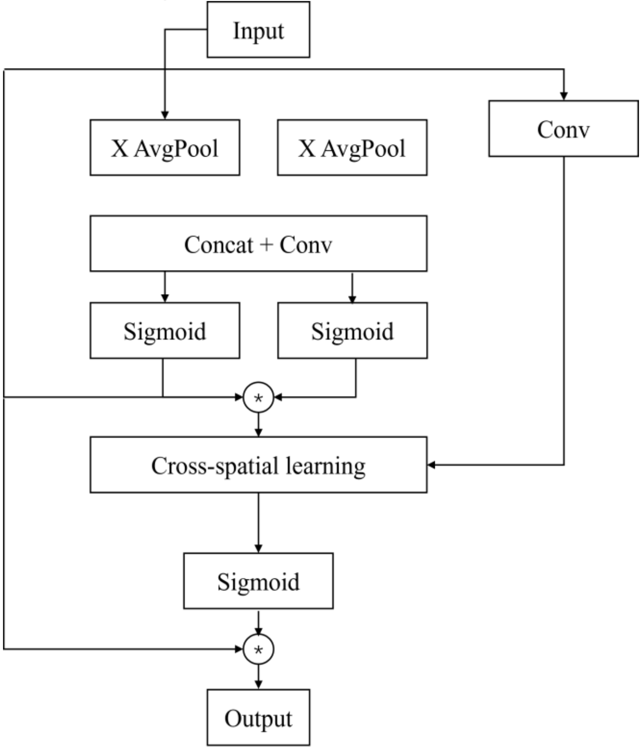

Fig. 9. Structure of the EMA.

EMA divides the input feature map X ∈ R^(C×H×W) into G sub-features across channel dimension directions, where sub-features $X = [X_0, X_0, \dots, X_{G-1}], X_i \in \mathbb{R}^{C//G\times H\times W}$. Three parallel paths are employed to extract the attention weight descriptors, which are subdivided into two 1×1 branches and one 3×3 branch. This parallel approach enables the effective utilization of resources and the optimization of the module's computational logic. However, there is a dearth of feature aggregation capability. To address this, EMA employs a cross-space information aggregation approach across different spatial dimensions. This is achieved by introducing two tensors in the output portion of the 1×1 branch and the 3×3 branch, and encoding the global spatial information using two-dimensional global average pooling. This is illustrated in the following equations:

$$Z_c = \frac{1}{H\times W}\sum_{j}^{H}\sum_{i}^{W} x_c(i,j) \tag{8}$$

The three attention mechanisms of SENet, CBAM, and EMA are sequentially embedded in the C2f-Faster model for model improvement. Their comprehensive performance is then compared through comparative experiments, and it is ultimately found that the selection of the EMA attention mechanism is integrated into the C2f-Faster model with the best effect. Accordingly, the EMA attention mechanism has been incorporated into the FasterNet Block module, resulting in the C2f-Faster-EMA configuration, as illustrated in Fig. 10.

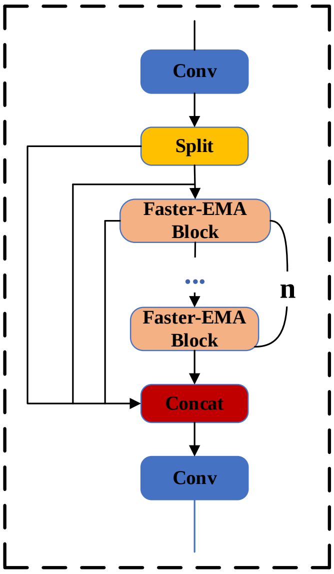

Fig. 10. Structure of the C2f-Faster-EMA.

