scientific reports

Check for updates

OPEN

A novel deep learning model based on YOLOv5 optimal method for coal gangue image recognition

Tongkai Gu[1,2], Haiyan Zhao[3], Yasheng Chang[2]✉, Sitong Yan[4], Feihan Cao[2] & Wei Liu[5]

Coal gangue recognition presents significant challenges in the mining industry due to its inefficient and costly traditional treatment methods. The advent of deep learning techniques has introduced novel solutions for automating and online coal gangue processing. Despite the potential of deep learning models, challenges such as overfitting and the need for extensive labeled datasets hinder their effectiveness. You Only Look Once version 5 (YOLOv5), with its rapid inference speed and high accuracy, offers a suitable solution for real-time coal gangue detection. This research investigates the application of YOLOv5 for coal gangue image recognition, involving data preprocessing, model training, and optimization. Experimental results demonstrate that incorporating the multiple channel attention mechanism and lightweight content-aware re-assembly of features up-sampling operator significantly improves model confidence and recognition performance.

Keywords Coal gangue detection, YOLOv5, Overfitting, Preprocessing, Detection accuracy

Coal is a significant energy resource; however, during the mining process, the inclusion of coal gangue reduces the calorific value of coal and compromises its quality. Coal gangue contains heavy metals and other hazardous substances, and its combustion releases many harmful gases, leading to severe environmental pollution. The separation of coal gangue not only enhances coal quality but also mitigates environmental pollution and promotes the efficient utilization of resources.

The traditional manual gangue sorting method is inefficient, labour-intensive and prone to misjudgment and omission. Although the mechanical methods of gangue sorting can significantly reduce misjudgment, they involve complex equipment[1,2], high costs, and stringent requirements regarding the particle size of the gangue[3]. Examples of such methods include heavy media sorting, jigging sorting, and other similar techniques.

Machine vision-based inspection technology does not require complex mechanical equipment[4,5]. It captures images of coal and gangue using image acquisition devices and extracts features through image processing algorithms[6,7], enabling the identification and localization of gangue[8]. However, the detection accuracy of this technology for coal gangue remains insufficient. With the rapid advancement of artificial intelligence, accurate detection of coal gangue has become feasible[9–11]. By employing artificial intelligence algorithms, such as convolutional neural networks (CNN)[12–14], you only look once (YOLO) series[15–17], and pyramid scene parsing network (PSPNet)[18], the detection accuracy can be enhanced by training models on large datasets of images. However, in the coal production process, coal gangue detection must be completed quickly to meet production efficiency requirements. Therefore, a lightweight network architecture is essential to optimize the detection algorithm, reduce computational complexity, and improve detection speed.

Traditional deep learning models can overfit the training data, especially if the dataset is small or not representative of real-world conditions, leading to poor generalization on new data. Therefore, traditional deep learning models require large amounts of labeled data to perform well[19–21]. However, collecting and labeling sufficient high-quality coal gangue images is expensive and challenging. In addition, deploying models for real-time image recognition in mining operations demands efficient algorithms and sufficient processing power to ensure timely and accurate recognition[22–24]. YOLOv5 offers very fast inference speeds while maintaining high accuracy, making it well-suited to the real-time and rapid requirements of detecting coal gangue[25–27]. Compared

[1]School of Mechanical and Electrical Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China. [2]School of Optical and Electronic Information & Suzhou Key Laboratory of Biophotonics & International Joint Metacenter for Advanced Photonics and Electronics, Suzhou City University, Suzhou 215104, China. [3]School of Architecture & Design, Kunshan Dengyun College of Science and Technology, Suzhou 215300, China. [4]School of Computer Science and Technology, Soochow University, Suzhou 215006, China. [5]School of Optoelectronic Information Engineering, Soochow University, Suzhou 215006, China. ✉email: cocys@126.com

with other versions, the YOLOv5 model can be quickly deployed on resource-constrained devices without sacrificing too much accuracy, meeting the lightweight requirements of detecting coal gangue.

The identification of coal gangue targets based on deep learning requires first recognizing and then locating the target. Therefore, the following steps are researched as follows: (1) Investigate the YOLOv5 model in deep learning to propose a method for coal gangue image recognition. (2) Preprocess the data, handle anomalies, and ensure that the images meet the requirements of the coal gangue recognition model. Establish an image dataset based on coal gangue image data and preprocessed image data. (3) Train the target recognition model using the dataset to locate coal gangue and annotate it with rectangular bounding boxes through deep learning target recognition. (4) Use optimization methods to enhance target recognition and process the data images obtained through experiments. (5) Validate the model. The overview of the workflow is shown in Fig. 1.

