International Journal of Advanced Engineering and Technology Research
ISSN: 3106-8677 (Print), ISSN: 3106-8685 (Online) | Vol. 2, No. 1, 2026

Research on a Lightweight Vision-based Monitoring Method for Rigid Guide Rail Joint Defects in Coal Mine Shafts

Bingchan Li *

School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, 454003, China

* **Corresponding author**: (Email: 1102177637@qq.com)

Abstract: Prolonged mechanical impacts on mine shaft rigid guide rails inevitably induce joint gaps and misalignments. However, existing visual inspection methods struggle to strike an optimal balance among detection accuracy, inference speed, and edge-deployability. To address these challenges, we propose YOLOv10n-LBC, a highly efficient, integrated model tailored for defect detection, gap measurement, and rail edge extraction on resource-constrained edge devices. Specifically, the network incorporates a Lightweight Ghost Convolutional CSP (LGC-CSP) module to compress the architecture, a Cross-Scale Decoupled (CSD) head to enhance multi-scale feature fusion, and a BiFormer attention mechanism to compensate for feature degradation caused by the lightweight design. Following detection, joint gap widths are quantified using a pixel-to-physical ratio methodology, while misalignments are identified via depth estimation-based edge extraction. Extensive experiments on a custom 6,100-image dataset demonstrate that YOLOv10n-LBC achieves a Precision of 98.6%, a Recall of 97.1%, and an mAP@0.5:0.95 of 72.2%. Despite its robust performance, the model maintains a highly compact footprint with only 1.57 M parameters and a 3.4 MB weight file. Real-world mine field tests validate the system's stability, maintaining maximum gap measurement errors below 1 mm, thereby offering a reliable solution for intelligent underground shaft inspection.

Keywords: Rigid guide rails, gap measurement, YOLOv10n, edge extraction, lightweight model.

1. Introduction

Rigid guide rails in coal mine hoisting systems operate under harsh conditions (e.g., impact loads and geological deformations), making them prone to excessive joint gaps, misalignment, and tilting. These defects induce severe vibrations, accelerate component wear, and compromise operational safety, potentially causing catastrophic accidents like conveyance derailment.

Although China's Coal Mine Safety Regulations mandate daily and monthly inspections, manual checks are inefficient and prone to subjective errors, exacerbated by harsh underground environments featuring high dust and humidity. This traditional approach fails to meet the real-time monitoring demands of intelligent mine construction [1]. Therefore, automated and intelligent inspection of shaft guide rails has become an inevitable trend.

Current rigid guide rail detection technologies primarily fall into three categories:

(1) Vibration characteristic analysis. This approach infers structural health from vibration signals using methods like time-frequency image conversion [2], multi-parameter feature extraction [3], and fiber-optic listening [4]. However, it suffers from poor defect localization, environmental noise sensitivity, and reliance on high-precision sensors.

(2) Laser scanning. This method captures high-precision 2D or 3D geometric profiles to extract deformation parameters [5, 6]. Despite its high measurement accuracy, this approach is limited by expensive equipment, massive point-cloud processing latency, and poor real-time performance.

(3) Vision-based object detection. This rapidly evolving approach utilizes deep learning to identify defects directly from images. Recent studies predominantly focus on optimizing YOLO architectures (e.g., YOLOv5, v7, and v8) by integrating lightweight modules and attention mechanisms [7-11]. These innovations aim to improve real-time processing, environmental robustness, and the balance between accuracy and computational efficiency.

Detecting guide rail joint gaps in coal mine shafts requires both high accuracy and real-time processing. While improved YOLO models [12–16] excel in controlled settings, their performance degrades in real underground environments. Here, environmental interference and limited edge computing resources make achieving a lightweight design without sacrificing accuracy a critical challenge. To address this, we propose YOLOv10n-LBC, an enhanced YOLOv10n [17] architecture for joint gap detection, validated through actual mine inspection videos.

