

Coal Mine Drill Pipe Detection Algorithm Integrating Large-Kernel Separation and Coordinate Locking Mechanisms

Canjun Xiao

Chengdu Technological University

Zaisheng Jiang

zaisheng_j@163.com

Chengdu Technological University

Chenyu Wang

Chengdu Technological University

Junxiang Yang

Chengdu Technological University

Siliang Zheng

Lubanshan North Mine, Sichuan Chuannan Coal Industry Co., Ltd.

Youping Li

Jiayang Coal Mine, Sichuan Jiayang Group Co., Ltd.

Dongdong Chen

China University of Mining and Technology - Beijing

Shengrong Xie

China University of Mining and Technology - Beijing

Article

Keywords:

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11177859/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Coal Mine Drill Pipe Detection Algorithm Integrating Large-Kernel Separation and Coordinate Locking Mechanisms

Canjun Xiao^{1,2}, Zaisheng Jiang^{1,2*}, Chenyu Wang^{1,3}, Junxiang Yang¹, Siliang Zheng⁴,
Youping Li⁵, Dongdong Chen⁶, Shengrong Xie⁶

¹(Mine Intelligence) Sichuan Digital Transformation Promotion Center, Chengdu Technological University, Chengdu, Sichuan, 611730, China.

²School of Computer Engineering, Chengdu Technological University, Chengdu, Sichuan, 611730, China.

³School of Computer Science, Chengdu University of Information Technology, Chengdu, Sichuan, 610225, China.

⁴Lubanshan North Mine, Sichuan Chuannan Coal Industry Co., Ltd., Yibin, Sichuan, 645250, China.

⁵Jiayang Coal Mine, Sichuan Jiayang Group Co., Ltd., Leshan, Sichuan, 614401, China.

⁶School of Energy and Mining Engineering, China University of Mining and Technology (Beijing), Beijing, 100083, China.

*Corresponding author(s). E-mail(s): zaisheng_j@163.com;
Contributing authors: canjunxiao@cdu.edu.cn;

Abstract

Underground coal mine operating scenes feature complex working conditions such as significant dynamic variations, frequent target occlusions, intense exposure interference, localized low illumination, and high concentrations of pervasive dust. Conventional object detection models tend to suffer from severe overfitting, leading to a substantial degradation in generalization performance for detecting slender drill pipes. To overcome the deficiencies of conventional models, this paper proposes a drill pipe detection algorithm based on the lightweight YOLO11n network, integrating large-kernel separation and coordinate locking mechanisms. At the end of the backbone network, Large Separable Kernel Attention (LSKA) and Coordinate Attention (CA) are integrated to construct the LSCA feature enhancement module; this module uses large-kernel convolutions to connect fractured features and alleviate frequent occlusions, and employs orthogonal pooling to precisely lock onto the physical edges of drill pipes against intense light, darkness, and dust interference. At the detection head, an interpolation-based flexible regression loss (InterpIoU) is introduced to reconstruct the bounding box localization mechanism; by establishing a transitional interpolation buffer, it suppresses gradient oscillations caused by extreme aspect ratios and effectively reduces the short-edge divergence sensitivity in slender drill pipe regression. At the validation stage, a validation set is constructed based on an Out-Of-Distribution (OOD) strategy with physical isolation in time and space. Experimental results demonstrate that: (1) The improved model exhibits superior generalization at unknown operating sites, where mAP@50 increases by 2.0% over the baseline (reaching 88.7%) and mAP@50-95 increases by 2.2% (reaching 43.7%); (2) The improved model demonstrates stronger anti-interference capability under complex conditions, with the miss rate reduced by 1.4% over the baseline (to

18.3%) and the recall increased by 1.4% (to 81.7%); (3) The improved model maintains lightweight characteristics, with parameters reduced by 0.029 M relative to the baseline (reaching 2.561 M) while maintaining high real-time inference speed (reaching 1303.8 fps); (4) When coupled with the ByteTrack tracker at the back end, the counting accuracy on five challenging video cases is improved by 9.1 percentage points over the baseline.

Keywords: Complex working conditions, Large separable kernel attention, Coordinate locking, Flexible regression, Out-of-distribution, Drill pipe detection

Underground coal mine drilling is a fundamental operation for gas control, disaster prevention, and the intelligent development of coal mining equipment [1, 2]. With the advancement of intelligent and reduced-staff coal mines, accurately obtaining drilling depth to evaluate gas drainage compliance rates and borehole pressure relief and energy release effects fundamentally relies on continuous and accurate drill pipe counting [3, 4, 5, 6]. Traditional contact-based sensing is prone to failure under harsh working conditions. Conversely, machine vision and drilling robots, leveraging the advantages of non-contact perception, provide reliable means for drilling operation monitoring and coal flow safety assurance. Accurately capturing drill pipe targets at the front end and linking with back-end logic to achieve automated counting have become key to the intelligence of coal drawing and drilling operations.

In recent years, extensive research has been conducted on intelligent visual perception and counting methods for drill pipes. Hu et al. [7] and Wang and Shi [8] combined improved YOLO models with the DeepSORT algorithm to count drill pipes by tracking the relative distance between the chuck and gripper or tracking drill pipe positions. Zhang et al. [9] constructed the Drill-YOLOv8 model and utilized a two-stage decision area to count the movement cycles of the rig carriage in real time. Zhang et al. [10] combined oriented object detection with a saliency peak inference mechanism to improve detection under variable poses of slender targets. Li et al. [11] proposed a dual-encoding edge enhancement network (DEED-Net) to achieve unified recognition of multiple targets in drilling scenarios. Jiang [12] integrated object detection with digital twin technology to realize visual servoing and intelligent monitoring of drill pipe joints. Fang et al. [13] and Dong et al. [14] explored counting schemes using conventional visual tracking algorithms such as ECO-HC and Camshift. Hao et al. [15] extracted human skeleton keypoints via YOLO-Pose and integrated a spatio-temporal graph convolutional network (ST-GCN) to classify drill rod loading and unloading actions. Du et al. [16] similarly identified

worker action sequences based on an improved ST-GCN. Zhang and Jiang [17] and Zong [18] employed improved MobileNetV2 and AlexNet classification networks, respectively, to count indirectly by determining rig operating states or worker rod-unloading postures. In summary, the established mainstream framework of “front-end detection + back-end temporal or action discrimination” has improved counting accuracy, laying a foundation for the intelligent and unmanned development of mining equipment. However, while these methods achieve high test accuracy on routinely distributed datasets, their generalization performance deteriorates under dynamic underground conditions characterized by temporal variations, frequent occlusions, sharp illumination changes, and pervasive dust. The fragile boundary features of drill pipes are prone to physical truncation and photometric distortion. Consequently, traditional front-end perception models suffer rapid generalization degradation when facing unknown dynamic working faces, making it difficult to maintain stable and continuous target tracking.

To address these issues, this paper constructs an out-of-distribution (OOD) validation mechanism based on the principle of spatiotemporal physical isolation to reflect the true generalization capability of the model, and proposes a lightweight drill pipe detection algorithm. Using YOLO11n as the baseline, the algorithm introduces a large-kernel separation and coordinate locking mechanism (LSCA) at the end of the backbone network to decouple background noise and enhance the topological connectivity of slender features. Simultaneously, an interpolation-based flexible regression loss (InterIoU) is introduced to suppress gradient oscillations, improving visual perception stability under extreme conditions and providing robust front-end support for automated drilling counting.

1 Method

1.1 Overall Model Architecture

Although multi-modal fusion sensing technologies combining X-rays or LiDAR point clouds offer higher theoretical accuracy limits, lightweight monocular vision-only detection remains the practical primary choice for large-scale engineering deployment due to the computational bottlenecks and deployment costs of underground edge computing platforms [19], and is supported by the general evolution of object detection [20, 21]. To address the generalization degradation caused by complex underground conditions, this paper adopts YOLO11n as the baseline—a choice in line with recent lightweight YOLO-based enhancements for slender structural targets such as bridge and pipeline cracks [22, 23]—and constructs a scene-adaptively restructured drill pipe detection network (overall architecture shown in Fig. 1).

The network restructuring focuses on two core pathways: feature extraction and bounding box regression. At the end of the backbone network, the original C2PSA module is replaced by the C2PSA_LSCA module (Fig. 2). This design enables the network to establish long-range semantic connectivity across frequent occlusions before features flow into the neck, precisely locking onto the physical edges of drill pipes to resist intense light and dust interference, thereby enhancing the model’s feature decoupling and perceptual generalization capabilities in complex environments.

At the detection head, to address the convergence and divergence difficulties of slender rigid bodies under complex conditions, the InterIoU flexible regression loss is introduced. By establishing a transitional interpolation buffer, this mechanism suppresses gradient oscillations caused by extreme aspect ratios, improving the compactness of bounding box wrapping and localization robustness under severe interference.