Problems were encountered during the experiment, such as the selection of data images during the production of the dataset and the optimization module added to the recognition model. These problems were theoretically feasible and logically rigorous, but they could not be carried out in practice. This was mainly because the initially selected optimization module did not have a significant recognition effect on the data image selection, which caused the prediction time to increase. To solve these problems, the parameters were continuously corrected, and the multiple channel attention (MCA) mechanism[28,29] and lightweight content-aware re-assembly of features (CARAFE) up-sampling operator[30,31] were added.

The YOLOv5 optimal model improves the precision (P) value for recognizing coal and rock from a baseline of 0.963 to 0.966, the recall (R) value from 0.954 to 0.959, and the mean average precision (mAP) value from 0.975 to 0.977. The results show that the confidence of the optimal model is significantly higher than that of the basic model, and the recognition effect is significantly improved.

Model

The experiment was based on the YOLOv5 model with improvements introduced in the feature enhancement stage, incorporating the MCA mechanism and the lightweight CARAFE up-sampling operator.

Principles of the YOLOv5 model

YOLOv5 is an object detection algorithm based on deep learning technology. It utilizes components such as the CSPDarknet53 backbone network, feature pyramid structure, lightweight detection head, and anchor boxes, among others, to achieve efficient object detection[32,33]. The model performs forward propagation to compute bounding boxes and class confidence scores, optimized through improved activation functions and loss functions, ultimately achieving high detection speed and accuracy.

Backbone network: YOLOv5 uses CSPDarknet53 as its backbone network, known for its lightweight Darknet architecture with high performance and computational efficiency. The network employs the cross-stage partial (CSP) network structure to split and process input feature maps in parallel, enhancing information propagation efficiency and feature reuse.

Feature pyramid: A computer vision technique used for multiscale object detection and image segmentation, YOLOv5 incorporates a feature pyramid structure to fuse feature maps from different levels, enabling the detection of objects at different scales. Detecting objects across different feature map levels improves the model's capability to handle objects of varying sizes.

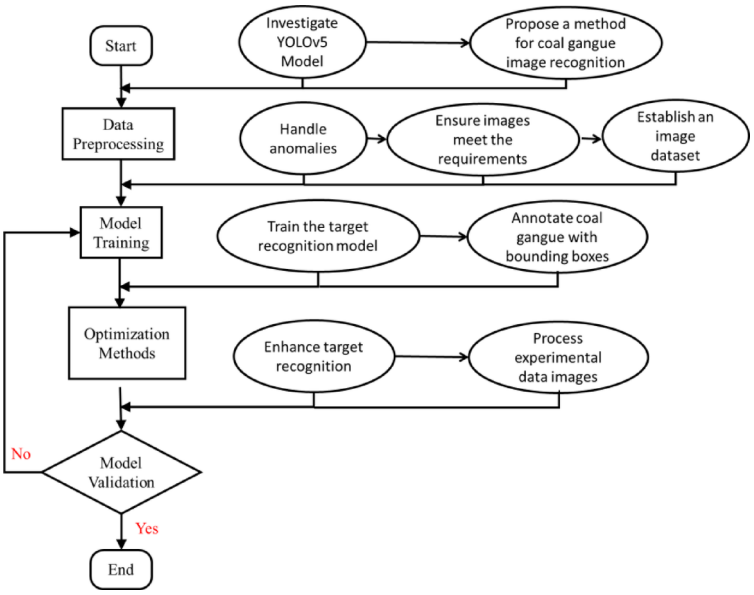

Fig. 1. The flowchart of this work.

Detection head: YOLOv5 adopts a lightweight detection head structure responsible for generating detection bounding boxes and class confidence scores. The detection head includes a series of convolutional layers, fully connected layers, and activation functions for predicting bounding box positions and class probabilities.

Anchor boxes: These are predefined bounding boxes used to adjust for object shape and scale by predicting offsets and scale information. YOLOv5 integrates Anchor boxes to enhance the model's detection capabilities across different object shapes and scales[34].

MCA attention mechanism

The MCA mechanism is an attention mechanism used in deep learning models. It aims to enhance the model's learning ability to correlate features across different channels, thereby improving its performance on specific tasks. The core idea is to dynamically learn the importance of each channel in the feature map using attention mechanisms. This mechanism then integrates these weighted features to extract richer and more effective feature representations. By effectively capturing channel correlations in image features, MCA enhances the representation capability of deep learning models.

In traditional attention mechanisms, attention weights are typically computed in the spatial dimension. In contrast, MCA focuses on weighting attention across channel dimensions. This approach allows the model to flexibly learn correlations between different channels, thereby improving its ability to represent input data. In practice, MCA typically involves the following steps: (1) Channel segmentation: First, the input features are segmented into multiple channels, each containing a set of features. (2) Compute attention weights: For each channel, attention weights are computed using an attention mechanism. Typically, this involves linear transformations of features within the channel to obtain the attention distribution. (3) Weighted feature fusion: Multiply the attention weights of each channel by the features within that channel. Sum these weighted features across all channels to obtain the final weighted feature representation. (4) Parallel computation with multiple heads: Multiple attention heads are often introduced to enhance representation capability. These heads compute attention weights and weighted feature representations in parallel. The outputs from these multiple heads are then concatenated or aggregated to obtain the final output feature representation.