The structure of the enhanced GD-YOLO model is illustrated in Fig. 11. Initially, the feature fusion methodology based on the FPN+PAN architectural configuration in the original algorithm is refined into a novel GD feature fusion mechanism by eliminating it. Subsequently, the Bottleneck is enhanced into a FasterNet Block-based Subsequently, the EMA attention mechanism is integrated into the enhanced C2f-Faster, followed by the incorporation of the SIoU loss function, which considers the three dimensions of angles, distances, and shapes, enabling the model to make predictions regarding coal and gangue.

C. *Improved SIOU loss function*

The SIoU Loss function is constructed based on three fundamental relationships between two frames: angle, distance, and shape. The angle loss is dynamically adjusted based on the angle difference between the centroid coordinates of the two frames, and this adjustment is prepared for by the distance loss, which is also dynamically adjusted based on the Euclidean distance between the centroid coordinates of the two frames. Finally, the shape loss is penalized based on the shape difference between the two frames, as illustrated in Fig. 12.

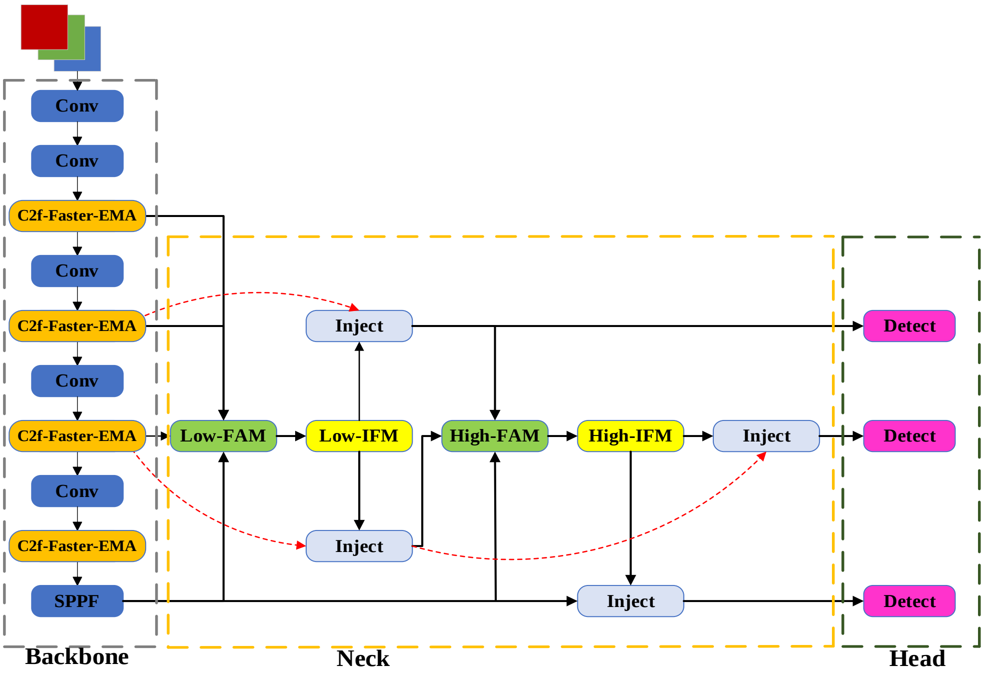

Fig. 11. Structure of the GD-YOLO.

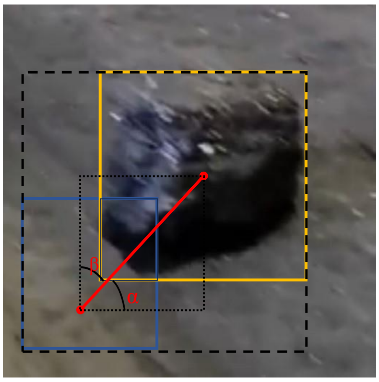

Fig. 12. SIoU Parameter Schematic.