2. YOLOv10n-LBC Detection Algorithm

2.1. Network Architecture of YOLOv10n-LBC

To balance detection accuracy and real-time performance for rigid guide rail joint gap detection in challenging mine environments, we propose YOLOv10n-LBC, a novel network whose architecture is shown in Figure 1. The proposed architecture introduces three key innovations. First, a Lightweight Ghost Convolutional CSP (LGC-CSP) module replaces the C2f structure, using dynamic convolution for inference-time efficiency and channel scaling to adjust model complexity. Second, a Cross-Scale Decoupled (CSD) head improves localization and classification accuracy for joint gaps. Third, the Biformer (Bi-Level Routing Attention) mechanism [18] is introduced to complement PSA attention, enhancing feature representation in complex scenes. Together, these innovations balance high accuracy with real-time performance, meeting the requirements of resource-

constrained edge deployment.

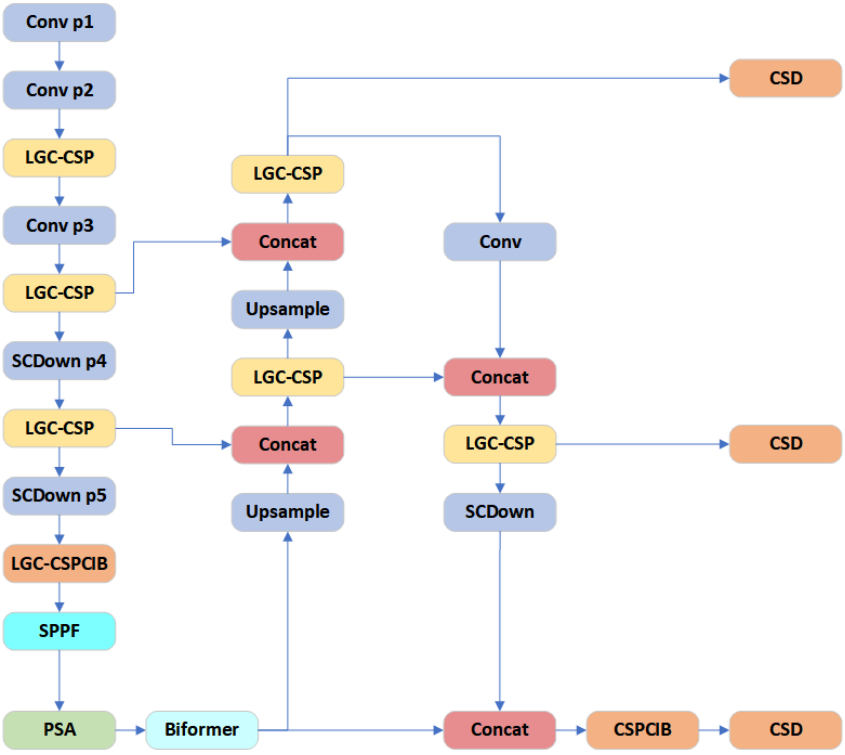


Figure 1. Structural framework of YOLOv10n-LBC

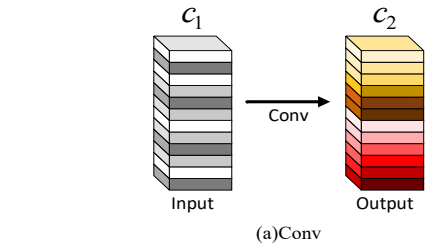
2.2. Lightweight Ghost Convolutional CSP (LGC-CSP)

Gap detection in mine shafts demands lightweight models with robust feature extraction to overcome inconsistent illumination and limited edge resources. However, YOLOv10n's standard C2f module incurs heavy parameter and computational overhead relative to these constraints, restricting its deployment on edge devices.

To address these constraints, we propose the Lightweight Ghost Convolutional CSP (LGC-CSP) module, building upon the CSPNet architecture [19].

The LGC-CSP module employs GhostConv [20, 21] as its core operator. Unlike standard dense convolutions, GhostConv generates a portion of the feature maps via linear transformations, substantially reducing computational complexity while effectively preserving representational capacity.

Specifically, GhostConv operates in two stages: a standard convolution initially generates half of the output channels as primary feature maps, followed by lightweight depthwise separable convolutions that produce the remaining channels. This feature reuse strategy inherently minimizes parameter overhead, driving the aforementioned computational efficiency. Figure 2 compares GhostConv with standard convolution.