Lightweight CARAFE up-sample operator

The lightweight CARAFE up-sampling operator is an up-sampling algorithm based on reversible convolution, aimed at enhancing the efficiency and accuracy of deep learning models during up-sampling[35]. Compared to traditional up-sampling methods like bilinear interpolation or transposed convolution, CARAFE offers lower computational complexity and higher up-sampling quality. Its core idea is to use reversible convolution for up-sampling while integrating a local feature reassembly mechanism to preserve detailed information in feature maps.

Specifically, CARAFE achieves more accurate and detailed up-sampling results by reassembling features from local receptive fields during the up-sampling process. This content-aware reassembly approach allows CARAFE to better preserve semantic information and spatial structure of feature maps, avoiding potential blurring and distortion issues associated with traditional up-sampling methods. In practice, CARAFE typically involves the following steps: (1) Reversible convolution up-sampling: First, the input feature map undergoes up-sampling using reversible convolution. Reversible convolution is a specialized convolution operation that can enlarge the size of feature maps without losing information. (2) Feature re-assembly: For each pixel position after up-sampling, CARAFE reassembles features from local receptive fields. Specifically, it calculates local receptive fields based on the pixel position in the original feature map and uses these local features for reassembly to generate the feature representation at the target position. (3) Re-assembly weight calculation: During feature reassembly, CARAFE computes reassembly weights for each pixel position to determine how local features are utilized for reassembly. These reassembly weights are typically calculated based on spatial position and feature similarity information to ensure accuracy and fidelity in reassembly. (4) Feature fusion: Finally, CARAFE integrates the feature map obtained from feature re-assembly with the up-sampled feature map to produce the final up-sampling result.

Optimized model based on YOLOv5

The YOLOv5 model has been enhanced to address the challenges of coal gangue image recognition tasks. As part of the improvement, the lightweight CARAFE up-sampling operator was chosen. CARAFE is a lightweight up-sampling operator that effectively increases the receptive field and enhances the utilization of semantic information from feature maps. This allows the model to maintain detection accuracy while reducing computational complexity and accelerating recognition.

Next, the MCA attention mechanism is introduced during the feature enhancement phase. The MCA mechanism aids the model in integrating high and low-level feature information more effectively, thereby enhancing feature representation and robustness. By incorporating MCA modules into the backbone network, spatial position information encoding is shared, facilitating the fusion of high and low-level feature information. This enhancement improves the network's feature extraction capability, enabling more accurate localization of target information and further enhancing recognition capability[36].

The optimized model[37] comprises the backbone network, neck network, and prediction network, as illustrated in Fig. 2. MCA modules are integrated into the neck network, while CARAFE modules replace the original up-sampling modules in the backbone. The placement of MCA modules is meticulously adjusted to enhance feature extraction by integrating information across channels, horizontal spatial dimensions, and vertical spatial dimensions. This refinement assists the backbone network in precisely locating target information and enhances its recognition capability.

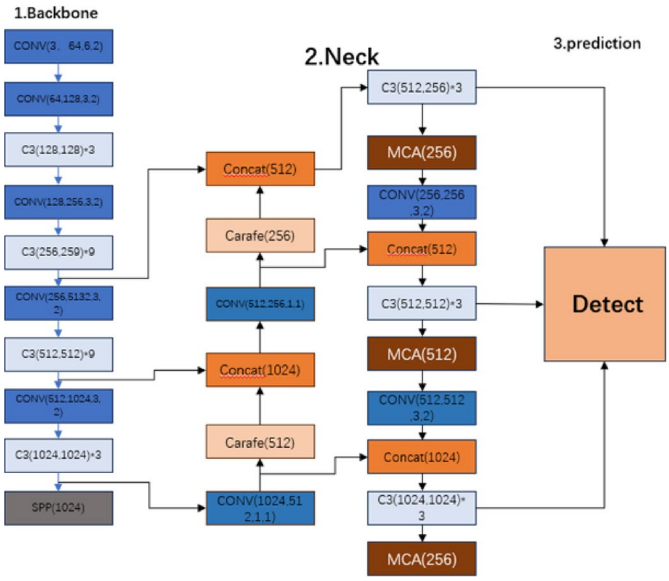

Fig. 2. The flowchart of the optimized model based on YOLOv5.

Fig.3. The raw images of coal gangue.