1.2 Feature Enhancement Mechanism LSCA Integrating Large-Kernel Separation and Coordinate Locking

At the end of the original YOLO11 backbone network, the C2PSA module is used for spatial feature extraction [24]. However, underground drill pipes are frequently occluded by pipelines and subjected to intense light and dust interference. The native module relies on local convolutions, whose limited receptive field

often causes topological fractures at occlusion boundaries; furthermore, it lacks explicit physical directional modeling, making features easily confused with complex backgrounds.

To address this, a large-kernel separable coordinate attention mechanism (LSCA) is constructed by integrating LSKA and CA into dual parallel branches and is embedded into the PSABlock. LSCA decouples cross-occlusion long-range contextual perception and interference-resistant physical edge locking into the two branches, completing feature fusion via a single multiplication (internal architecture shown in Fig. 3). The individual attention operators are both established designs; the contribution of this work lies in their parallel multiplicative fusion and the embedding position, which are tailored to the occlusion and glare characteristics of underground drill pipe scenes, and their effectiveness is verified against the single-branch counterparts in the ablation study (Section 3.3). For an input feature map, the calculation procedure is as follows:

Branch A: Large-Kernel Spatial Attention (extracting long-range contextual modulation features) To bridge physical occlusion boundaries, the input feature X enters a large-kernel separable convolution module based on Large Separable Kernel Attention (LSKA) [22, 25] along Branch A. Large-kernel separable attention has recently proven effective for slender elongated targets such as bridge and pavement cracks [22, 26], sharing the extreme-aspect-ratio perception challenge of drill pipes. This branch decomposes a dense large kernel into depthwise convolutions in the horizontal $1 \times k$ and vertical $k \times 1$ directions, coupled with depthwise dilated convolutions with dilation rate $d = 2$ (or $d = 5$), substantially expanding the receptive field with minimal parameters. Subsequently, through a 1×1 pointwise convolution and Sigmoid activation, a spatial modulation map M_{spatial} bridging physical occlusions is generated. The complete forward process is formulated as:

$$F_{dw} = \text{DWConv}_{k \times 1}(\text{DWConv}_{1 \times k}(X)) \quad (1)$$

$$F_{dwd} = \text{DW-DConv}_{d \times 1}(\text{DW-DConv}_{1 \times d}(F_{dw})) \quad (2)$$

$$M_{\text{spatial}} = \sigma(\text{Conv}_{1 \times 1}(F_{dwd})) \quad (3)$$

where σ denotes the Sigmoid activation function. The receptive field size directly determines the ability to correlate fractured segments. Based on the kernel size optimization results in the ablation study (Table 3),

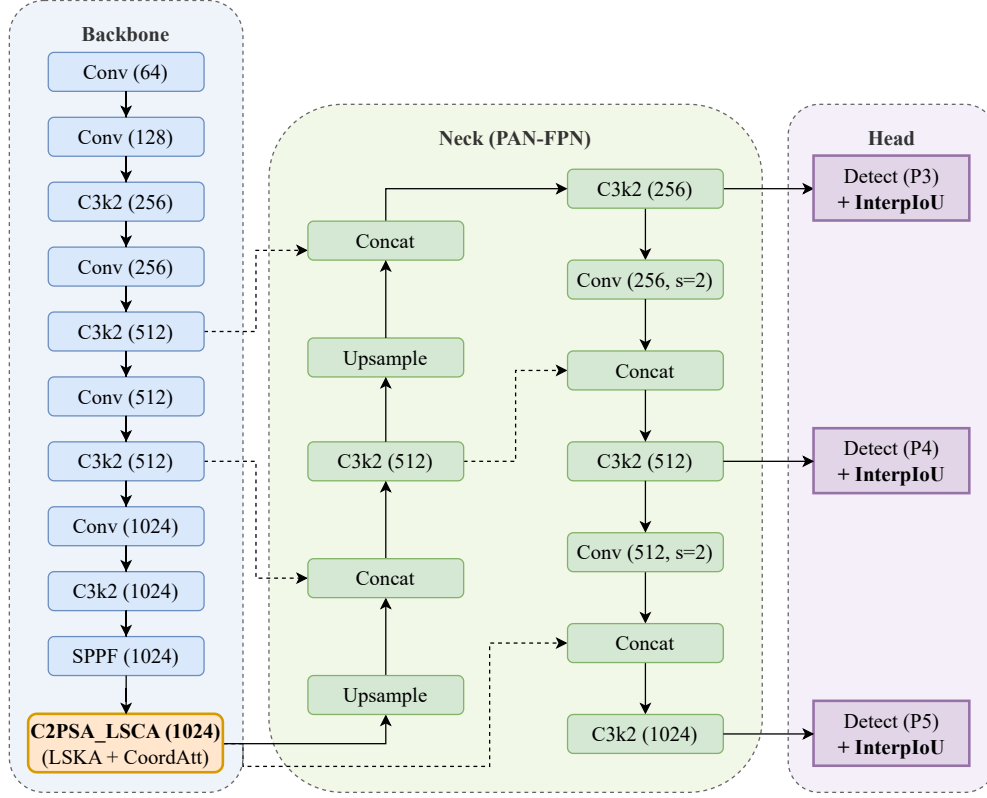


Fig. 1: Overall architecture of the improved detection network

C2PSA_LSCA Macro-architecture and PSABlock_LSCA Internal Topology

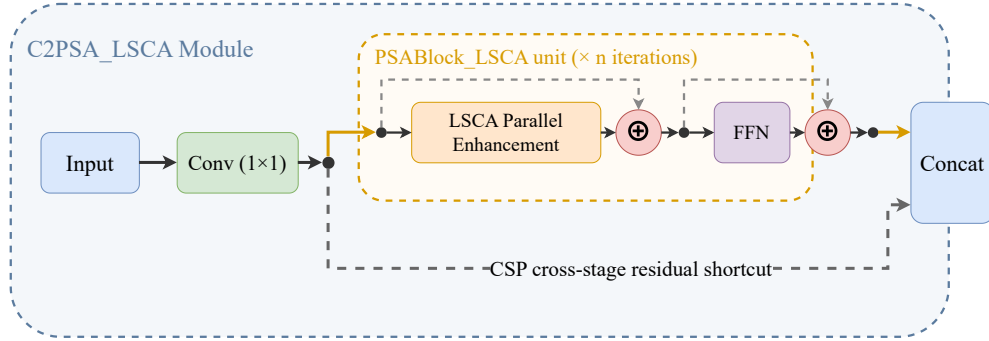


Fig. 2: Macro-architecture of C2PSA.LSCA and internal topology of PSABlock.LSCA

$k = 11$ (corresponding to $d = 5$) is selected as the default configuration, where the effective receptive

field covers typical occlusion widths without introducing excessive background noise from an oversized kernel.

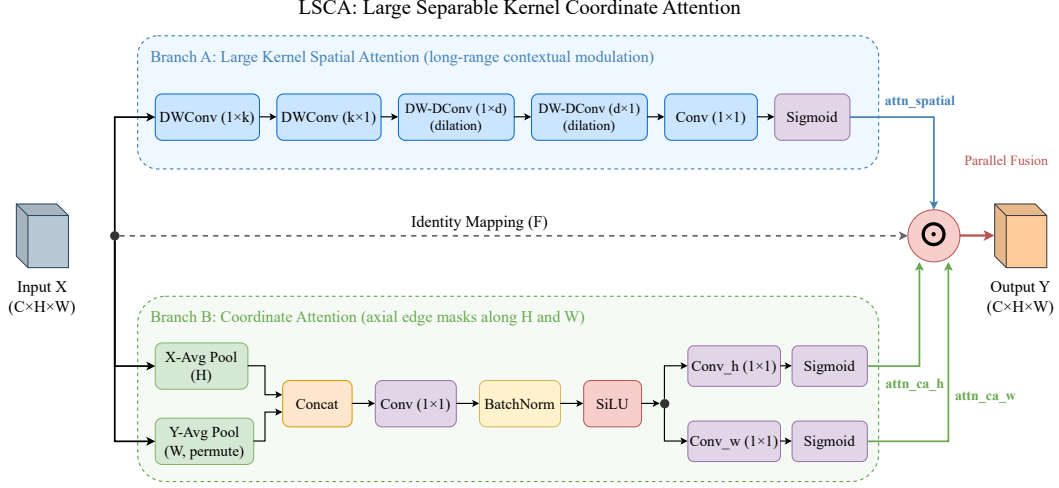


Fig. 3: Structure of the LSCA large-kernel separable coordinate attention mechanism

Branch B: Coordinate Attention (precisely locking axial edge masks along H and W) The input feature X simultaneously enters the Coordinate Attention (CA) [27] branch, where adaptive average pooling is applied along the horizontal and vertical directions, extracting feature vectors $z_h \in \mathbb{R}^{C \times H \times 1}$ and $z_w \in \mathbb{R}^{C \times 1 \times W}$ containing directional priors:

$$z_h = \frac{1}{W} \sum_{i=1}^W X(h, i) \quad (4)$$

$$z_w = \frac{1}{H} \sum_{j=1}^H X(j, w) \quad (5)$$

The extracted height and width features are concatenated, and cross-channel intermediate features Y_{ca} are derived through a shared convolution, BatchNorm, and SiLU activation:

$$Y_{ca} = \text{SiLU}(\text{BatchNorm}(\text{Conv}_{ca}(\text{Concat}(z_h, z_w)))) \quad (6)$$