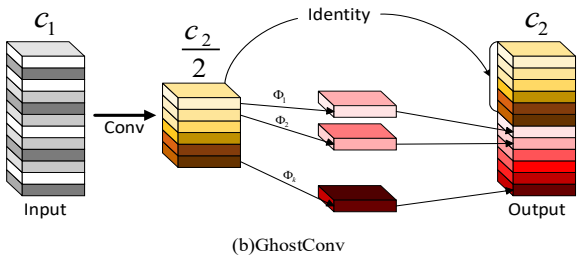


Figure 2. Structure of GhostConv and Conv

Given an input feature map $X \in R^{C_{in} \times H \times W}$, the computational costs of standard convolution (Cal_c) and GhostConv (Cal_g) are formulated as follows:

$$Cal_c = C_{in} \times K^2 \times H \times W \times C_{out} \quad (1)$$

$$Cal_g = C_{in} \times K^2 \times H \times W \times C_{out} \times s^{-1} + (s-1) \times K_2^2 \times H \times W \times C_{out} \times s^{-1} \quad (2)$$

Where C_{in} and C_{out} denote the number of input and output channels, $H \times W$ represents the spatial dimensions, K is the standard kernel size, s is the expansion factor, and K_2 is the kernel size of the GhostConv branch.

The module incorporates Reparameterized Convolution (RepConv) [22] to resolve the trade-off between feature representation capacity and inference speed via a heterogeneous architecture. Specifically, multi-branch fusion enhances feature diversity during training, which is

structurally converted into a single 3×3 convolution during inference to ensure high computational efficiency. Furthermore, to accommodate diverse deployment environments, the core innovation of LGC-CSP lies in its flexible, scalable multi-branch design. The number of parallel branches is governed by a parameter n, enabling precise adjustments to model complexity based on available computational resources. A smaller n minimizes parameter overhead for resource-constrained devices, whereas a larger n maximizes performance for high-accuracy scenarios. Additionally, an integrated Dropout2d layer applies channel-level regularization, bolstering the model's robustness against illumination variations and occlusions.

2.3. Cross-Scale Decoupled (CSD) Head

The standard YOLOv10n detection head often exhibits insufficient localization accuracy and low bounding box precision when detecting joint gaps in complex underground environments. To address these limitations, we propose the Cross-Scale Decoupled (CSD) head based on a decoupled design, offering superior structural efficiency and performance. As shown in Figure 3, the CSD head integrates convolutional layers, shared convolutions (ShareConv), bounding box regression and classification modules, a Distribution Focal Loss (DFL) processing module, and a decoding module. By employing shared convolutions and multi-level feature processing, the CSD head minimizes computational redundancy while enhancing both accuracy and efficiency. Additionally, the combination of dynamic anchor generation and DFL processing optimizes target localization and bounding box prediction, particularly for rigid guide rail joint gaps. To further streamline deployment, a heterogeneous training-inference architecture is utilized, retaining only the one-to-one branch during inference to significantly reduce computational complexity.

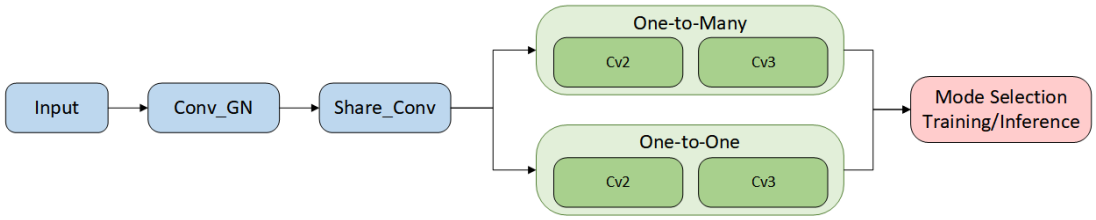


Figure 3. Structure of CSD Head

Initially, input features are processed through multiple convolutional layers equipped with normalization to ensure training stability and enhance detection accuracy. The extracted features are then fed into the shared convolution module, which comprises depthwise and pointwise convolutions. This configuration effectively captures detailed feature representations while alleviating the computational burden. Subsequently, the feature maps are split into a one-to-many branch and a one-to-one branch. To optimize joint gap detection, the classification and localization tasks are strictly decoupled, significantly improving precision. During bounding box regression, the Distribution Focal Loss (DFL) module transforms predictions into coordinate offsets. Finally, an integrated decoding and post-processing mechanism translates these offsets into actual bounding box coordinates.