Experiment and analysis

Collecting and preprocessing of coal gangue image data

The pictures of coal and gangue in the manuscript were obtained by our own camera. We collected 3200 different pictures of coal and gangue for model training and testing. The sample are illustrated in Fig. 3.

Methods were employed to augment the original coal gangue images, enhancing the dataset. Data labeling involves annotating image data by adding information about the location and category of targets in each image, facilitating model training and evaluation. Before training, coal gangue images need to be labeled. The LabelImg tool was used for annotation, as shown in Fig. 4.

After completing the annotation, place the images of the test set and training set into the 'images' folder under the 'train' and 'val' directories, respectively. Similarly, place the image labels into the 'labels' folder under the 'train' and 'val' directories. This completes the creation of the target dataset, as shown in Fig. 5.

Experiment details

The operating system used for this experiment is Windows 10, with an Intel(R) Core (TM) i7-8700 CPU as the core processor and an NVIDIA GTX 2070 as the graphics processor (GPU). The development framework for the program includes Python 3.8.5, CUDA 10.2.89, cuDNN 7.6.5, and PyTorch 1.6.0. Using the concept of transfer learning, the dataset is employed to train the model and obtain pre-trained weight parameters. The batch size for model training is set to 16, and the momentum of learning rate and weight decay are set to 0.934 and 0.0005, respectively. The optimizer used is SGD, with an initial learning rate of 1×10^{-2}. The parameters are detailed in Table 1.

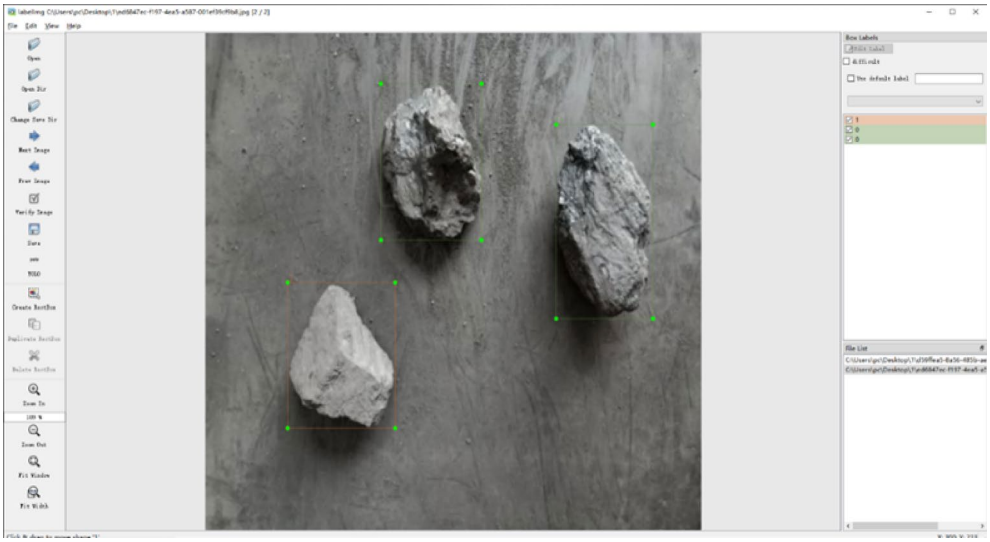

Fig. 4. Annotation of coal and gangue dataset using LabelImg tool.

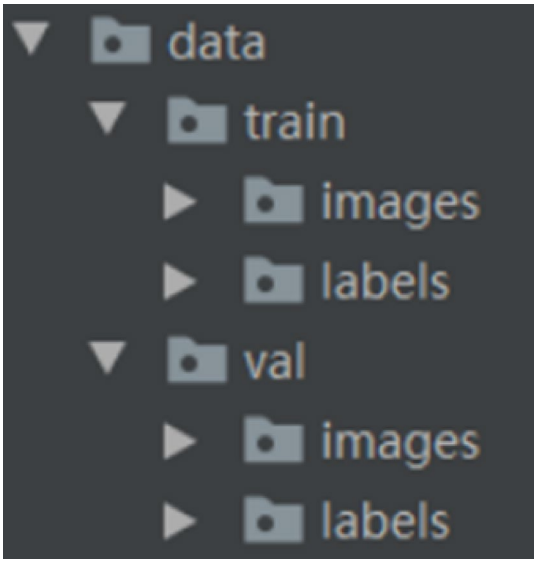

Fig. 5. The dataset classification.

Parameter name	Value
Input resolution (lr0)	640 × 640
Initial learning rate	0.01
Cyclical learning rate (lrf)	0.02
Weight decay coefficient	0.0005
Learning rate momentum	0.934
Detection box position loss coefficient	0.005
Classification loss coefficient	0.5
Object loss coefficient	0.5
Number of training epochs	150
Batch size	16

Table 1. Model training parameters.

