Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Research on Multispectral Human Target Detection Algorithm in Underground Mining

Wei Li

Heilongjiang University of Science and Technology

Haocheng He

h17352180952@hotmail.com

Heilongjiang University of Science and Technology

Research Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11325346/v1

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

Additional Declarations: No competing interests reported.

Research on Multispectral Human Target Detection Algorithm in Underground Mining

Wei Li, Haocheng He*

School of Energy and Mining Engineering. Heilongjiang University of Science and Technology, Harbin 150022, China

**Corresponding author. Email: h17352180952@hotmail.com*

Abstract To address the instability of single-modal detection caused by weak illumination, dust occlusion, and thermal-source interference in underground coal mines, a dual-stream detection framework integrating visible-light and infrared images was developed. Independent preprocessing pipelines were designed for the two modalities, and early fusion was performed along the channel dimension to construct a unified bimodal input representation. On this basis, a gated feature enhancement mechanism was introduced into the You Only Look Once version 5s network to adaptively select informative modal features and improve target perception in complex underground environments. A public paired visible-light and infrared dataset collected under low-light conditions was used to simulate underground operating scenarios, and both ablation and comparative experiments were conducted. The results show that the infrared unimodal model outperformed the visible-light unimodal model, with the accuracy increasing from 94.33% to 96.11% and the mean average precision improving by 2.61%, indicating the higher reliability of infrared imaging in low-light conditions. In addition, the proposed bimodal fusion scheme outperformed all unimodal baselines, and the split-stream four-channel fusion configuration achieved a mean average precision of 98.41%. Although the proposed model showed slightly lower localization performance than the comparison model at stricter evaluation thresholds, it reduced the number of parameters, computational cost, and memory consumption by 43.3%, 38.8%, and 44.7%, respectively, demonstrating a better balance between detection accuracy and model complexity.

Keywords Infrared thermal imaging, Visible light, Gating mechanism, feature enhancement, dual-stream input

Funding declaration: This work was supported by the Basic Scientific Research Business Expense Project of Heilongjiang Provincial Universities (Grant No. 2025-KYYWF-ZR0613).

1 Introduction

In response to the limited computational resources and real-time requirements in underground coal mines, the development of efficient vision computing architectures and intelligent object detection algorithms that jointly achieve high accuracy and lightweight deployment has become an important research direction for intelligent mine safety monitoring. Such methods are expected to enable real-time perception, identification, and warning of personnel activities around hazardous areas (CHENG et al. 2025; CHENG et al. 2023). Compared with open above-ground environments, underground video surveillance scenarios are characterized by greater environmental complexity and higher perception difficulty. On the one hand, underground illumination is generally insufficient, and suspended dust, local water accumulation, and reflections from equipment surfaces jointly degrade image quality. As a result, visible-light images are prone to uneven brightness, reduced contrast, degraded texture details, and blurred target boundaries. On the other hand, pedestrian movement, equipment operation, conveyor motion, and slight camera vibration may further introduce motion blur, leading to additional degradation of target shape information. This problem becomes more severe under long-distance, small-scale, and occluded conditions, thereby increasing the likelihood of missed and false detections (Yang et al. 2025; Diana et al. 2025). In addition, underground backgrounds usually contain complex structures such as supports, pipelines, cables, transportation equipment, and stacked materials. These background objects may share

similarities with personnel in terms of shape, scale, and local texture, which further increases the difficulty of visual recognition.

Infrared imaging does not rely on external visible illumination and can highlight human targets through thermal radiation information under low-light or even no-light conditions. Therefore, it provides important complementary perception capability in underground scenarios (HWANG et al. 2015). However, the infrared modality does not exhibit stable advantages under all operating conditions. During continuous operation, underground mechanical equipment often forms high-temperature radiation sources, while transfer points, motors, rollers, and friction zones may generate locally bright thermal regions. These factors can lead to thermal-source aggregation, halo diffusion, and high background responses in infrared images. When a human target is located close to high-temperature equipment, its thermal signature may overlap with the mechanical heat source or be submerged by the locally strong thermal background, resulting in blurred target boundaries, incomplete contours, and increased localization deviations (LI et al. 2019). Therefore, neither visible-light imaging nor infrared imaging alone can stably balance target saliency, boundary integrity, and background interference resistance in complex underground environments. This also indicates that personnel intrusion detection in hazardous areas requires visual perception methods capable of comprehensively exploiting the complementary information of multiple modalities.