The angular loss of SIoU first determines the size of α. If $\alpha \leq 90°$, the prediction frame is moved along the y-axis to minimize α, and vice versa, it is moved along the x-axis to minimize β, the formula is as follows:

$$\Lambda = 1-2*\sin^2\left(\arcsin(x)-\frac{\pi}{4}\right) \tag{9}$$

$$x=\frac{c_h}{\sigma}=\sin(\alpha) \tag{10}$$

$$\sigma=\sqrt{\left(b^{gt}_{c_x}-b_{c_x}\right)^2+\left(b^{gt}_{c_y}-b_{c_y}\right)^2} \tag{11}$$

$$c_h=\max\left(b^{gt}_{c_y},b_{c_y}\right)-\min\left(b^{gt}_{c_y},b_{c_y}\right) \tag{12}$$

Distance loss is calculated based on angular, the formula is as follows:

$$\Delta=\sum_{t=x,y}\left(1-e^{-\gamma\rho_t}\right) \tag{13}$$

$$\rho_x=\left(\frac{b^{gt}_{c_x}-b_{c_x}}{c_w}\right)^2,\rho_y=\left(\frac{b^{gt}_{c_y}-b_{c_y}}{c_h}\right)^2,\gamma=2-\Lambda \tag{14}$$

The formula for calculating shape loss is shown below:

$$\Omega=\sum_{t=w,h}\left(1-e^{-\omega_t}\right)^{\theta} \tag{15}$$

$$\omega_w=\frac{\left|w-w^{gt}\right|}{\max\left(w,w^{gt}\right)},\omega_h=\frac{\left|h-h^{gt}\right|}{\max\left(h,h^{gt}\right)} \tag{16}$$

In this formula, θ affects the degree of control of shape loss in SIoU.

III. Experiments

A. Dataset

At the time of writing, there is no publicly available, high-quality dataset of coal and gangue. Furthermore, the complete laboratory environment does not provide a superior response to that of the coal mine site environment. Consequently, the coal and gangue dataset used in this paper is derived primarily from the Inner Mongolia coal mine site conditions and laboratory simulation environment and labeling.

The dataset presented in this paper comprises 2,775 original images, distributed as follows: 2,369 images in the training set, 304 images in the validation set, and 102 images in the test set. All images in the dataset have been resized to 640×640 pixels. The manual image annotation process has been conducted using LabelImg, and the resulting labeling division of the training set is illustrated in Fig. 13.

The uneven distribution of coal and gangue on the conveyor belt, as observed on the coal mine site, has been taken into account in the construction of the dataset. The actual number of coal has been increased in order to achieve a closer match with the on-site production situation. Due to the inherent ambiguity in the spatial distribution of coal and gangue on the conveyor belt, these materials are represented in the image at varying positions and weights to elicit the model's comprehensive attention to the image as a whole. The number of specific labels is presented in Table I.

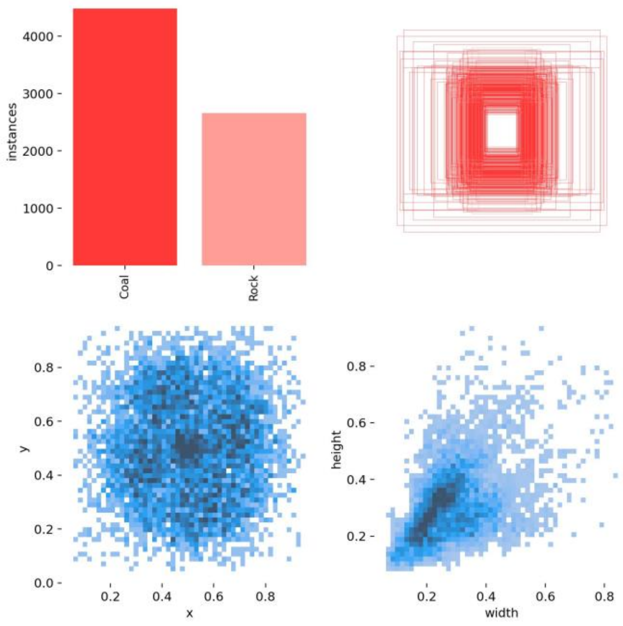

Fig. 13. Visualization of the dataset labels.