This paper evaluates the detection performance of each model on the ARDs-5-TE dataset. For each test image, precision (*P*) and recall (*R*) are calculated by comparing the detection results with the ground truth labels. Further metrics include the F_1 score for each class, which is the harmonic mean of P and R, and the average precision (*AP*) for each class, representing the area under the precision-recall curve. The mean average precision (*mAP*) is then computed by averaging *AP* across all classes, providing an overall measure of the model's detection performance in complex scenarios. Computational complexity is indicated by the number of algorithm parameters (Par, unit in Mb) and FLOPs (floating-point operations, unit in G), where higher Par values imply longer training and inference times, and FLOPs represent the total number of floating-point operations required by the model to process an input instance. Reducing FLOPs helps to improve the speed and efficiency of the model's operation. Inference efficiency is measured by the average inference time per image in the test set (in ms), all computed on a GTX 2070.

Here *TP* represents true positive predictions, *TN* represents true negative predictions, *FP* represents false positive predictions, and FN represents false negative predictions, the calculations are as follows:

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

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

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{3}$$

Experimental results

This experiment is divided into four sets of data: YOLOv5 basic model experiment, YOLOv5-MCA experiment, YOLOv5-CARAFE experiment, and YOLOv5 optimal model experiment. Each experiment tested 310 images, with 2472 images used for training and 309 for validation.

During validation, different types of coal gangue were identified under varying backgrounds. Each image contained numerous coal gangue pieces of different sizes and arrangements. Through the experiments, all targets were successfully detected, validating the model's ability to simultaneously detect multiple types of coal gangue. Figure 6 demonstrates the recognition performance of the YOLOv5 optimal model.

Results analysis

This study conducted a statistical analysis of experimental results, including P, R, and mAP values for three categories: coal, rock, and all. Additionally, it evaluated Par values, FLOPs, and time values (prediction time per single image). The experimental data results from the four experiments are summarized, with partial weight results shown in Table 2 and Fig. 7.

The experimental results are shown in Table 3 and Fig. 8. From top to bottom are the four groups ranging from the basic model to the optimal model. The YOLOv5 optimal model improves the P value for recognizing coal and rock from a baseline of 0.963 to 0.966, with a relative improvement of 0.31% ($(0.966 - 0.963)/0.963 \times 100\% \approx 0.31\%$). The R value improves from 0.954 to 0.959, corresponding to an improvement of 0.52% ($(0.959 - 0.954)/0.954 \times 100\% \approx 0.52\%$), indicating a reduced omission rate in recognizing coal and rock, thereby detecting all relevant targets more comprehensively. The mAP value improves from 0.975 to 0.977, with a relative improvement of 0.2% ($(0.977 - 0.975)/0.975 \times 100\% \approx 0.20\%$). However, due to the increased model complexity, the prediction time per single image has slightly increased. The experimental results demonstrate that the design of this optimal model structure is reasonable, and its recognition performance has been improved.

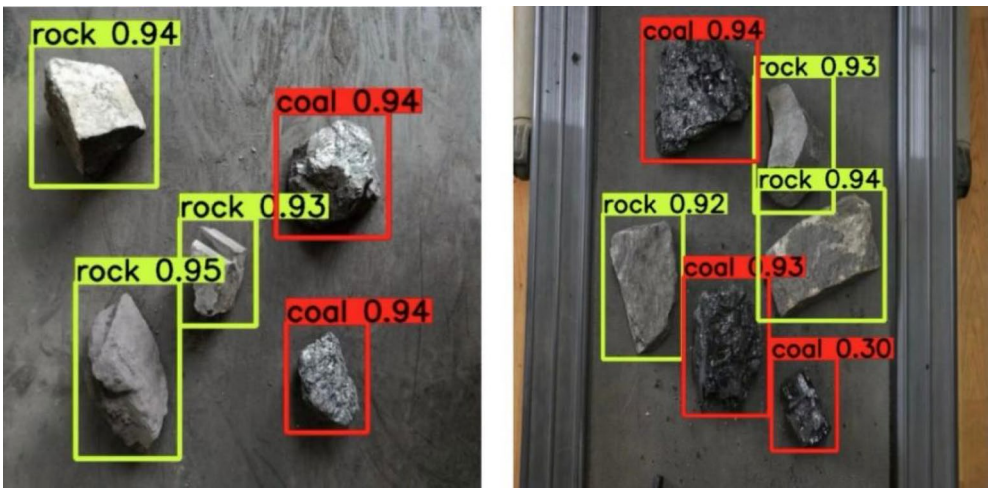

Fig. 6. Recognition results of YOLOv5 optimal model.