Liang et al. (2025) adopted a dynamically gated adaptive pruning-based object detection model, in which adaptive pruning was triggered dynamically and the optimized images were then fed into a feature enhancer for cross-modal fusion. Jin. (2025) improved the YOLOv5s-based equipment detection model and significantly enhanced both the detection accuracy and efficiency of YOLOv5s in complex environments through network structure optimization. Zhang Jiewen et al. (2025) introduced a dynamic detection head, a pyramid split attention module, partial convolution, and grouped depthwise separable convolution to establish a lightweight network that reduced model complexity while maintaining recognition accuracy. Nan Jingjing et al. (2025) employed a Retinex-based image enhancement method to effectively separate and remove low-quality features while preserving discriminative reflectance features, thereby improving low-light enhancement performance. Although these improved models have enhanced detection accuracy, optimized network structures, and improved the overall detail quality of input images, several issues still remain to be addressed.

(1) Existing methods usually rely on a single image source and thus lack a comprehensive description of the environment, which affects detection accuracy under different operating conditions.

(2) Existing methods commonly process infrared and visible-light images independently after separate input. Although this strategy may reduce memory consumption, unified preprocessing may damage the temperature gradient information inherent in infrared imaging because of the distinct physical characteristics of the two modalities.

To address the above issues, a dual-modal sensing scheme and a gated YOLOv5 model with dual-stream early fusion were developed. Key functions, including image processing, target detection and calibration, moving-target detection, and decision-making, were systematically designed. By introducing dual-end signal input and optimizing the overall algorithmic architecture, real-time infrared thermal perception-based personnel protection monitoring in hazardous areas around underground conveyors was achieved. Experimental results show that the multimodal gated YOLOv5 model with front-end dual-stream input provides faster inference speed, more stable deployment, and lower dependency on external conditions. The proposed system can accurately analyze a variety of typical operating conditions and demonstrates promising application potential.

2 Visible-Light and Infrared Dual-Modal Object Detection

2.1 YOLOv5 Model

At present, the YOLO series has been updated to YOLOv12 (LIAO et al. 2025), with substantial improvements introduced in successive versions. For

example, YOLOv6 redesigned the backbone network and was specifically optimized for industrial applications. YOLOv7 introduced an efficient layer aggregation network and achieved high overall performance. YOLOv8 further upgraded the backbone architecture and exhibited stronger comprehensive capability (WANG et al. 2025). Although the YOLO series (WU et al. 2022) has undergone multiple generations of evolution and later versions have shown better overall detection performance, YOLOv5 still offers distinct advantages in terms of its mature open-source ecosystem, adaptability to lightweight deployment in practical applications, and balanced trade-off between detection speed and accuracy. Therefore, it is well suited for rapid engineering implementation and secondary development. The basic idea of YOLOv5 is to formulate object detection as a regression problem, in which the category, location, and size of target objects are directly predicted. Specifically, the algorithm first extracts features from the input image and then performs object detection on the generated feature maps.

The YOLOv5 network consists of four main components, namely the input layer, backbone, neck, and head. The backbone adopts a three-stage feature extraction architecture composed of convolution-batch normalization-SiLU blocks, C3 modules, and the spatial pyramid pooling-fast module, thereby constructing the fundamental feature space. The convolution-batch normalization-SiLU block is formed by stacking a two-dimensional convolution layer, a batch normalization layer, and a SiLU activation function. The C3 module establishes a feature reuse and compensation mechanism through the collaborative design of three convolutional layers and a deep residual path, which enhances model stability. The spatial pyramid pooling-fast module (Kaiming et al. 2015) employs a progressively increasing pooling-window strategy to aggregate multi-scale features while maintaining size consistency (SHI et al. 2023).

In the neck, the upsampling branch reconstructs spatial resolution through deconvolution operations, enabling high-level semantic information to guide low-level features. The downsampling path compresses feature scales through strided convolution, allowing precise localization cues to be transmitted to higher-level features and forming a dynamic collaborative correction mechanism within the feature pyramid. In the head, three groups of prediction grids with progressively increasing resolutions（20^2/40^2/80^2）are deployed. During forward propagation, geometric intersection-over-union optimization is used for bounding-box regression, and an adaptive confidence-aware non-maximum suppression algorithm is finally applied to eliminate redundant detections and generate the final prediction results. The overall architecture of YOLOv5 is shown in Fig.1.