Furthermore, a learnable scaling factor is applied to the bounding box predictions at each feature level to adjust the $Scale_i$, as formulated in Eq. (3):

$$outport_i = Concat\left(Scale_i(cv2(x_i)), cv3(x_i)\right) \quad (3)$$

Where i denotes the feature level, x_i is the input feature map, cv_2 and cv_3 represent the bounding box regression and classification predictions, respectively, and $Scale_i$ is the level-specific learnable scaling factor.

2.4. BiFormer Attention Mechanism

The complex backgrounds of underground mine shafts frequently cause false positives and missed detections of joint gaps. To effectively isolate critical gap features amidst such interference, we integrate the BiFormer attention module into the YOLOv10n backbone. As illustrated in Figure 4, BiFormer employs a dynamic, query-aware sparse attention mechanism based on bi-level routing. This architecture enables highly flexible computation allocation and content awareness, substantially boosting detection performance without incurring significant computational overhead.

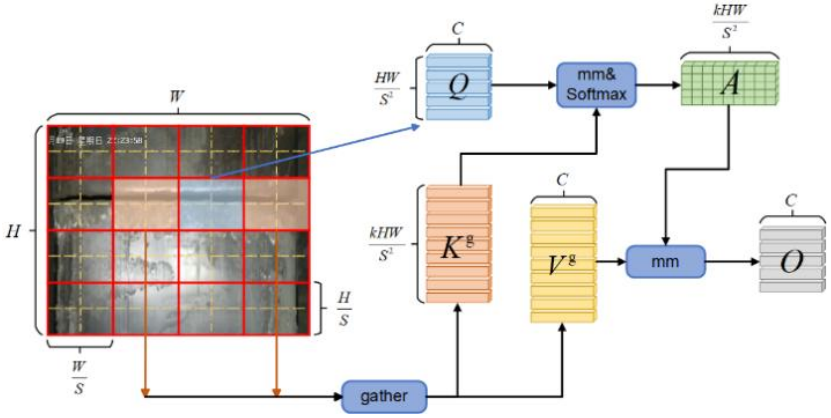


Figure 4. Structure of Biformer

The module comprises depthwise separable convolutions, layer normalization, bi-level routing attention, and a Multi-Layer Perceptron (MLP), all integrated via residual connections. Its operation proceeds in three primary stages.

Initially, region partitioning is applied. Given a 2D input feature map $X \in R^{C \times H \times W}$, it is divided into $S \times S$ non-overlapping regions, with each region containing $\frac{HW}{S^2}$ feature

vectors. After reshaping these partitioned features into X^r, the query (Q), key (K), and value (V) vectors are computed via linear projections, as formulated in Eq. (4):

$$Q = X^r W^q, K = X^r W^k, V = X^r W^v \quad (4)$$

Where $W^q, W^k, W^v \in R^{C\times C}$ denote the projection weight matrices for Q, K, and V, respectively.

Next, to capture inter-region correlations, the adjacency matrix $Q^r, K^r \in R^{S^2\times C}$ is derived by multiplying the region-averaged query and key vectors ($A^r \in R^{S^2\times S^2}$), as formulated in Eq. (5):

$$A^r = Q^r (K^r)^T \quad (5)$$

At a coarse level, low-relevance regions are filtered out using a top-k operation, retaining only the k most relevant regions for each query to generate the routing index matrix $I^r \in N^{S^2\times k}$. Based on I^r, the key and value vectors of these selected regions are gathered to form K^g and V^g, respectively. The final bi-level routing attention output O is computed via Eqs. (6)–(8):

$$K^g = gather(K, I^r) \quad (6)$$

$$V^g = gather(V, I^r) \quad (7)$$

$$O = Attention(Q, K^g, V^g) + LCE(V) \quad (8)$$

Where LCE employs a 5×5 depthwise separable convolution for local context enhancement.