Model Summary: 271 layers, 7,182,383 par, 0 gradients, 16.4 G FLOPs						
Class	**Images**	**Labels**	**P**	**R**	**mAP**	**Times (ms)**
All	310	1096	0.966	0.959	0.977	0.801
Coal	310	709	0.962	0.96	0.98	0.793
Rock	310	387	0.971	0.958	0.975	0.81
Speed: 0.6 ms pre-process, 10.2 ms inference, 0.8 ms NMS per image at shape (1,3,640,640)						

Table 2. The Table of partial weight results.

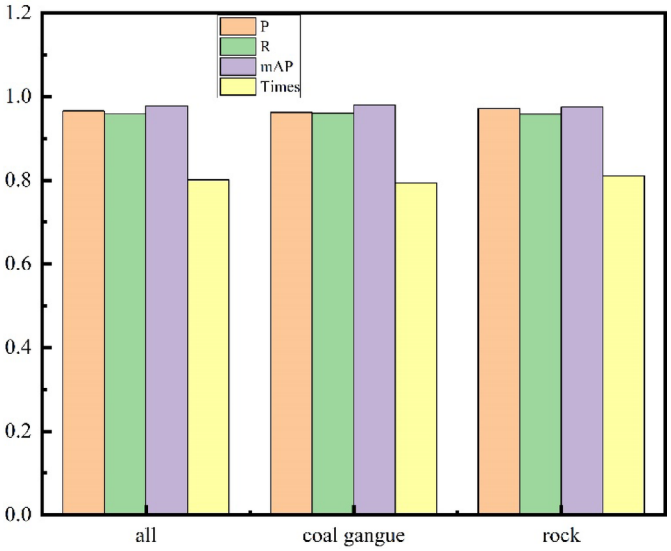

Fig. 7. The bar chart of partial weight results.

		P	R	mAP	Par	FLOPs	Times
YOLOv5 basic	Coal	0.963	0.949	0.973	–	–	–
	Rock	0.963	0.959	0.977	–	–	–
	All	0.963	0.954	0.975	7.03 Mb	16.0 G	6.9 ms
YOLOv5-CARAFE	Coal	0.951	0.961	0.972	–	–	–
	Rock	0.961	0.966	0.979	–	–	–
	All	0.956	0.956	0.976	7.16 Mb	16.4 G	8.0 ms
YOLOv5-MCA	Coal	0.965	0.947	0.974	–	–	–
	Rock	0.959	0.959	0.979	–	–	–
	All	0.971	0.936	0.976	7.06 Mb	16.1 G	7.7 ms
YOLOv5 optimal	Coal	0.962	0.959	0.98	–	–	–
	Rock	0.971	0.96	0.975	–	–	–
	All	0.966	0.959	0.977	7.19 Mb	16.6 G	9.0 ms

Table 3. Experimental data results.

In four verification experiments, different recognition effects of the same image are compared, as shown in Fig. 9. The models of YOLOv5 basic, YOLOv5-MCA, YOLOv5-CARAFE and the YOLOv5 optimal are shown in Fig. 9a–d, respectively. The results show that the confidence of the optimal model is significantly higher than that of the basic model, and the recognition effect is significantly improved, indicating that the structure of this optimized model is reasonable and the recognition ability is feasible.

Conclusions

Based on the construction of a coal gangue image dataset, this research integrates deep learning theories and methodologies to achieve accurate identification of targets within coal gangue images. In the process of target

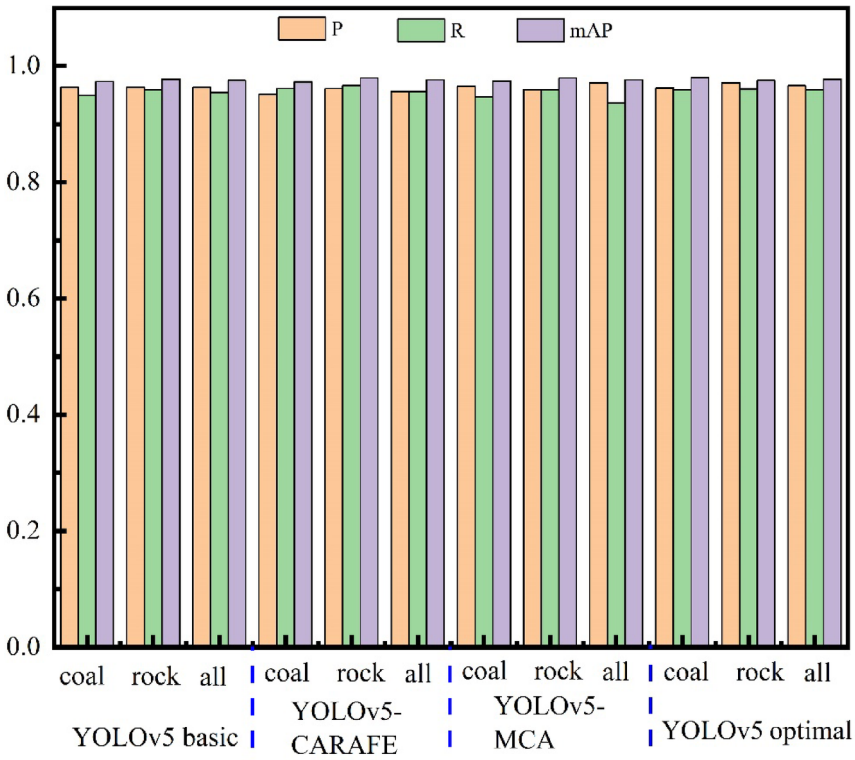

Fig. 8. The bar chart of experimental results.