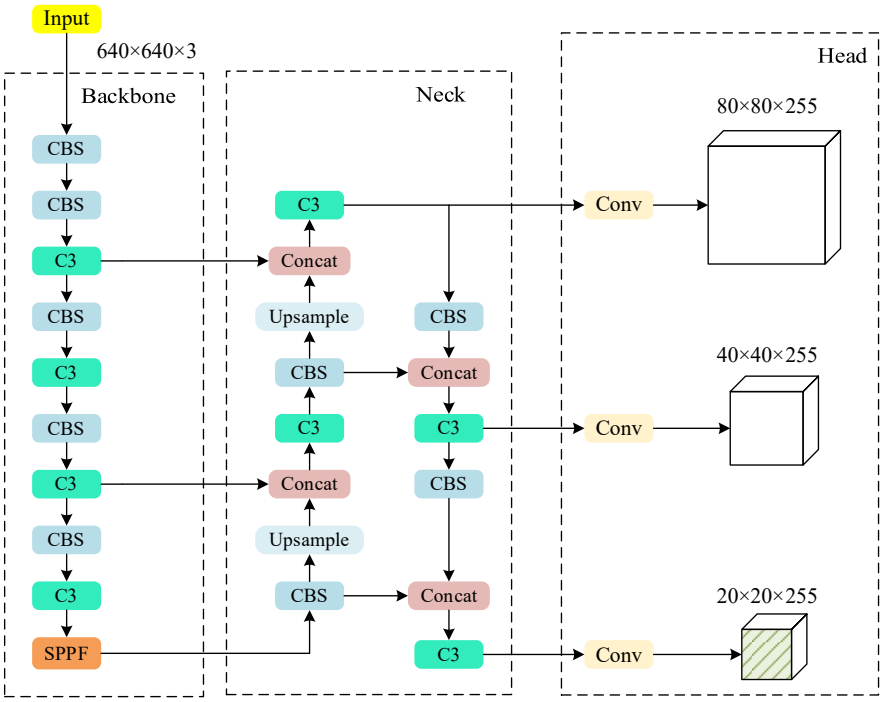


Fig.1 YOLOv5 network structure

2.2 Dual-Modal Input and Feature Fusion Design

2.2.1 Dual-Modal Preprocessing and Alignment

Visible-light images mainly provide texture, edge, and structural information of targets, whereas infrared images primarily reflect the thermal response intensity of targets. Significant differences therefore exist between the two modalities in terms of data distribution and representation patterns (HWANG et al. 2015). To reduce the influence of cross-modal scale differences on network training, a dual-modal input strategy with separate preprocessing branches was adopted, in which visible-light and infrared data were independently read and preprocessed. In particular, to suppress the influence of high-response regions and extreme values in infrared images on training stability, the infrared images were first normalized. The expression is given as follows:

$$I'_{ir}(x,y)=\frac{I_{ir}(x,y)-I_{ir}^{\min}}{I_{ir}^{\max}-I_{ir}^{\min}} \qquad (1)$$

Where $I_{ir}^{\max}$ and $I_{ir}^{\min}$ denote the maximum and minimum pixel values of the infrared image, respectively, and $I^{*}_{ir}(x,y)$ denotes the normalized pixel value.

To ensure that the visible-light and infrared images remain spatially aligned before fusion, the same Letterbox resizing and padding parameters were applied to both modalities, and both were uniformly adjusted to the network input size of $S\times S$。 Let the original image size be $H\times W$ and the target input size be $S\times S$. Then, the scaling ratio r is defined as follows:

$$r=\min\left(\frac{S}{H},\frac{S}{W}\right) \qquad (2)$$

By applying a consistent scaling ratio and padding strategy to both visible-light and infrared images, a one-to-one correspondence between the two modalities can be maintained at the pixel level, thereby providing a spatial alignment basis for subsequent fusion. After preprocessing, the two input streams are uniformly converted into tensors in CHW format and concatenated along the channel dimension to form a unified bimodal input representation, which is then fed into the shared backbone network for subsequent feature extraction, as shown in Fig.2. This strategy preserves the original information differences between the two modalities while achieving early fusion at the input level, thereby enabling the model to simultaneously exploit the texture information from visible-light images and the thermal features from infrared images, and ultimately improving the stability of human target detection in complex environments.