By optimizing computational efficiency, the integration of BiFormer enables YOLOv10n-LBC to precisely detect joint gaps in harsh environments characterized by low illumination, heavy dust, and complex backgrounds. Specifically, its bi-level routing and sparse computation paradigms substantially enhance the network's feature representation capacity, excelling particularly in the extraction of fine edge details. Ultimately, this attention mechanism guarantees both the high detection accuracy and the real-time processing capabilities critical for underground mine inspection.

3. Experiments and Results Analysis

3.1. Dataset

The dataset was derived from actual mine inspection videos captured by an explosion-proof fiber-optic camera (Hikvision KBA127-XYF-LZ). Because the cage traverses the entire shaft during operation, the camera was mounted atop the cage to record comprehensive guide rail footage. The finalized dataset comprises 6,100 images: 4,500 real joint gap images extracted from the videos, alongside 1,600 augmented images simulating heavy dust (800) and low-light (800) conditions. Ground truth bounding boxes were manually annotated using LabelImg, and the entire dataset was randomly partitioned into training, validation, and testing sets at an 8:1:1 ratio.

3.2. Experimental Setup

All experiments were conducted on a Windows 11 (64-bit) workstation equipped with an NVIDIA GeForce RTX 4060 Ti GPU (16 GB VRAM). The proposed model was implemented using the PyTorch 2.0.1 deep learning framework, accelerated by CUDA 11.8 within a Python 3.9.21 environment.

3.3. Evaluation Metrics

To comprehensively evaluate the proposed model, we employ Precision (P), Recall (R), mean Average Precision (mAP), model size (Weight), and parameter count (Parameters) as baseline performance metrics. The quantitative calculations for these metrics are formulated in Eqs. (9)–(11):

$$P = \frac{TP}{TP+FP} \quad (9)$$

$$R = \frac{TP}{TP+FN} \quad (10)$$

$$AP = \int_0^1 P(R)dR \quad (11)$$

Where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively; P represents precision, R represents recall, and AP stands for Average Precision.

3.4. Ablation Study

To verify the individual effectiveness of the proposed modules and their combined contribution to the baseline YOLOv10n, a series of ablation experiments were conducted. To ensure fairness, all models were trained using identical hyper-parameters on the rigid guide rail joint gap dataset. Table 1 summarizes the objective performance metrics of the baseline YOLOv10n, three single-module improvements, and four multi-module combinations, culminating in the final YOLOv10n-LBC.

The experimental results demonstrate that the LGC-CSP module effectively compresses the model, reducing the weight by 20.0% compared to the baseline. However, its independent application incurs a 6.2% drop in mAP@0.5, necessitating collaborative optimization with other modules to compensate for this accuracy loss. Conversely, the CSD head delivers a significant performance boost, elevating mAP@0.5 by 9.7% to 98.8% while simultaneously reducing the model weight by 27.3%. When LGC-CSP and CSD are combined—yielding YOLOv10n-(LGC+CSD)—the model maintains an excellent mAP@0.5 of 98.7% and a weight of 4.0 MB, achieving a solid balance between accuracy and lightweight design. Ultimately, the YOLOv10n-LBC architecture, which integrates all three modules, slashes the parameter count to 1.57 M with a compact 3.4 MB weight file. Furthermore, it achieves an mAP@0.5:0.95 of 72.2% and an mAP@0.5 of 98.8%, demonstrating optimal model compression and robust detection performance.