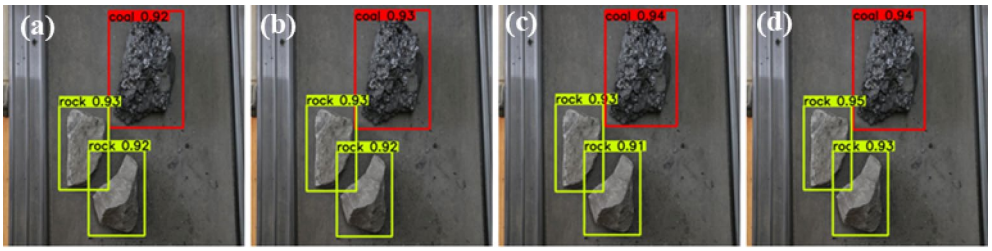

Fig. 9. Experimental verification comparison chart showing results of (**a**) YOLOv5 basic, (**b**) YOLOv5-MCA, (**c**) YOLOv5-CARAFE, and (**d**) YOLOv5 optimal.

recognition, convolutional neural networks, particularly the YOLOv5 optimal model, are employed. Ample and high-quality data support for model training is ensured through data preprocessing and annotation. The novel YOLOv5 optimal model is proposed by adding the MCA attention mechanism and the lightweight CARAFE up-sampling operator. Experimental tests show that the optimal model finally achieve the expected design goal through testing, training, and prediction of the dataset. The YOLOv5 optimal model improves the precision (*P*) value for recognizing coal and rock from a baseline of 0.963 to 0.966, with a relative improvement of 0.31%, the recall (*R*) value from 0.954 to 0.959, corresponding to an improvement of 0.52%, and the mean average precision (mAP) value from 0.975 to 0.977, with a relative improvement of 0.2%. The results can be utilized to identify coal and coal gangue accurately and quickly, with notable improvements in recognition effectiveness.

Data availability

Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the Corresponding author（Yasheng Chang, email: cocys@126.com）upon reasonable request.

Received: 7 September 2024; Accepted: 5 May 2025
Published online: 28 May 2025

References

1. Wang, Z. et al. An online flexible sorting model for coal and gangue based on multi-information fusion. *IEEE Access* **9**(1), 90816 (2021).