TABLE I
STATISTICS OF DATASET LABELS

Class	Coal	Rock
train	4436	2654
validation	562	323
test	186	101

B. *Experimental configurations*

The GD-YOLO model and other deep learning models are characterised by higher complexity and the utilisation of a greater number of datasets. Consequently, GPUs are employed to accelerate the computation. The details of the experimental equipment employed in this study are presented in Table II.

TABLE II
COMPUTER ENVIRONMENT

Item	Value
CPU	Intel(R) Core(TM) i5-12490F
GPU	NVIDIA GeForce RTX 4060
CPU Clock Speed	3.0GHz
Memory	DDR5 5600MHz 32GB
PyTorch	2.0.1
Python	3.8.18
CUDA	11.8.0
Operating System	Win10

The training process of each parameter is of significant importance; thus, the selection of appropriate parameters is of paramount importance, directly influencing the model's ability to detect the final results of coal and gangue. Following comprehensive experimentation, this paper employs the model parameters as illustrated in Table III.

TABLE III
PARAMETERS

Parameters	Value
Epochs	200
Input size	640
Optimizer	Adam
Lr0	0.01
Weight decay	0.0005
Batch	16

C. *Evaluation Metrics*

In order to evaluate the model performance in a more comprehensive and accurate manner, this paper employs a range of metrics, including the number of parameters (Params), the number of billion floating point operations per second (GFLOPs), precision, recall, average precision (AP, mAP), and frames per second (FPS).

The precision and recall results are comprised of the four results presented in the confusion matrix in Table IV.

TABLE IV
PARAMETERS

Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The aforementioned four results are calculated in order to ascertain the precision and recall. The calculation process is illustrated in the following section.

$$\Pr ecision = \frac{TP}{TP + FP} \tag{17}$$

$$\operatorname{Re} call = \frac{TP}{TP + FN} \tag{18}$$

The AP value is calculated by integrating the precise value corresponding to each recall point on the precision-recall curve. The calculation is as follows:

$$AP = \int_0^1 p(r)dr \tag{19}$$

The mAP is calculated by averaging the AP values of each category, which is an important indicator for evaluating the effectiveness of the model in detecting coal and gangue. The calculation process is outlined below:

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \tag{20}$$

D. *Position comparison experiments*

Once the C2f module has undergone enhancements and optimizations, it is essential to consider the issue of location selection in order to fully leverage the potential of the model. To this end, an experiment was conducted, in which the C2f module was replaced in various locations within the Backbone, as illustrated in Fig. 14. In this experiment, the C2f-Faster-EMA was implemented in place of the C2f module in locations a, b, c, d, and in the entirety of the C2f module within the Backbone.

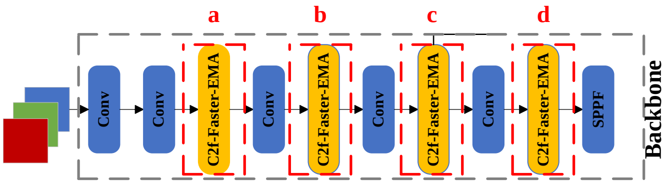

Fig. 14. C2f-Faster-EMA Replacement Location.

Following the implementation of the aforementioned replacement module in accordance with the prescribed methodology, the resulting experimental outcomes are illustrated in Fig. 15.