Table 1. Comparison of objective indexes in ablation experiments

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5-0.95/%	Parameters/M	Weight/MB
YOLOv10n	89.1	88.8	89.1	14.7	2.27	5.5
YOLOv10n-LGC	84.3	83.8	82.9	12.3	1.68	4.4
YOLOv10n-Bi	87.5	87.4	85.8	14.1	2.53	6.0
YOLOv10n-CSD	93.8	96.3	98.8	69.9	1.83	4.0
YOLOv10n-(LGC+Bi)	86.4	85.6	84.1	12.8	2.42	6.1
YOLOv10n-(CSD+Bi)	96.7	96.6	99.0	71.2	2.07	4.2
YOLOv10n-(LGC+CSD)	98.2	94.6	98.7	70.8	1.83	4.0
YOLOv10n-LBC	98.6	97.1	98.8	72.2	1.57	3.4

In summary, the comprehensive ablation results conclusively validate the effectiveness of the proposed enhancements to the YOLOv10n model. The complementary strengths of these modules synergize to ensure a lightweight footprint without compromising detection accuracy, thereby providing a superior solution for practical, real-world deployment.

3.5. Comparison with State-of-the-Art Models

To evaluate the performance of YOLOv10n-LBC for underground guide rail joint gap detection, we benchmarked it against current mainstream object detection models—YOLOv5s, YOLOv8n, YOLOv9t, YOLOv10n, YOLOv11n, and YOLOv12n—under strictly identical experimental settings and datasets. As summarized in Table 2, the proposed YOLOv10n-LBC exhibits distinct advantages across Precision (P), Recall (R), mAP@0.5, mAP@0.5:0.95, parameter count, and model size. By simultaneously enhancing detection accuracy and achieving a highly compact architecture, these results compellingly substantiate the efficacy of our method. However, these structural modifications incur a minor inference speed trade-off, reducing the frames per second (FPS) to 291.5.

Table 2. Performance comparison of different detection algorithms

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5-0.95/%	Parameters/M	Weight/MB	FPS/(frames·s^{-1})
YOLOv5s	91.7	91.7	92.4	16.7	2.50	5.0	442.5
YOLOv8n	94.9	94.9	94.8	17.0	3.01	6.0	457.5
YOLOv9t	82.0	88.5	86.6	14.9	1.97	4.4	188.2
YOLOv10n	89.1	88.8	89.1	14.7	2.27	5.5	365.2
YOLOv11n	91.4	91.5	91.7	15.8	2.58	5.2	367.9
YOLOv12n	90.3	90.3	90.9	16.0	2.56	5.2	367.2
YOLOv10n-LBC	98.6	97.1	98.8	72.2	1.57	3.4	291.5

4. Joint Gap Width Measurement and Edge Extraction

We employ a pixel-to-physical ratio methodology to compute the absolute width of the rigid guide rail joint gaps. Initially, the proposed YOLOv10n-LBC algorithm performs high-precision detection to generate the spatial bounding box for each target. By correlating the pixel dimensions of these bounding boxes with the known, calibrated physical width of the guide rails, the actual gap dimensions are accurately quantified.

Furthermore, with the trained model achieving a remarkable detection accuracy of 98.5% on the validation set, these bounding boxes serve as a highly reliable geometric foundation for width estimation. To validate the practical efficacy of this approach, we evaluated the system using continuous video footage captured during the uniform inspection descent of the cage. The corresponding real-world detection results, highlighting eight distinct joint gaps encountered throughout the inspection run, are presented in Figure 5.

Figure 5. Detection results of guide rail inspection video

5. Conclusion

To address the challenges of small targets, blurred edges, and intense environmental interference in underground mine shaft inspections—while balancing detection accuracy and model lightweighting for edge devices—we propose YOLOv10n-LBC, an enhanced detection algorithm based on YOLOv10n. By integrating the LGC-CSP module for dynamic complexity adjustment, the CSD head for multi-scale feature fusion and precise localization, and the BiFormer attention mechanism for critical feature extraction, the proposed algorithm achieves an optimal trade-off between high-precision detection and efficient deployment in complex underground environments.

Experimental results demonstrate that YOLOv10n-LBC significantly outperforms the baseline YOLOv10n. On the rigid guide rail joint gap dataset, the model achieves an mAP@0.5:0.95 of 72.2% with only 1.57 M parameters and a compact 3.4 MB weight file. Compared with state-of-the-art algorithms, our method maintains a lightweight footprint while delivering superior detection accuracy, making it highly suitable for practical embedded vision systems in mines. Furthermore, a visual measurement methodology based on bounding box results was implemented, achieving a measurement error fluctuation of less than 1 mm in real-world testing. This, coupled with the proposed edge extraction algorithm, validates the stability and reliability of the system, providing robust technical support for detecting guide rail misalignments in practical engineering scenarios.