Subsequently, Y_{ca} is split back into height feature X_h and width feature X_w , which pass through independent 1×1 convolutions and Sigmoid activations to generate height mask M_h and width mask M_w that lock the drill pipe topology:

$$M_h = \sigma(\text{Conv}_h(X_h)) \quad (7)$$

$$M_w = \sigma(\text{Conv}_w(X_w)) \quad (8)$$

Parallel Multiplicative Fusion To avoid signal attenuation caused by sequential structures, LSCA abandons multi-stage cascading and directly utilizes identity mapping, performing a single element-wise multiplication among the original feature X , spatial modulation map M_{spatial} , height mask M_h , and width mask M_w to yield the enhanced feature Y_{out} :

$$Y_{\text{out}} = X \odot M_{\text{spatial}} \odot M_h \odot M_w \quad (9)$$

While preserving initial high-frequency edge details, this parallel fusion strategy synergistically injects dual feature priors of cross-occlusion semantic connectivity and physical edge locking, enhancing feature decoupling and perceptual generalization under extreme conditions.

1.3 Interpolation-Based Flexible Regression Loss InterIoU

Bounding box regression at the detection head is sensitive to extreme aspect ratios. The aspect ratio penalty in the original CIoU [28] tends to distort the short edge under absolute errors, encompassing background noise and causing oscillations. Subsequent works such as EIoU also proposed decoupling penalty terms to address this sensitivity [29], and geometry-aware IoU variants tailored to elongated targets, such as WIoU-based designs for slender bridge cracks [30], further confirm the benefit of explicit geometric constraints.

For drill pipe targets, pixel counts along the width dimension are typically small; thus, minute deviations on the short edge can be substantially amplified in the regression loss, leading to oscillations where the predicted box fluctuates between wide and narrow during training. This not only impairs the tight wrapping of the bounding box around the drill pipe body, but also undermines the stability of subsequent event determinations based on box position and length changes.

To alleviate this problem, the flexible regression loss InterIoU [31, 32] is introduced. Let the predicted box be B_p and the ground-truth box be B_g . A transitional interpolated box B_i is constructed:

$$\begin{aligned} x_i &= \lambda x_p + (1 - \lambda)x_g, & y_i &= \lambda y_p + (1 - \lambda)y_g & (10) \\ w_i &= \lambda w_p + (1 - \lambda)w_g, & h_i &= \lambda h_p + (1 - \lambda)h_g & (11) \end{aligned}$$

where λ is set to 0.5, as determined by the sensitivity analysis presented in Section 3.3. The interpolated Intersection over Union is computed as:

$$\text{IoU}_{\text{interp}} = \frac{|B_i \cap B_g|}{|B_i \cup B_g|} \quad (12)$$

The loss function is formulated as:

$$\mathcal{L}_{\text{InterIoU}} = 1 - \text{IoU}_{\text{interp}} + \frac{\rho^2(B_p, B_g)}{c^2} + \alpha v \quad (13)$$

The original InterIoU directly replaced hand-crafted geometric penalties with the interpolated IoU [32]. However, considering that center deviation and shape mismatch under extreme aspect ratios still require explicit constraints for drill pipe targets, the center distance penalty $\rho^2(B_p, B_g)/c^2$ and aspect ratio consistency term αv from CIOU [28] are retained within the interpolation framework, forming a flexible regression scheme tailored to slender rigid bodies. In other words, the proposed loss replaces only the overlap term with its interpolated counterpart, while the center-distance and aspect-ratio penalties are inherited from CIOU; this hybrid retains the smooth short-edge gradients provided by the interpolation without abandoning explicit geometric regularization. Here, $\rho^2(B_p, B_g)$ denotes the Euclidean distance between center points, c is the diagonal length of the smallest enclosing box covering both predicted and ground-truth boxes, α is a positive weight parameter, and v measures aspect ratio consistency. The interpolated

box provides a smooth transition that supplies continuous gradients to the network, reducing the probability of short-edge dilation.

From the perspective of training stability, this design is suited for bounding box regression of slender rigid objects like drill pipes. The long-edge information of drill pipes is generally stable, whereas the short edge is more vulnerable to minor annotation discrepancies, blurry occlusion boundaries, and image compression noise. If the loss function is overly sensitive to short-edge variations, the model tends to expand the predicted box to secure a higher overlap ratio. InterIoU attenuates this aggressive compensation tendency through interpolated transitions, guiding the predicted box to form a compact boundary along the drill pipe body and providing support for the interpretability and operational stability of downstream detection results.

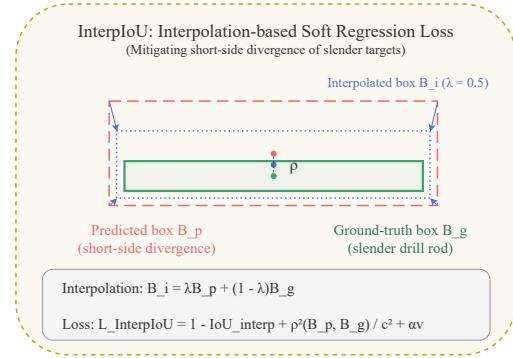


Fig. 4: Geometric diagram of interpolation-based flexible regression loss for InterIoU

2 Dataset Construction and Preprocessing

2.1 Data Acquisition and Out-of-Distribution Partitioning

The dataset is designed for a single-class drill pipe detection task. Raw videos were collected across nine distinct drilling operation sites. After filtering out similar frames, a coal mine drill pipe visual detection dataset covering diverse working conditions was constructed. Rather than using conventional random partitioning of mixed data, an out-of-distribution (OOD) grouping strategy based on spatial-site and

Table 1: Statistics of the drill pipe detection dataset

Split	Images	Instances
Training (before augmentation)	4,820	4,821
Training (after augmentation)	14,458	15,791
Test (OOD)	1,182	1,183

acquisition-time isolation was adopted for training and test splits (sample images are shown in Fig. 5). Underground roadway environments undergo substantial variations over time, and background configurations and lighting conditions differ markedly across mining sites. Structuring the dataset with strict OOD splits by time and location objectively evaluates the model’s true feature extraction capability on unseen physical working faces, aligning with practical deployment scenarios where the model is directly deployed at new working faces.

The statistics of the dataset are summarized in Table 1. The nine drilling operation sites comprise four sites at the Lubanshan North Mine—site 1438 (one working face), site 1448 (two working faces), site 1458 (two working faces), and site 1468 (three working faces)—together with one additional site at another mine, totaling nine collection points. After filtering out similar frames, 6,002 original images containing 6,004 drill pipe instances were retained. According to the OOD grouping strategy, the training and test splits were assigned at the site and working-face level with acquisition-time isolation, yielding a test set of 1,182 images with 1,183 instances and a training pool of 4,820 images with 4,821 instances; the latter was expanded to 14,458 images (15,791 instances) by the non-destructive augmentation described in Section 2.2. All annotations were manually delineated and verified frame-by-frame; boxes with ambiguous boundary fitting were re-annotated to ensure consistent wrapping along the target major axis.

In terms of data composition, the samples encompass drill pipe targets across various drilling angles, feeding poses, and occlusion levels, incorporating representative challenging conditions such as specular reflections, dust interference, close-range human operations, and dense equipment occlusions. For images that may lead to the identification of personal information, this manuscript has obtained informed consent from the relevant personnel in the images, agreeing to publish the personal images involved in Figures 5, 6,

9, and 10 in open access publications. This distribution enables the model to capture the appearance variation range of slender targets in real underground environments. Because drill pipes typically occupy a minimal portion of the field of view, data annotation prioritized consistency in wrapping along the major axis of the target to reduce interference caused by excessive bounding box width discrepancies during regression learning.

2.2 Non-Destructive Data Augmentation

Given the extremely narrow rigid-body characteristics of drill pipes, a conservative non-destructive augmentation strategy combining controlled geometry and global photometric transformations is adopted:

(1) Controlled geometric transformations: Discrete angle sets $\{-30^\circ, -10^\circ, 10^\circ, 30^\circ, 50^\circ, 70^\circ\}$ and scale factor sets $s \in \{0.8, 0.9, 1.1, 1.2\}$ are defined. The pixel coordinates (x, y) are affinely mapped as:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s \cos \theta & -s \sin \theta & t_x \\ s \sin \theta & s \cos \theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (14)$$

where t_x and t_y denote translations along the horizontal and vertical directions, respectively.