2. Li, M. et al. Image positioning and identification method and system for coal and gangue sorting robot. *Int. J. Coal Prep. Util.* **4**, 1–19 (2020).
3. Lv, Z. et al. Fine-grained object detection method using attention mechanism and its application in coal–gangue detection. *Appl. Soft Comput.* **113**(1), 107891 (2021).
4. Zhang, K. et al. Design and application of coal gangue sorting system based on deep learning. *Sci. Rep.* **14**, 16508 (2024).
5. Xudong, Wu. et al. Scheme evaluation method of coal gangue sorting robot system with time-varying multi-scenario based on deep learning. *Sci. Rep.* **14**, 28063 (2024).
6. Li, D. et al. An image-based hierarchical deep learning framework for coal and gangue detection. *IEEE Access* **7**, 184686–184699 (2019).
7. Sun, Z., Huang, L. & Jia, R. Coal and gangue separating robot system based on computer vision. *Sensors* **21**(4), 1349 (2021).
8. Li, Z. et al. 3D location of gangue by point cloud segmentation with RG-TCF. *Int. J. Coal Prep. Util.* **27**, 1–25 (2025).
9. Liu, Q. et al. Recognition methods for coal and coal gangue based on deep learning. *IEEE Access* **9**(99), 77599 (2021).
10. Luo, Q. et al. Adaptive image enhancement and particle size identification method based on coal and gangue. *Meas. Sci. Technol.* **34**(10), 105403 (2023).
11. Gao, J. et al. A coal and gangue detection method for low light and dusty environments. *Meas. Sci. Technol.* **35**(3), 035402 (2024).
12. Lai, W. et al. The study of coal gangue segmentation for location and shape predicts based on multispectral and improved Mask R-CNN. *Powder Technol. Int. J. Sci. Technol. Wet Dry Part. Syst.* **407**, 117655 (2022).
13. Yuanyuan, Pu. et al. Image Recognition of coal and coal gangue using a convolutional neural network and transfer learning. *Energies* **12**(9), 1735 (2019).
14. Lv, Z. et al. Cascade network for detection of coal and gangue in the production context. *Powder Technol.* **377**, 361–371 (2021).
15. Yang, D. et al. Improved YOLOv7 network model for gangue selection robot for gangue and foreign matter detection in coal. *Sensors* **23**(11), 5140 (2023).
16. Yan, P. et al. Detection of coal and gangue based on improved YOLOv5.1 which embedded scSE module. *Measurement* **188**, 110530 (2022).
17. Qin, Y. et al. Intelligent gangue sorting system based on dual-energy X-ray and improved YOLOv5 algorithm. *Appl. Sci.* **14**(1), 98 (2024).
18. Wang, Xi. et al. Rapid detection of incomplete coal and gangue based on improved PSPNet. *Measurement* **201**, 111646 (2022).
19. Zhang, Q. et al. A survey on deep learning for big data. *Inf. Fusion* **42**(1), 146–157 (2018).
20. Safonova, A. et al. Ten deep learning techniques to address small data problems with remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **125**(1), 103569 (2023).
21. Wang, J. et al. Co-training neural network-based infrared sensor array for natural gas monitoring. *Sens. Actuators A* **335**, 113392 (2022).
22. Wei, D. et al. A fast recognition method for coal gangue image processing. *Multimedia Syst.* **29**(4), 2323 (2023).
23. Wang, Xi. et al. Rapid detection of incomplete coal and gangue based on improved PSPNet. *Measurement* **201**(1), 111646 (2022).
24. Xue, G. et al. Coal gangue recognition during coal preparation using an adaptive boosting algorithm. *Minerals* **13**(3), 329 (2023).
25. Yan, P. et al. Detection of coal and gangue based on improved YOLOv5.1 which embedded scSE module. *Measurement* **188**(1), 110530 (2022).
26. Wen, X. et al. A swin transformer-functionalized lightweight YOLOv5s for real-time coal–gangue detection. *J. Real-Time Image Proc.* **20**(3), 47 (2023).
27. Wang, S. et al. Coal gangue target detection based on improved YOLOv5s. *Appl. Sci.* **13**(20), 11220 (2023).
28. Tang, H., Torr, P. H. S. & Sebe, N. Multi-channel attention selection gans for guided image-to-image translation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(5), 6055–6071 (2022).
29. Han, Q., Dan, Lu. & Chen, R. Fine-grained air quality inference via multi-channel attention model. *JCAI* **8**, 2512–2518 (2021).
30. Wang, J., Chen, K., Xu, R. et al. Carafe: Content-aware reassembly of features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019).
31. Wang, J. et al. CARAFE++: Unified content-aware reassembly of features. *IEEE Trans. Pattern Anal. Mach. Intell.* **9**(1), 44 (2022).
32. Li, F. et al. MSF-CSPNet: A specially designed backbone network for faster R-CNN. *IEEE Access* **12**, 52390 (2024).
33. Sun, Yu. et al. A dense feature pyramid network for remote sensing object detection. *Appl. Sci.* **12**(10), 4997 (2022).
34. Gao, M. et al. Adaptive anchor box mechanism to improve the accuracy in the object detection system. *Multimed. Tools Appl.* **78**, 27383–27402 (2019).
35. Wang, X. et al. Deep-learning-based sampling position selection on color doppler sonography images during renal artery ultrasound scanning. *Sci. Rep.* **14**(1), 11768 (2024).
36. Wang, L. et al. Enhancing hazardous material vehicle detection with advanced feature enhancement modules using HMV-YOLO. *Front. Neurorobot.* **18**, 1351939 (2024).
37. Li, He. et al. Design of field real-time target spraying system based on improved yolov5. *Front. Plant Sci.* **13**, 1072631 (2022).

Acknowledgements

This work is supported in part by the School-level Scientific Research Start-up Project (No.007/1960323099) and National Natural Science Cultivation Project (No. x20230039), Xi'an University of Architecture and Technology, the China Postdoctoral Science Foundation(No. 2024MD753960), the Xi'an Beilin District 2024 Applied Technology R&D Reserve Project(No. GX2421), and in part by the Open Projects of State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University (Nos. sklms2023011 and sklms2023009), and in part by the Open Project of State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines under Grant (SKLMRDPC23KF20), and in part by Suzhou Basic Research Project (SJC2023003).

Author contributions

Tongkai Gu: manuscript text, Funding Support Haiyan Zhao: manuscript text, figures, Funding Support Yasheng Chang: Algorithm development Feihan Cao: Algorithm development Sitong Yan: Algorithm development Wei Liu: Algorithm development.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.C.