References

[1] WANG Guofa, PANG Yihui, REN Huaiwei, et al. System engineering and key technologies research and practice of smart mine [J]. Journal of China Coal Society, 2024, 49(1):181−202.

[2] DU Fei, MA Tianbing, HU Weikang, et al. Fault diagnosis of rigid guide based on wavelet transform and improved convolutional neural network [J]. Industry and Mine Automation, 2022, 48(09):42-48+62.

[3] MA Tianbing, WANG Xinquan, WANG Xiaodong. Fault diagnosis method for rigid tank channel based on EMD-PNN network [J]. Journal of Electronic Measurement and Instrumentation, 2019, 33(03):58-64.

[4] ZHANG Yi, KONG Xianggao, GUO Mengda. Research and application of optical fiber sense technology in auxiliary shaft hoisting system [J]. Coal Engineering, 2025, 57(09):175-182.

[5] YANG Wanlin. Research on Wellbore Safety Detection Based on Inertial and 3D Laser Technology [D]. Tai'an:Shandong University of Science and Technology, 2020,12-25.

[6] NIU Weifeng, ZONG Liangliang, WU Feng, et al. Two-dimensional laser scanner based coal mine shaft guide deformation detection device [J]. Industry and Mine Automation, 2021, 47(03):101-104+111.

[7] Pang Meng. Research on Wellbore Inspection System of Deep Vertical Shaft Based on Machine Vision [D]. Huai'nan: Anhui University of Science and Technology, 2023, 35-54.

[8] ZHAO Baiting, WU Jundong, JIA Xiaofang. Lightweight guide defect detection algorithm based on feature enhancement [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(06):159-168.

[9] Sun Zhicheng. Rigid Tank Channel Defect Detection Based on Improved YOLOv5 and Embedded Platform[D]. Huai'nan: Anhui University of Science and Technology, 2024, 27-48.

[10] Fang Jiaxin. Research on Fault Detection of Guiding Device in Vertical Shaft Hoisting System Based on Machine Vision [D]. Huai'nan: Anhui University of Science and Technology, 2024, 41-58.

[11] WANG Manli, YANG Shuang, ZHANG Changsen. An improved YOLOv8n based dislocation detection algorithm for shaft rigid tank channel joint [J]. Coal Science and Technology, 2024, 52(S2):236-248.

[12] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 779-788.

[13] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 7263-7271.

[14] Redmon J, Farhadi A. YOLOv3: An incremental improvement [EB/OL]. arXiv preprint, arXiv:1804.02767, 2018.

[15] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection [EB/OL]. arXiv preprint, arXiv:2004.10934, 2020.

[16] WANG C, YEH I, LIAO H. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information [EB/OL]. arXiv preprint, arXiv:2402.13616, 2024.

[17] Wang A, Chen H, Liu L, et al. YOLOv10: Real-time End-to-End Object Detection [EB/OL]. arXiv preprint arXiv: 2405.14458, 2024.

[18] Zhu L, Wang X, Ke Z, et al. BiFormer: Vision Transformer with Bi-Level Routing Attention [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 10323-10333.

[19] Wang Chien-Yao, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN [C]// Proceedings of the IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 390-391.

[20] Han Kai, Wang Yunhe, Tian Qi, et al. Ghostnet: More features from cheap operations [C]// Proceedings of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 1580-1589.

[21] Tang Yehui, Han Kai, Guo Jianyuan, et al. GhostNetv2: Enhance cheap operation with long-range attention [C]//Proc of the 36th Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2022: 9969-9982.

[22] Soudy M, Afify Y, Badr N. RepConv: A novel architecture for image scene classification on Intel scenes dataset [J]. International Journal of Intelligent Computing and Information Sciences, 22(2):63-73, 2022.