(2) Global photometric transformations: Linear contrast adjustments and non-linear gamma corrections are introduced:

$$P'_c(x, y) = \alpha P_c(x, y) + \beta \quad (15)$$

$$P'_{\text{out}}(x, y) = 255 \times \left(\frac{P'_c(x, y)}{255} \right)^\gamma \quad (16)$$

The correction factor γ is sampled randomly between 0.5 and 1.5 to simulate specular overexposure or underexposure in dark areas. Comparisons before and after augmentation are illustrated in Fig. 6.

Following these non-destructive augmentation strategies, the final training set was expanded to 14,458 images, providing the network with photometric and pose samples that adhere to underground physical principles. Unlike conventional data augmentation pipelines, aggressive random mosaics, partial cropping, and severe perspective distortions that disrupt the topology of slender rigid bodies were discarded, strictly preserving the longitudinal continuity of drill pipe projections, physical boundary integrity, and photometric fidelity in confined spaces.

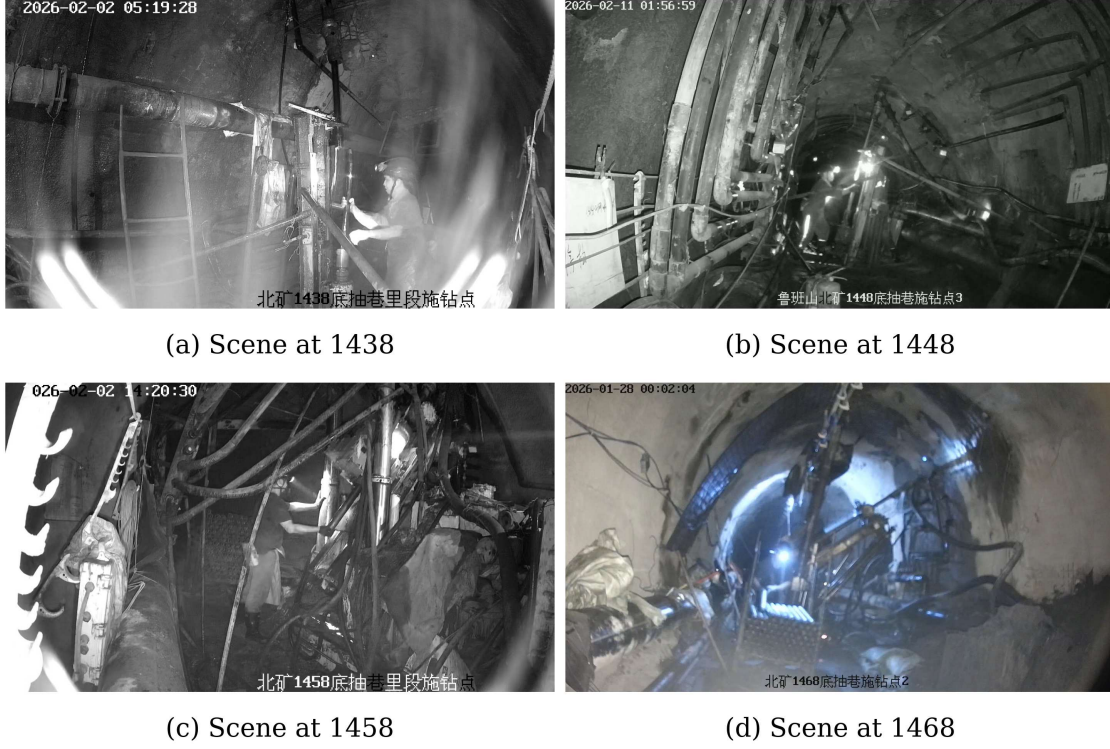


Fig. 5: Dataset samples of drill pipe operating scenes at different mining sites

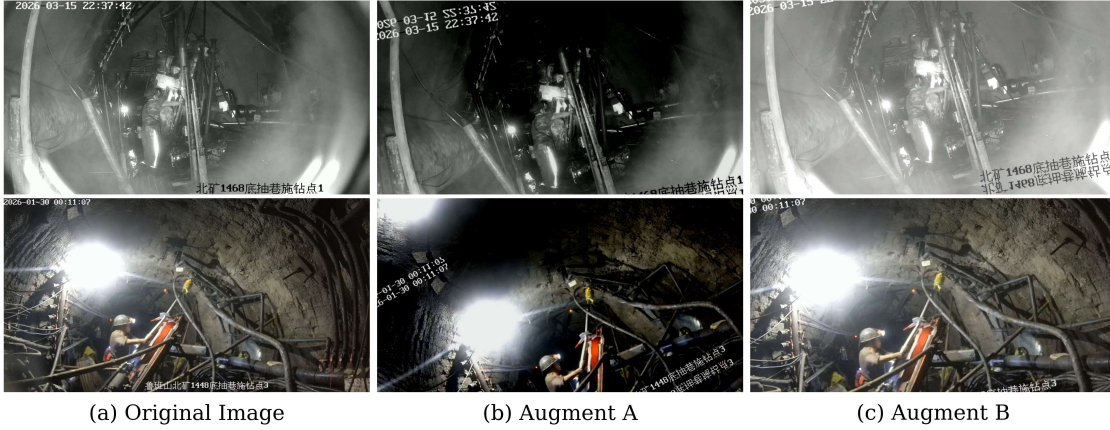


Fig. 6: Comparison of non-destructive data augmentation based on controlled geometric and photometric transformations

3 Experiments and Result Analysis

3.1 Experimental Setup

Experiments were implemented using the PyTorch 2.0 framework on an NVIDIA RTX 5880 Ada GPU. To establish a standardized baseline for comparison, the network input resolution was set to 640×640 pixels, batch size was 64, data loader workers were 4, and total training epochs were 200. The

optimizer employed an adaptive matching strategy (Auto) to dynamically adjust the learning rate. During training, image rotation (Degrees=0), image blending (MixUp=0), and hue/saturation/value perturbations (HSV=0) were disabled, retaining only Mosaic augmentation with a probability of 0.5.

For evaluation metrics, parameter count (Params), floating-point operations (GFLOPs), and inference speed (FPS) were selected to assess model complexity

and operational efficiency. Mean average precision at an IoU threshold of 0.50 (mAP@50) and mean average precision across IoU thresholds from 0.50 to 0.95 (mAP@50-95) served as core accuracy metrics. Inference speed was measured on the same NVIDIA RTX 5880 Ada GPU (48 GB memory) with default batch size and precision settings. All baseline and comparative models were trained and evaluated independently under identical hardware, software, and evaluation protocols to ensure horizontal comparability.

3.2 Comprehensive Comparison with Mainstream Lightweight Detectors

To evaluate the relative advantages of the proposed framework among lightweight architectures, it was benchmarked horizontally against mainstream lightweight models; comprehensive evaluation results are listed in Table 2 and Fig. 7. The models exhibit distinct trade-offs among accuracy, computational complexity, and inference speed, whereas our method delivers higher generalization under underground conditions characterized by frequent occlusions and dust-illumination interference.

In terms of detection accuracy, the proposed model achieves 88.7% mAP@50 and 43.7% mAP@50-95, outperforming the baseline YOLO11n by 2.0% and 2.2%, respectively. This indicates that the structural restructuring improves bounding box localization tightness under strict IoU thresholds. Compared with competitive models such as YOLOv8n and YOLO12n, our method maintains an advantage across core metrics. Regarding complexity and inference speed, although YOLO26n achieves lower parameter count (2.504 M), lower computational cost (5.8 GFLOPs), and higher inference frame rate (2013.7 fps), its mAP@50 drops to 84.4%. This indicates that excessive model compression restricts the network’s capacity to delineate drill pipe contours and model cross-occlusion dependencies. In contrast, our proposed model maintains parameter count (2.561 M) and computational cost (6.4 GFLOPs) comparable to the baseline while achieving the highest detection accuracy at an inference speed of 1303.8 fps, balancing lightweight design and real-time execution. It is noted that the inference speed of the improved model is about 16% lower than that of the baseline (1303.8 vs. 1556.1 fps), which stems from the additional attention computation introduced at the end of the backbone. Considering that the absolute speed still substantially exceeds the frame rate of underground

surveillance cameras (typically 25–30 fps), this overhead has negligible impact on real-time deployment, and the accuracy improvement justifies the modest computational cost.