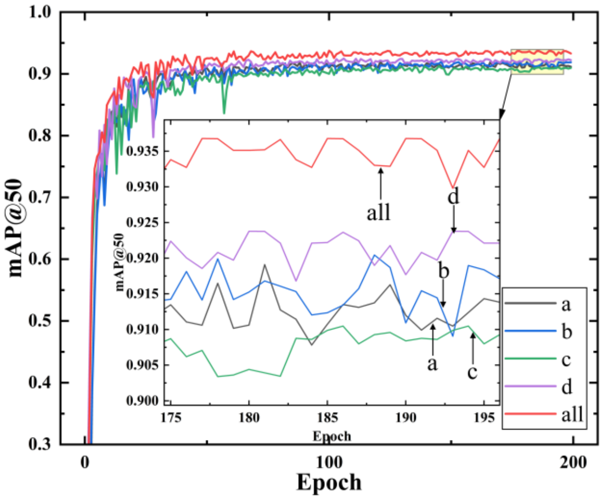

Fig. 15. Comparison of Replacement Positions.

A review of the experimental results indicates that the replacement of all C2f modules in Backbone with C2f-Faster-EMA yields the highest accuracy, with an mAP value of 93.5%.

E. Ablation Experiments

The GD-YOLO model proposed in this paper exhibits enhanced accuracy and convergence speed. However, the comprehensive performance of the overall improved model requires verification through the design of ablation experiments for comprehensive performance testing. In this paper, YOLOv8n is utilized as the base model, and the parameters employed in the model are presented in Table 3.7. The four improvements implemented as follows:

(1) T: Neck part of FPN+PAN architecture is improved to GD architecture.

(2) U: Bottleneck improved to FasterNet Block.

(3) V: EMA mechanism is introduced to enhance the model's feature information extraction ability for coal and gangue.

(4) W: CIoU Loss is improved to SIoU Loss.

The model performance evaluation indexes are carried out by Params, GFLOPs, Precision, Recall, AP, mAP@50, and FPS. The experimental results are shown in Table V.

TABLE V
RESULTS OF ABLATION EXPERIMENTS

T	U	V	W	P(%)	mAP@50(%)	Params (M)	GFLOPs(G)	FPS (f/s)
×	×	×	×	91.2	91.7	3.1	8.7	82.1
√	×	×	×	94.7	94.9	6.5	13.8	75.6
×	√	√	×	92.3	93.5	2.5	6.9	70.2
×	×	×	√	91.2	91.7	3.1	8.7	81.6
√	√	√	√	95.7	96.6	6.1	12.3	61.5

The ablation experiments demonstrate that after improving the traditional FPN+PAN architecture in the model to a GD architecture, there is a 3.5% improvement in precision and a 3.3% improvement in recall, resulting in a 3.2% improvement in mAP@50. These findings indicate that enhancing the feature fusion network can significantly enhance the accuracy of the model. Following the enhancement of the Bottleneck to FasterNet Block, the C2f-Faster-EMA exhibited a reduction in Params and GFLOPs by 19.4% and 20.6%, respectively, in comparison to C2f. This indicates a decrease in the model's parameters and complexity, which can mitigate the adverse impact of superfluous information on the model and enhance its capability for feature extraction. Following the improvement of CIou to SIou, the precision, recall, and mAP@50 of the model exhibited varying degrees of enhancement. The integration of these modalities demonstrated the optimal performance of the model, with precision improving by 4.5%, recall improving by 4.4%, and mAP@50 improving by 4.9%. Despite the reduction in model detection speed, these outcomes still satisfy industrial production requirements.

To more effectively illustrate the impact of the enhanced GD-YOLO model, a comparative analysis was conducted between the improved GD-YOLO model and the initial YOLOv8 model. This involved the generation of mAP curves based on 200 epochs, with identical parameters and operational environments, as illustrated in Fig. 16.