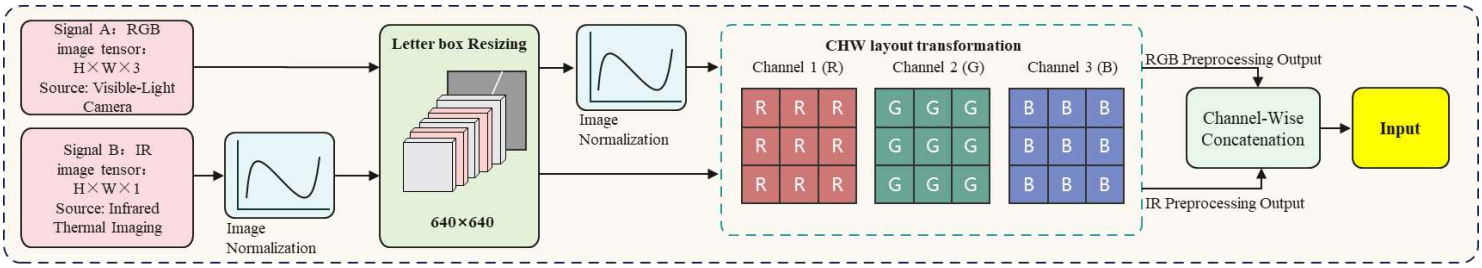


Fig.2 Dual-Modal Data Preprocessing

2.2.2 Early Fusion Method at the Input Layer

During data input, OpenCV was uniformly used to read the visible-light and infrared data through two separate branches. Since images and video frames are read by OpenCV in the default BGR order, the channel order of visible-light images needs to be converted from B (blue), G (green), and R (red) to R, G, and B to avoid color distortion. In contrast, a single-channel grayscale infrared image does not contain the trichromatic combination of reflected visible light. Its data essentially represent the thermal radiation intensity emitted by the object itself, rather than separated and recombined color components. Therefore, it contains only one physical quantity, and channel order does not need to be considered.

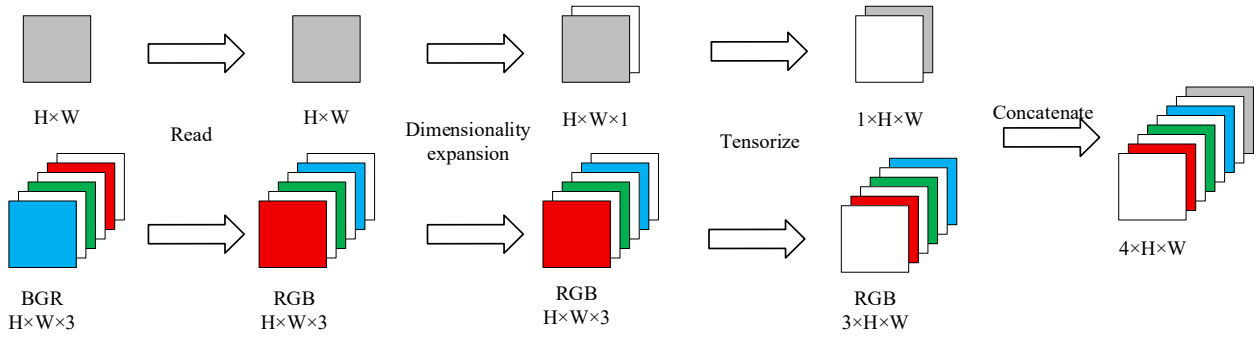


Fig.3 Early Fusion Process of Dual-Modal Inputs

Performing infrared normalization in advance, immediately after data loading, enables the infrared data to be brought to the same scale before concatenation with the visible-light channels, thereby reducing differences in gradient magnitude and activation distribution between the two modalities. By constraining the infrared intensity values to a stable range at an early stage, the dominance of one modality and the resulting weight bias in the first convolutional layer after multimodal concatenation can be effectively avoided. This strategy is also more conducive to convergence when the first-layer weights are extended for concatenated inputs. In addition, applying normalization to the infrared data prior to cropping and denoising helps suppress the influence of extreme values and improves the stability of subsequent enhancement and model learning. The two branches share the same Letterbox parameters and random geometric transformations. Specifically, the visible-light and infrared images are subjected to identical resizing and padding strategies, and all pixels are uniformly adjusted to 640×640, thereby ensuring that each pixel corresponds to the same scene location after channel concatenation. To satisfy the input tensor format required by the network, the single-channel infrared image is further expanded to a tensor with an explicit channel dimension. Therefore, the infrared image is unsqueezed by adding one channel, so that the grayscale image conforms to the input requirements of the deep learning model.