3.3 Ablation Study on Core Modules

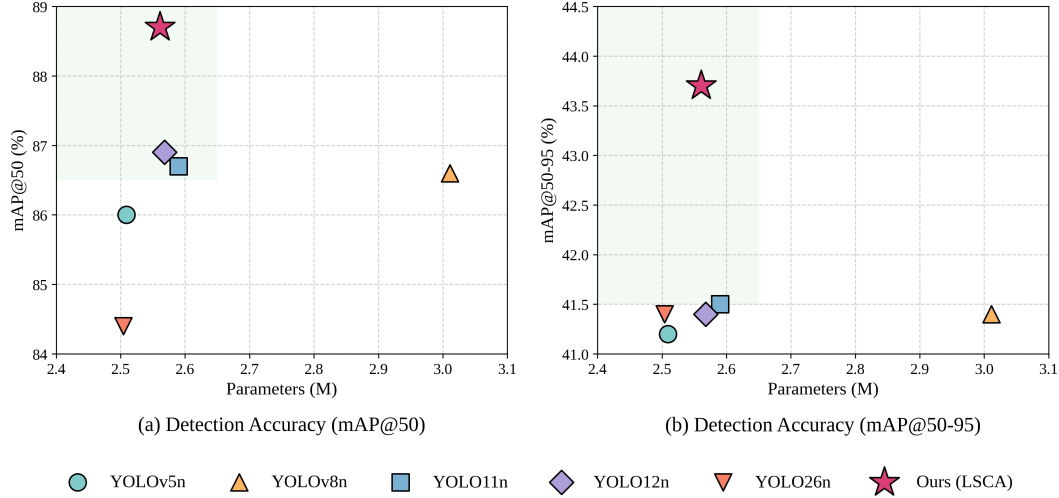
This set of experiments adopts a restructuring strategy prioritizing the loss function followed by the feature network to evaluate the individual contributions of each module to model generalization under complex working conditions. First, under the configuration of C2PSA-LSKA with CIOU loss, the large-kernel size k in Branch A was optimized. The results are presented in Table 3: $k = 11$ yields the highest mAP@50 (87.5%) and recall (81.2%) among the tested sizes, whereas $k = 23$ reduces precision to 85.4%, indicating that an excessively large receptive field introduces background noise. It is noted that the differences among $k \in \{3, 7, 11, 17\}$ are small and fall within the range of training stochasticity quantified in Section 3.4; $k = 11$ is adopted as the default configuration owing to its best recall and mAP@50.

Building on this, progressive ablation experiments were conducted (Table 4). Comparing Step 1 with Step 2, while retaining the native C2PSA in the backbone, replacing CIOU with InterIoU increases mAP@50 and mAP@50-95 from 86.7% and 41.5% to 86.9% and 43.3%, respectively, with recall improving to 81.8%, demonstrating that InterIoU improves bounding box localization accuracy.

To isolate the contribution of each attention branch, Steps 3–4 evaluate the coordinate branch alone (C2PSA-CA), and Steps 5–6 the large-kernel branch alone (C2PSA-LSKA), each under both loss functions. Under CIOU, the coordinate branch alone raises mAP@50 from 86.7% to 87.3% (Step 3) and the large-kernel branch alone to 87.5% (Step 5); under InterIoU, the corresponding values are 87.8% (Step 4) and 88.2% (Step 6). In both cases, the parallel combination of the two branches (C2PSA-LSCA, Steps 7–8) further exceeds either single-branch configuration, reaching 88.1% and 88.7% mAP@50, respectively. This consistent pattern—combined larger than either single branch—verifies the complementarity of the cross-occlusion contextual modeling and the physical edge locking, and supports the parallel multiplicative fusion design. The full model (Step 8) achieves the best overall performance, with mAP@50

Table 2: Comprehensive performance comparison of mainstream lightweight models

Model	Params (M)	GFLOPs	Inference Speed (fps)	mAP@50	mAP@50-95
YOLOv5n	2.509	7.2	1677.7	86.0%	41.2%
YOLOv8n	3.011	8.2	1624.6	86.6%	41.4%
YOLO11n	2.590	6.4	1556.1	86.7%	41.5%
YOLO12n	2.568	6.5	1367.8	86.9%	41.4%
YOLO26n	2.504	5.8	2013.7	84.4%	41.4%
Ours	2.561	6.4	1303.8	88.7%	43.7%

**Fig. 7:** Pareto frontier comparison of parameter count and detection accuracy across different models**Table 3:** Performance comparison and optimization of different kernel sizes in LSKA

Kernel Size	Precision	Recall	Params (M)	Speed (fps)	mAP@50
$k = 3$	88.2%	80.7%	2.556	1297.9	87.1%
$k = 7$	87.4%	81.1%	2.557	1476.4	87.3%
$k = 11$	87.5%	81.2%	2.557	1366.0	87.5%
$k = 17$	87.3%	80.8%	2.558	1332.1	87.1%
$k = 23$	85.4%	80.6%	2.559	1339.4	87.3%

and mAP@50-95 reaching 88.7% and 43.7%, respectively. The comparative dynamic training curves are depicted in Fig. 8.

The interpolation coefficient λ determines the position of the transitional box between the prediction and the ground truth, and thus the balance of the regression gradient. To justify the default setting, the full model (Step 8) was retrained with $\lambda \in \{0.3, 0.5, 0.7\}$, and the results are listed in Table 5. The intermediate value $\lambda = 0.5$ achieves the best mAP@50 and mAP@50-95: a smaller λ biases the interpolated box toward the ground truth and weakens the corrective gradient transmitted to the prediction,

whereas a larger λ retains more of the short-edge deviation of the prediction and re-introduces oscillation. Therefore, $\lambda = 0.5$ was adopted in all experiments.

3.4 Validation of Data Augmentation Effectiveness

To quantify the efficacy of the non-destructive data augmentation strategy described in Section 2.2, Table 6 presents baseline performance comparisons before and after data augmentation in panels (a) and (b), respectively. After applying augmentation, mAP@50 across baseline models generally improves by 1.5% to 2.2% (YOLO11n improves from 84.9%

Table 4: Progressive ablation experiments of improved modules

Step	Backbone End	Loss Function	Precision	Recall	Params (M)	GFLOPs	Inference Speed	mAP@50	mAP@50-95
							(fps)		
Step 1	Native C2PSA	CIoU	85.2%	80.3%	2.590	6.4	1556.1	86.7%	41.5%
Step 2	Native C2PSA	InterIoU	86.1%	81.8%	2.590	6.4	1307.3	86.9%	43.3%
Step 3	C2PSA-CA	CIoU	86.7%	82.2%	2.542	6.4	1244.7	87.3%	43.5%
Step 4	C2PSA-CA	InterIoU	88.2%	82.7%	2.542	6.4	1064.4	87.8%	43.3%
Step 5	C2PSA-LSKA	CIoU	87.5%	81.2%	2.557	6.4	1366.0	87.5%	42.4%
Step 6	C2PSA-LSKA	InterIoU	87.6%	81.4%	2.557	6.4	1242.2	88.2%	43.5%
Step 7	C2PSA-LSCA	CIoU	88.1%	81.0%	2.561	6.4	1357.7	88.1%	42.7%
Step 8	C2PSA-LSCA	InterIoU	88.3%	81.7%	2.561	6.4	1303.8	88.7%	43.7%

Table 5: Sensitivity analysis of the interpolation coefficient λ in InterIoU

λ	Precision	Recall	mAP@50	mAP@50-95
0.3	89.2%	80.6%	87.7%	43.5%
0.5	88.3%	81.7%	88.7%	43.7%
0.7	88.0%	82.8%	88.4%	43.4%

to 86.7%). Meanwhile, mAP@50-95 remains broadly stable or slightly improved, with isolated baseline fluctuations within $\pm 0.3\%$ (e.g., YOLO11n varying from 41.8% to 41.5%), which falls within normal testing variance. Because data augmentation does not modify network architectures, parameter counts and computational loads remain identical before and after augmentation; minor differences in reported inference speeds arise from environmental fluctuations across independent testing runs and lack statistical significance. This demonstrates that the augmentation strategy effectively expands training diversity and enhances detection accuracy without degrading bounding box regression quality or inference efficiency.

To further examine the robustness of the reported improvements against training randomness, two supplementary analyses were conducted. First, the baseline and the final model were each retrained once under identical settings: the retrained baseline achieved 85.9% mAP@50 and 41.8% mAP@50-95, and the retrained final model 88.5% and 44.3%, reproducing margins of 2.6 and 2.5 percentage points over the retrained baseline, while the run-to-run variation of each configuration itself remained within 0.8 percentage points. Second, the same comparison was cross-checked on the pre-augmentation training regime, where the improved model reaches 87.1% mAP@50 versus 84.9% for the baseline (a 2.2

percentage-point margin), compared with the 2.0-point margin (88.7% vs. 86.7%) under the augmented regime. In both regimes, the module-induced gains substantially exceed the 0.2–0.4 percentage-point differences among individual ablation configurations, indicating that the reported improvements are robust to training stochasticity.

3.5 Detection Results Analysis in Complex Scenarios

To visually verify the anti-interference capability of the model under real dynamic conditions, three representative extreme scenarios—intense glare, mechanical occlusion, and complex pipeline backgrounds—were selected for visual comparison (Fig. 9). Detection performance in extreme scenarios directly reflects operational robustness during field deployment. As indicated by the detection results: In Scenario 1 (intense glare interference), glare spots conceal local textures, causing YOLO11n to miss the target, whereas the proposed model and YOLO26n overcome photometric distortion and successfully detect the drill pipe; in Scenario 2 (mechanical occlusion), the middle segment of the drill pipe is severed, causing YOLO26n to miss the target, while the proposed model and YOLO11n maintain successful detection; in Scenario 3 (complex pipelines and low illumination), facing a low signal-to-noise ratio and parallel bundle interference, both YOLO11n and YOLO26n fail, whereas our model isolates background noise and accurately detects the drill pipe using the long-range connectivity and directional locking mechanisms of the LSCA module.