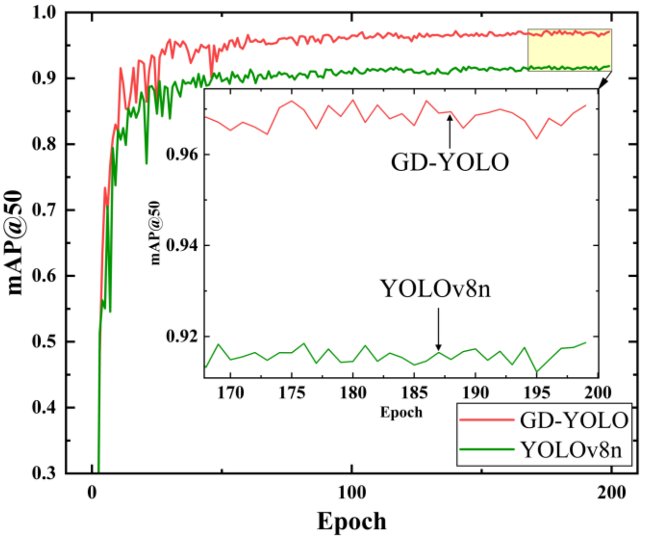

Fig. 16. Comparison of model mAP values.

To enhance the interpretability of the enhanced model, we conducted experiments utilizing a GradCAM visualization to ascertain the model's responsiveness to coal and gangue, as well as its capacity to integrate features. The findings from these experiments are presented in Fig. 17.

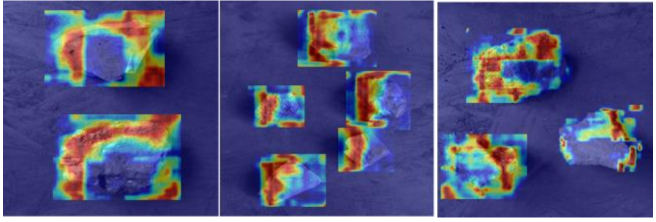

Fig. 17. GradCAM Experiment.

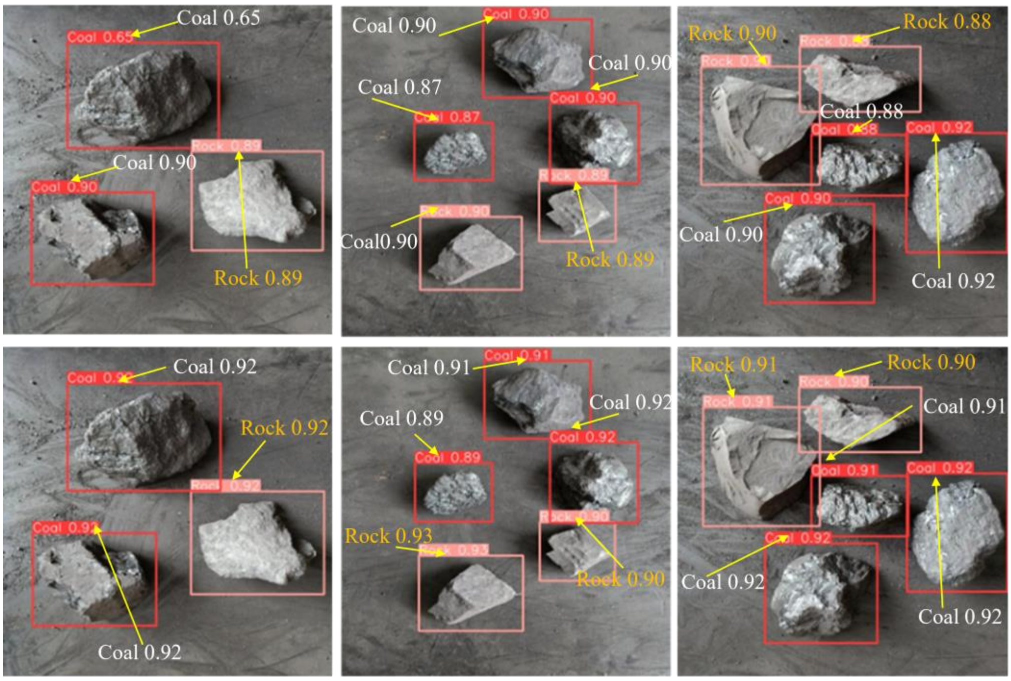

Fig. 18. Comparison of model detection results. (The initial row presents the outcomes of the YOLOv8n model detection, while the subsequent row illustrates the results of GD-YOLO model detection.)