In summary, the proposed model reduces the miss rate and maintains bounding box consistency under complex working conditions. Because automated passive counting logic relies on the temporal evolution of

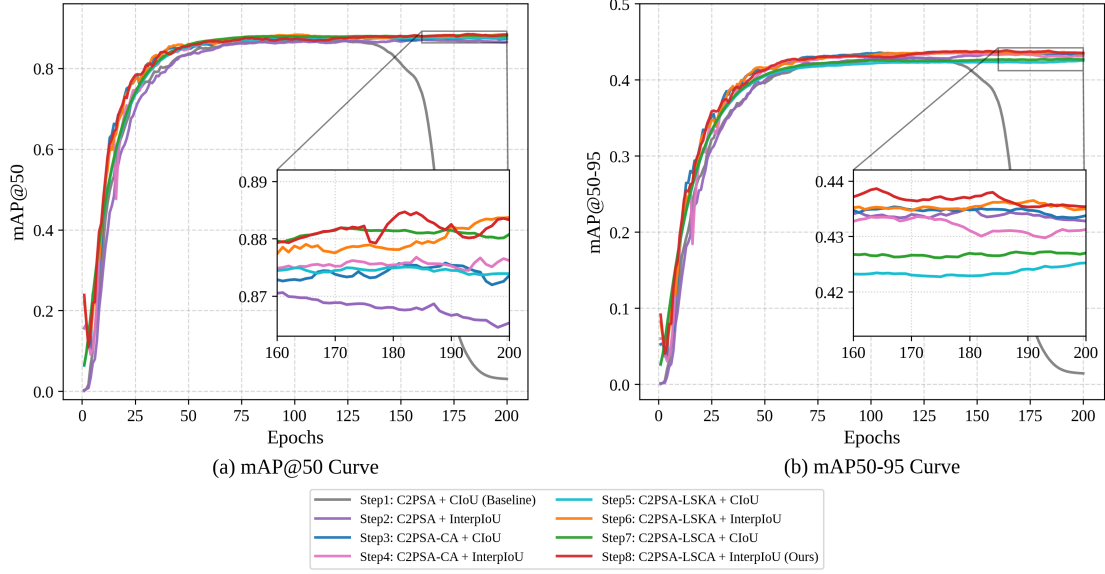


Fig. 8: Comparison of dynamic mAP training curves for the progressive ablation study of core modules

Table 6: Baseline model performance before and after non-destructive data augmentation

Model	Params (M)	GFLOPs	Speed (fps)	mAP@50	mAP@50-95
<i>(a) Before augmentation</i>					
YOLOv5n	2.509	7.2	1658.8	84.2%	40.6%
YOLOv8n	3.011	8.2	1728.6	84.7%	41.0%
YOLO11n	2.590	6.4	1645.4	84.9%	41.8%
YOLO12n	2.568	6.5	1386.5	84.7%	41.3%
YOLO26n	2.504	5.8	1857.2	82.9%	40.2%
<i>(b) After augmentation</i>					
YOLOv5n	2.509	7.2	1677.7	86.0%	41.2%
YOLOv8n	3.011	8.2	1624.6	86.6%	41.4%
YOLO11n	2.590	6.4	1556.1	86.7%	41.5%
YOLO12n	2.568	6.5	1367.8	86.9%	41.4%
YOLO26n	2.504	5.8	2013.7	84.4%	41.4%

bounding boxes, this model mitigates tracking interruptions caused by front-end misses or box jitter, providing a robust perceptual foundation for intelligent drill pipe counting.

3.6 Quantitative Layered Evaluation under Illumination and Scale Conditions

To provide quantitative, rather than solely visual, support for the anti-interference claims, the OOD test set was stratified by two image-level illumination statistics—the mean grayness $\bar{g}$ and the ratio of highlight pixels (gray value > 230)—as well as by target scale (per-image median instance area). Each

dimension was partitioned into three equal-frequency strata (394 images per stratum), yielding objective and reproducible boundaries; the miss rate is defined as $1 - \text{recall}$. The layered results are listed in Tables 7 and 8.

Under illumination stratification (Table 7), the improvements concentrate exactly where photometric interference is most severe. In the brightest grayness stratum, precision rises from 86.6% to 92.0% (+5.4 pp) with the miss rate reduced by 1.5 pp; in the medium and high highlight-ratio strata, the miss rate decreases by 3.8 pp and 1.0 pp, respectively. The dim stratum improves consistently across all metrics. In the normal stratum the two models perform comparably, with the largest adverse difference being 1.3 pp in

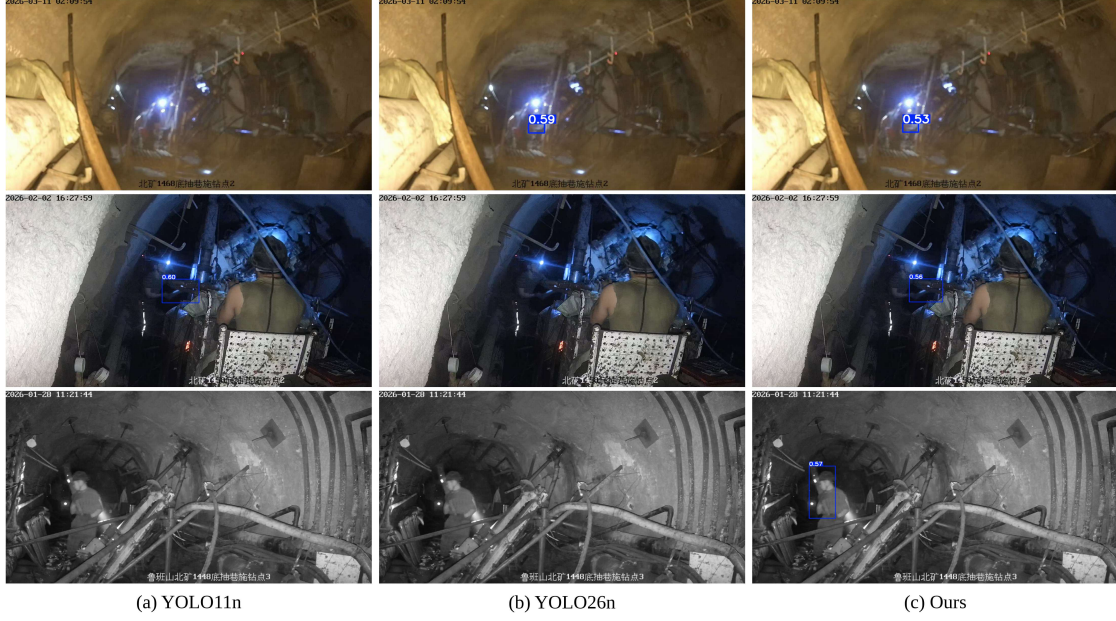


Fig. 9: Detection visualization comparison of different models under complex scenarios

precision, while recall still improves by 1.9 pp, indicating that the gains under extreme conditions are not obtained at the cost of benign scenes. These results quantitatively corroborate the glare-suppression capability of the coordinate-locking branch analyzed in Sections 3.5 and 3.8.

Under scale stratification (Table 8), precision improves in every stratum and most markedly for the smallest targets (+5.3 pp), while the remaining metrics of the small stratum stay comparable (within 1.3 pp). Meanwhile, the absolute performance of both models remains lowest in the small-target stratum (mAP@50 of 81.5% and mAP@50-95 of 34.5% for the improved model), which is consistent with the counting degradation in the distant-small-target video case (Video 8, Table 9) and is mainly attributable to the feature-resolution loss of the deeper layers under the 640×640 input. Long-range small drill pipes therefore mark the current boundary of the proposed method; higher-resolution inputs and a dedicated shallow detection head are left as future work.

3.7 Feature Heatmap Distribution Analysis

To investigate the underlying mechanisms behind the enhanced anti-interference capability of the improved model under extreme conditions, feature activation heatmaps corresponding directly to the scenarios

in Fig. 9 were extracted for comparative analysis (Fig. 10). The heatmaps were generated via forward feature activation visualization: for the multiscale feature maps (P3, P4, P5) received by the detection heads of the three models, maximum activation values were computed pixel-wise along the channel dimension; the activation maps from all scales were upsampled to the original image resolution and fused via element-wise maximum pooling, followed by normalization and Gaussian smoothing before being overlaid on the original image in JET colormap format. This visualization reflects the distribution of spatial evidence utilized during network inference.