TABLE VI
COMPARISON OF EXPERIMENTAL RESULTS

Model	P(%)	R(%)	Map@50(%)	Params(M)	GFLOPs(G)	FPS(f/s)
YOLOv5n	83.3	82.6	84.5	3.1	8.7	84.5
YOLOv7-Tiny	87.6	86.8	88.7	6.2	5.8	92.4
YOLOvX-Tiny	88.5	87.6	89.1	5.1	6.5	89.1
YOLOv6n	91.3	90.2	91.8	4.7	11.4	99.2
YOLOv8n	91.2	90.1	91.7	3.1	8.7	82.1
GD-YOLO	95.7	94.5	96.6	6.1	12.3	61.5

As illustrated by the heat map visualization results, the GD-YOLO model exhibits high feature sensitivity, demonstrating an ability to discern and attend to the distinctive characteristics of coal and gangue. Additionally, it displays a noteworthy aptitude for integrating feature information, which further substantiates the superiority of the GD-YOLO algorithm model.

In order to directly reflect the actual effect of the final model for the detection of coal and gangue, the photos in the test set for detection were randomly selected for analysis. The resulting experimental results are presented in Fig. 18.

F. Comparative Experiments

In order to evaluate the efficacy of the GD-YOLO algorithm in coal and gangue detection, a comparative analysis was conducted with other models, including YOLOv5n, YOLOv7, YOLOv8n, and Faster-RCNN. The experiments were conducted under identical configuration environments and the results are presented in Table VI. The mAP curves for each model are presented in Figures 19.

While the YOLOv5n, YOLOv6n, YOLOv7-Tiny, YOLOvX-Tiny, and YOLOv8n models demonstrate enhanced detection speeds, they exhibit reduced accuracy, which is inadequate for meeting the specified application requirements. In contrast, the results of the validation experiments demonstrate the efficacy of GD-YOLO, with a mean accuracy (mAP) of 96.6%, which is 4.9% higher than that of the traditional YOLOv8n.