A comparison reveals that the high-response regions (red and yellow zones) of the baseline YOLO11n exhibit pronounced divergence, being easily distracted by miner lamp glare (Scenario 1), personnel or machinery contours (Scenario 2), or background pipeline bundles (Scenario 3). Meanwhile, the overly compressed YOLO26n displays widespread ineffective activations across complex backgrounds. In contrast, the high-response centers of the proposed model concentrate precisely on the drill pipe surface. Even under intense glare or severe occlusion, the model maintains a continuous and focused response along the principal axis of the drill pipe, confirming the practical contribution of dual physical priors to generalization under challenging operational conditions.

Table 7: Layered detection performance under different illumination conditions on the OOD test set (394 images per stratum; miss rate = 1 − recall)

Stratum	Model	Precision	Recall	Miss Rate	mAP@50	mAP@50-95
<i>(a) Stratified by mean grayness $\bar{g}$ (tertiles)</i>						
Dim ($\bar{g} \leq 85.1$)	YOLO11n	86.0%	81.3%	18.7%	87.2%	42.9%
	Ours	88.1%	82.1%	17.9%	88.0%	44.5%
Normal ($85.1 < \bar{g} \leq 97.6$)	YOLO11n	88.9%	78.9%	21.1%	88.3%	44.3%
	Ours	87.6%	80.8%	19.2%	87.6%	45.1%
Bright ($\bar{g} > 97.6$)	YOLO11n	86.6%	79.9%	20.1%	86.1%	41.3%
	Ours	92.0%	81.4%	18.6%	87.9%	42.2%
<i>(b) Stratified by highlight-pixel ratio (gray value >230, tertiles)</i>						
Low ($\leq 3.29\%$)	YOLO11n	88.1%	79.1%	20.9%	86.7%	42.9%
	Ours	88.9%	80.2%	19.8%	86.9%	43.4%
Medium ($3.29-5.33\%$)	YOLO11n	84.3%	79.7%	20.3%	86.8%	42.5%
	Ours	87.3%	83.5%	16.5%	88.0%	43.7%
High ($>5.33\%$)	YOLO11n	88.4%	81.2%	18.8%	88.2%	43.1%
	Ours	89.5%	82.2%	17.8%	88.4%	44.4%

Table 8: Layered detection performance across target scales on the OOD test set, stratified by per-image median instance area (tertiles)

Scale Stratum	Model	Precision	Recall	Miss Rate	mAP@50	mAP@50-95
Small ($\leq 0.785\%$)	YOLO11n	77.9%	75.6%	24.4%	81.4%	34.1%
	Ours	83.2%	74.3%	25.7%	81.5%	34.5%
Medium ($0.785-1.595\%$)	YOLO11n	87.5%	83.5%	16.5%	88.0%	41.1%
	Ours	90.9%	83.5%	16.5%	88.3%	42.1%
Large ($>1.595\%$)	YOLO11n	91.8%	86.1%	13.9%	91.4%	51.9%
	Ours	92.9%	86.8%	13.2%	92.6%	53.7%

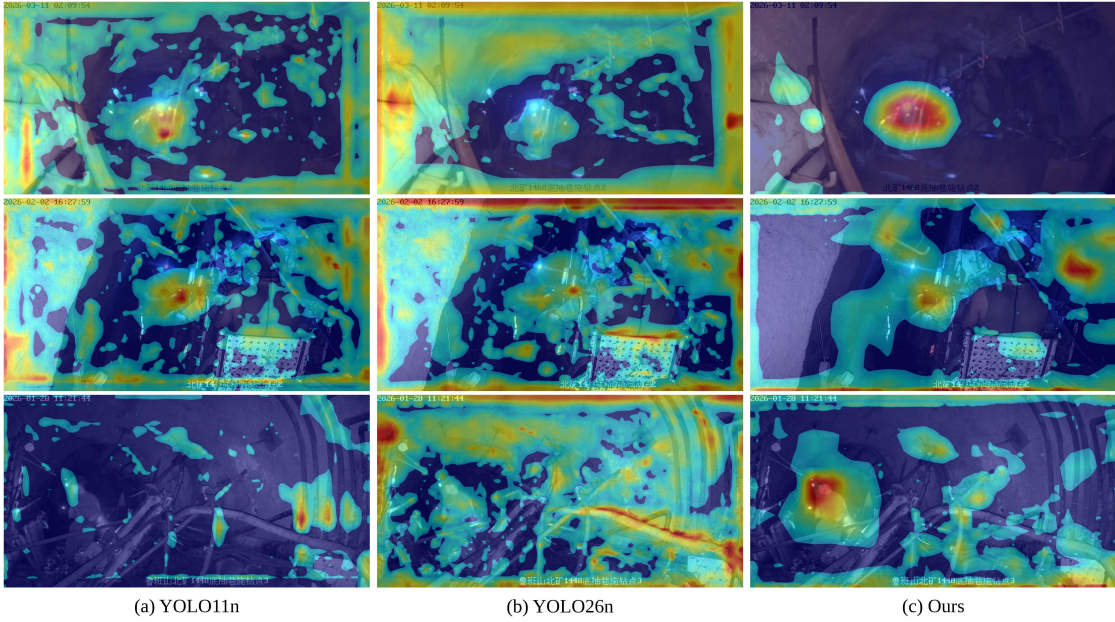


Fig. 10: Comparison of feature focus heatmaps among different models under complex scenarios

3.8 Evaluation of Drill Pipe Counting Accuracy

To verify the accuracy improvement in overall system counting brought by front-end detection optimization, five low-interference videos and five complex-interference videos were selected for tracking and

counting tests using the same back-end tracking algorithm, ByteTrack. In Table 9, manual count represents the frame-by-frame verified ground truth N_g ; counting accuracy is defined as $A = \min(N_m, N_g) / \max(N_m, N_g)$, where N_m is the model count—a formulation that penalizes both undercounting and overcounting; counting error is defined as $N_m - N_g$. The tracker was employed with its default configuration, without task-specific parameter tuning, and identical settings were applied to both models to ensure a fair comparison.

In regular scenarios (Video 1 to Video 5), both baseline and proposed models achieve high accuracy. However, when confronting complex interferences in Video 6 to Video 10 (severe occlusion, overexposure, distant small targets, etc.), the baseline model’s accuracy drops sharply to 52.9% under “distant small targets” (Video 8) and 67.5% under “overexposure” (Video 7). In contrast, across the five challenging video cases, the proposed model achieves an average counting accuracy of 78.7%, representing an improvement of 9.1 percentage points over the baseline’s 69.7%, verifying that enhanced front-end visual perception provides essential support for closed-loop back-end temporal tracking. It is noted that in Video 5, the proposed model yielded 96.2% (51 vs. 53), slightly below the baseline. Both configurations operate near the ceiling in this interference-free short video; the two-pipe deficit of the proposed model results from occasional brief front-end misses, which fragmented one tracked trajectory. For short videos with small ground-truth counts ($N_g = 53$), one or two missed detections translate into a visible accuracy drop (roughly 2 pipes $\approx$ 3–4 pp), whereas the same absolute jitter is negligible in longer videos. This level of fluctuation falls within normal short-video counting jitter, and the advantage of the proposed model emerges consistently under interference conditions (Videos 6–10).

4 Conclusion

(1) A drill pipe detection algorithm integrating large-kernel separation and coordinate locking mechanisms is proposed. Out-of-distribution (OOD) evaluation based on spatial and temporal isolation shows that the improved model achieves 88.7% mAP@50 and 43.7% mAP@50-95, representing improvements of 2.0% and 2.2% over the baseline, respectively, and improves counting accuracy under challenging conditions by 9.1 percentage points.

(2) The restructured C2PSA LSCA module combined with the InterIoU flexible regression loss effectively suppresses background noise under conditions of intense glare, pipeline slicing, and mechanical occlusion, reducing the miss rate of drill pipes and maintaining stable bounding box morphology. Quantitative layered evaluations further show that the gains concentrate under glare and highlight conditions (e.g., a 5.4 pp precision gain in the brightest stratum and a 3.8 pp miss-rate reduction in the medium highlight stratum), while distant small targets remain the current boundary of the method, to be addressed in future work.

(3) The improved model contains 2.561 M parameters and requires 6.4 GFLOPs, attaining an inference speed of 1303.8 fps, providing a robust and readily deployable visual perception solution for automated drill pipe operations in underground coal mines.

Consent for publication. For images that may lead to the identification of personal information, this manuscript has obtained informed consent from the relevant personnel in the images, agreeing to publish the personal images involved in Figures 5, 6, 9, and 10 in open access publications. In addition, the figures and tables in this manuscript do not contain any personal names or other identity information, and no blurring or color blocks have been used for anonymous processing.

Data Availability. The experimental data and customized dataset generated in this study are subject to coal mine safety and confidentiality regulations, but are available from the corresponding author upon reasonable request.

Funding. This work was supported by the National Natural Science Foundation of China (Grant Nos. 52074296 and 52004286) and the Talent Program of Chengdu Technological University (Grant No. 2025RC096).

Author Contributions. **Canjun Xiao:** Conceptualization, overall scheme design, funding acquisition, project administration, manuscript review and editing. **Zaisheng Jiang:** Corresponding author. Methodology design, algorithm implementation, experimental design and execution, original draft preparation. **Chenyu Wang:** Experimental execution, data curation, visualization. **Junxiang Yang:** Software development, data processing. **Siliang Zheng:** Resources (field data collection and operational site investigation). **Youping Li:** Resources (field data collection and

Table 9: Evaluation of drill pipe counting accuracy

Video ID	Challenge	Manual Count	Baseline Count	Baseline Accuracy	Counting Error	Proposed Count	Proposed Accuracy	Counting Error
Video 1	None	137	130	94.9%	-7	135	98.5%	-2
Video 2	None	86	89	96.6%	+3	83	96.5%	-3
Video 3	None	128	126	98.4%	-2	125	97.6%	-3
Video 4	None	144	142	98.6%	-2	143	99.3%	-1
Video 5	None	53	53	100%	0	51	96.2%	-2
Video 6	Partial occlusion	132	97	73.5%	-35	110	83.3%	-22
Video 7	Overexposure	83	56	67.5%	-27	60	72.3%	-23
Video 8	Distant small target	136	72	52.9%	-64	82	60.3%	-54
Video 9	Blur	122	83	68.0%	-39	97	79.5%	-25
Video 10	Large spatiotemporal shift (OOD)	59	51	86.4%	-8	58	98.3%	-1

operational site investigation). **Dongdong Chen:** Validation, manuscript review. **Shengrong Xie:** Supervision, manuscript review and editing. All authors read and approved the final manuscript.

Conflict of Interest. The authors declare that they have no conflict of interest.

Correspondence. Correspondence and requests for materials should be addressed to Z. Jiang (zaisheng_j@163.com).

References

- [1] ZHAO Y, YAN Y, LIU K, et al. Analysis of coal mine safety accident features in China, 2017–2022[J]. *Geohazard Mechanics*, 2024, 2(2): 108–120.
- [2] FAN C, LAI X, WEN H, et al. Coal and gas outburst prediction model based on principal component analysis and improved support vector machine[J]. *Geohazard Mechanics*, 2023, 1(4): 319–324.
- [3] LI Zhen, LIU Yiming, FENG Guorui, et al. Development and application of visual observation system for gas-solid coupling damage and catastrophe in coal and rock[J/OL]. *Journal of Mining and Strata Control Engineering*, 2025. DOI: 10.13532/j.jmsce.cn10-1638/td.2025-1406 (in Chinese).
- [4] GUO Weiyao, GUO Yuhao, WANG Xiangyu, et al. Pressure relief and energy release of large-diameter drilling method in the variable region of coal thickness[J]. *Journal of Mining and Strata Control Engineering*, 2026, 8(3): 033051. DOI: 10.13532/j.jmsce.cn10-1638/td.2025-1223 (in Chinese).
- [5] ZHONG Taoping, LI Zhenlei, HE Xueqiu, et al. Research on rock burst prevention by shallow rock mass weakening in steeply inclined and extremely thick coal seams[J]. *Journal of Mining and Strata Control Engineering*, 2024, 6(3): 033039. DOI: 10.13532/j.jmsce.cn10-1638/td.2024.03.003 (in Chinese).
- [6] LIN Haifei, LIU Sibao, SHUANG Haiqing, et al. Overburden rock fracture evolution law and pressure relief gas extraction technology of gob-side entry retaining mining[J]. *Journal of Mining and Strata Control Engineering*, 2024, 6(1): 013537. DOI: 10.13532/j.jmsce.cn10-1638/td.20240019.001 (in Chinese).
- [7] HU Jiaheng, DONG Lihong, QIN Yi. Drill pipe counting method based on working state discrimination of underground drilling rig in coal mine[J]. *Industry and Mine Automation*, 2025, 51(12): 56–62. DOI: 10.13272/j.issn.1671-251x.2025050080 (in Chinese).
- [8] WANG Xiangqian, SHI Ce. Drill pipe recognition and tracking counting method based on improved YOLOv5 and DeepSORT[J]. *Industry and Mine Automation*, 2022, 48(9): 45–52 (in Chinese).
- [9] ZHANG Fukai, SUN Yiran, WU Xufeng, et al. Real time counting method for coal mine drill pipes based on deep learning[J]. *Coal Science and Technology*, 2025, 53(6): 493–504. DOI: 10.12438/cst.2024-0366 (in Chinese).
- [10] ZHANG Fukai, ZHAO Shuo, ZHANG Qiang, et al. Coal mine underground drill pipe counting method based on rotating object detection and saliency peak inference[J]. *Journal of China Coal Society*, 2025, 50(S2): 1389–1403 (in Chinese).
- [11] LI Jianhua, ZHANG Maosheng, LI Baochun, et al. Dual-encoding edge enhancement object detection method for underground coal mine drilling[J/OL]. *Coal Science and Technology*, 2026 (in Chinese).
- [12] JIANG Longhao. Research and implementation of intelligent monitoring system for rescue drilling rig

- based on digital twin[D]. Xuzhou: China University of Mining and Technology, 2025 (in Chinese).
- [13] FANG Jie, LI Zhenbi, XIA Liang. Drill pipe counting method based on ECO-HC[J]. Coal Technology, 2021, 40(11): 186–189 (in Chinese).
 - [14] DONG Lihong, HU Jiaheng, QIN Yi, et al. Automatic tracking and counting system for underground drill pipe based on Camshift algorithm[J]. Coal Engineering, 2022, 54(8): 155–160 (in Chinese).
 - [15] HAO Tianxuan, FENG Huiyan, LI Xu, et al. Coal mine drill rod operation behavior recognition and counting method based on YOLO11 and spatio-temporal graph convolutional network[J]. Coal Science and Technology, 2026, 54(7): 349–363. DOI: 10.12438/cst.2025-1082 (in Chinese).
 - [16] DU Jingyi, DANG Mengke, QIAO Lei, et al. Drill pipe counting method based on improved spatio-temporal graph convolutional neural network[J]. Industry and Mine Automation, 2023, 49(1): 90–98 (in Chinese).
 - [17] ZHANG Dong, JIANG Yuanyuan. Drill pipe counting method based on improved MobileNetV2[J]. Industry and Mine Automation, 2022, 48(10): 69–75 (in Chinese).
 - [18] ZONG Zhanqiang. Underground drill pipe drilling action recognition based on AlexNet network[J]. Electronics World, 2021(17): 118–120 (in Chinese).
 - [19] MITTAL P. A comprehensive survey of deep learning-based lightweight object detection models for edge devices[J]. Artificial Intelligence Review, 2024, 57(9): 242. DOI: 10.1007/s10462-024-10877-1.
 - [20] ZOU Z, CHEN K, SHI Z, et al. Object detection in 20 years: A survey[J]. Proceedings of the IEEE, 2023, 111(3): 257–276.
 - [21] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 7464–7475.
 - [22] BAO J, XIAO C, GUO D, et al. CLGDS: Robust bridge crack detection with YOLO enhanced feature fusion and SIoU optimization[J]. Scientific Reports, 2026, 16: 16337. DOI: 10.1038/s41598-026-42727-1.
 - [23] LI Z, XIAO L, SHEN M, et al. A lightweight YOLOv8-based model with Squeeze-and-Excitation Version 2 for crack detection of pipelines[J]. Applied Soft Computing, 2025, 177: 113260. DOI: 10.1016/j.asoc.2025.113260.
 - [24] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8(3): 331–368.
 - [25] LAU K W, PO L M, REHMAN Y A U, et al. Large separable kernel attention: Rethinking the large kernel attention design in CNN[J]. Expert Systems with Applications, 2024, 236: 121352.
 - [26] ZHANG H, LIU J, HU G. FCN attention enhancing asphalt pavement crack detection through attention mechanisms and fully convolutional networks[J]. Scientific Reports, 2025, 15: 25257. DOI: 10.1038/s41598-025-92971-0.
 - [27] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 13708–13717.
 - [28] ZHENG Z, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020: 12993–13000.
 - [29] ZHANG Y F, REN W Q, ZHANG Z, et al. Focal and efficient IOU loss for accurate bounding box regression[J]. Neurocomputing, 2022, 506: 146–157.
 - [30] XIONG C, ZAYED T, ABDELKADER E M. A novel YOLOv8-GAM-wise-IoU model for automated detection of bridge surface cracks[J]. Construction and Building Materials, 2024, 414: 135025. DOI: 10.1016/j.conbuildmat.2024.135025.
 - [31] LIU H, WATANABE H. InterIoU: Rethinking bounding box regression with interpolation-based IoU optimization[J/OL]. arXiv preprint arXiv:2507.12420, 2025. DOI: 10.48550/arXiv.2507.12420.
 - [32] LIU H, WATANABE H. InterIoU: Robust bounding box regression loss within an interpolation-based IoU framework[J]. Neurocomputing, 2026, 666: 132230. DOI: 10.1016/j.neucom.2025.132230.