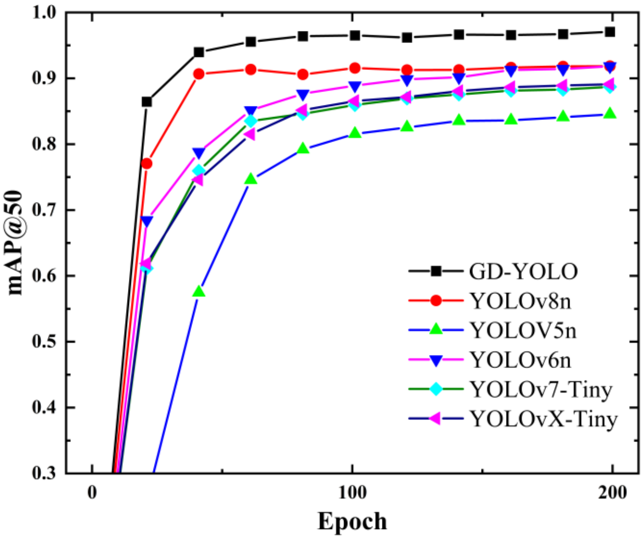

Fig. 19. The mAP curves for each model

IV. Conclusion

In order to address the issue of gangue detection, a novel GD-YOLO algorithm model has been proposed. This comprises a feature extraction network, a feature fusion network, a detection layer, and a loss function. Additionally, the bottleneck has been enhanced through the incorporation of a FasterNet Block, which is based on PConv convolution, with the objective of enhancing the model's capacity to extract spatial features. Concurrently, the EMA attention mechanism is integrated into the enhanced C2f-Faster, resulting in the creation of C2f-Faster-EMA. Subsequent position comparison experiments are conducted to ascertain the optimal replacement position, thereby further enhancing the detection performance. Furthermore, the issue of the model's slow convergence speed and poor performance can be effectively addressed by modifying the loss function to SIoU. Ablation and comparison experiments demonstrate that the mAP of the algorithm is 96.6%, a 4.9% improvement over the traditional YOLOv8n algorithm. Additionally, the GD-YOLO algorithm achieves a FPS of 61.5, conferring a notable advantage in terms of accuracy compared to current mainstream algorithms. These findings indicate that the algorithm has considerable potential for industrial applications.

References

[1] Yutao, W. (2022). Status and prospect of harmless disposal and resource comprehensive utilization of solid waste of coal gangue. Coal Geology & Exploration, 50(10), 54-66.

[2] 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.

[3] Xian, Y., Tao, Y., Ma, F., & Zhou, Y. (2022). The study of enhanced gravity concentrator for maceral enrichment of low-rank coal with heavy medium. International Journal of Coal Preparation and Utilization, 42(12), 3777-3793.

[4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[5] Xie, L., & Yuille, A. (2017). Genetic cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1379-1388).

[6] Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS journal of photogrammetry and remote sensing, 173, 24-49. Alzubaidi L, Zhang J, Humaidi A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. Journal of big Data, 2021, 8: 1-74.

[7] Alexey D. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv: 2010.11929, 2020.

[8] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[9] Song, B., Wu, Y., & Xu, Y. (2024, March). ViTCN: Vision Transformer Contrastive Network For Reasoning. In 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) (pp. 452-456). IEEE.

[10] Yang, J., Liu, J., Xu, N., & Huang, J. (2023). Tvt: Transferable vision transformer for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 520-530).

[11] 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.

[12] Yan, P., Sun, Q., Yin, N., Hua, L., Shang, S., & Zhang, C. (2022). Detection of coal and gangue based on improved YOLOv5. 1 which embedded scSE module. Measurement, 188, 110530.

[13] Guo, Y., Zhang, Y., Li, F., Wang, S., & Cheng, G. (2023). Research of coal and gangue identification and positioning method at mobile device. International Journal of Coal Preparation and Utilization, 43(4), 691-707.

[14] Wang, S., Zhu, J., Li, Z., Sun, X., & Wang, G. (2024). GDPs-YOLO: an improved YOLOv8s for coal gangue detection. International Journal of Coal Preparation and Utilization, 1-14.

[15] Talaat, F. M., & ZainEldin, H. (2023). An improved fire detection approach based on YOLO-v8 for smart cities. Neural Computing and Applications, 35(28), 20939-20954.

[16] Xiao, B., Nguyen, M., & Yan, W. Q. (2024). Fruit ripeness identification using YOLOv8 model. Multimedia Tools and Applications, 83(9), 28039-28056.

[17] Soylu, E., & Soylu, T. (2024). A performance comparison of YOLOv8 models for traffic sign detection in the Robotaxi-full scale autonomous vehicle competition. Multimedia Tools and Applications, 83(8), 25005-25035.

[18] Chen, X., Wang, M., Ling, J., Wu, H., Wu, B., & Li, C. (2024). Ship imaging trajectory extraction via an aggregated you only look once (YOLO) model. Engineering Applications of Artificial Intelligence, 130, 107742.

[19] Gaikwad, D. P., Sejal, A., Bagade, S., Ghodekar, N., & Labade, S. (2024). Identification of cervical spine fracture using deep learning. Australian Journal of Multi-Disciplinary Engineering, 1-9.

[20] Hussain, M. (2023). YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines, 11(7), 677.

[21] Sohan, M., Sai Ram, T., Reddy, R., & Venkata, C. (2024). A review on yolov8 and its advancements. In International Conference on Data Intelligence and Cognitive Informatics (pp. 529-545). Springer, Singapore.