Dimension expansion only changes the tensor shape and does not alter the pixel values. For the single-channel infrared image, dimension expansion was implemented by adding an explicit channel

dimension, rather than by fixed-value padding or replicated expansion. The tensor form of the original infrared image can be expressed as follows:

$$X \in \mathbb{R}^{H \times W} \tag{3}$$

By introducing an additional dimension of length 1 along the channel axis, it can be represented as follows:

$$X \in \mathbb{R}^{H \times W \times 1} \tag{4}$$

Where $\boldsymbol{R}$ denotes the real number domain, indicating that $\boldsymbol{X}$ is an H×W×1 matrix in which each element is a real number.

Unlike fixed-value channel padding, the channel-dimension expansion method does not introduce additional non-informative channels that may interfere with feature learning, and the first convolutional layer is not affected by constant-value bias. The convolutional weights in the first layer act only on the single effective channel, thereby ensuring the statistical stability and distributional consistency of early feature extraction. Meanwhile, this strategy is consistent with the input requirements of YOLOv5 and common convolutional neural networks, and can therefore be regarded as a standard procedure for processing single-channel infrared images. After the infrared image is expanded to H×W×1, it is concatenated with the visible-light image of size H×W×3 along the channel dimension, thereby ensuring the correctness of the fused C×H×W layout. After the two input streams are transformed from HWC to CHW or NCHW , early fusion can be directly performed along the channel dimension, making the fused representation compatible with deep learning models.

2.3 Hybrid Gated Attention Mechanism

2.3.1 Spatial Gating in Backbone Fusion

Spatial gating (LI et al. 2026) is a dynamic attention mechanism based on the spatial positions of feature maps, and its core idea is to enable the neural network to adaptively focus on key regions in the feature map. Fusion spatial gating performs spatial gating on already fused features. In the proposed design, a dual-stream backbone is adopted for visible-light and infrared inputs, where modal features are first extracted separately in the shallow layers. Cross-modal information interaction and enhancement are then achieved in the backbone stage through the gating mechanism and fusion modules. This design is intended to address the specific limitations of infrared images, including low resolution, low contrast, and low signal-to-noise ratio, while suppressing pseudo-textures and noise in the visible-light branch, thereby facilitating feature learning in the downstream neck and head.

The introduction of fusion spatial gating into the backbone is mainly aimed at enhancing source features and filtering noise. During fusion within the backbone, shallow backbone features retain stronger generality, and the gating operation only performs lightweight modulation rather than fundamentally altering low-level feature representations. In this way, invalid features can be prevented from being propagated and further amplified by the detection head.

In the improved backbone design, the original 10-layer structure is retained. The dual-stream architecture is introduced at the first stage, and feature fusion begins at the third stage. The core modules are upgraded to gated versions. The backbone structures before and after improvement are shown in Fig. 4

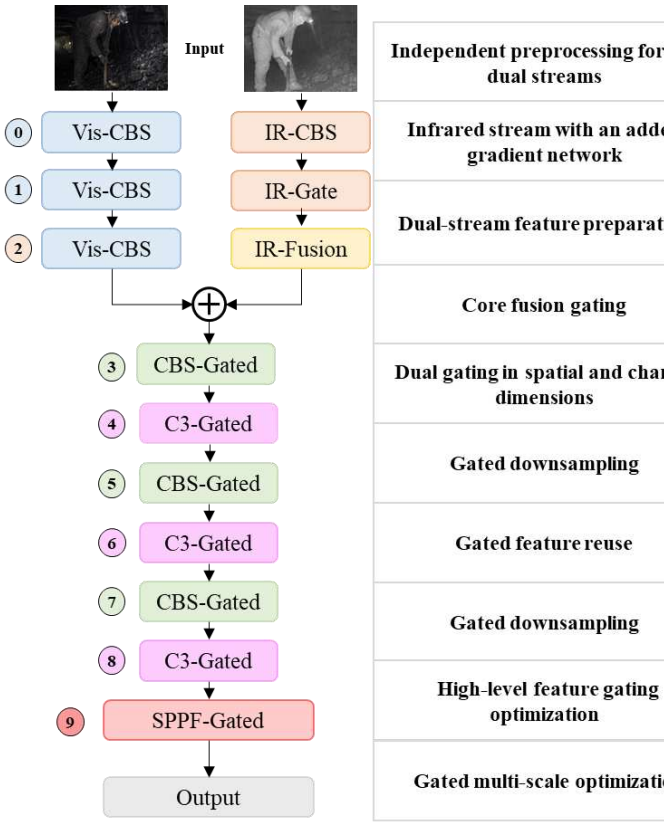
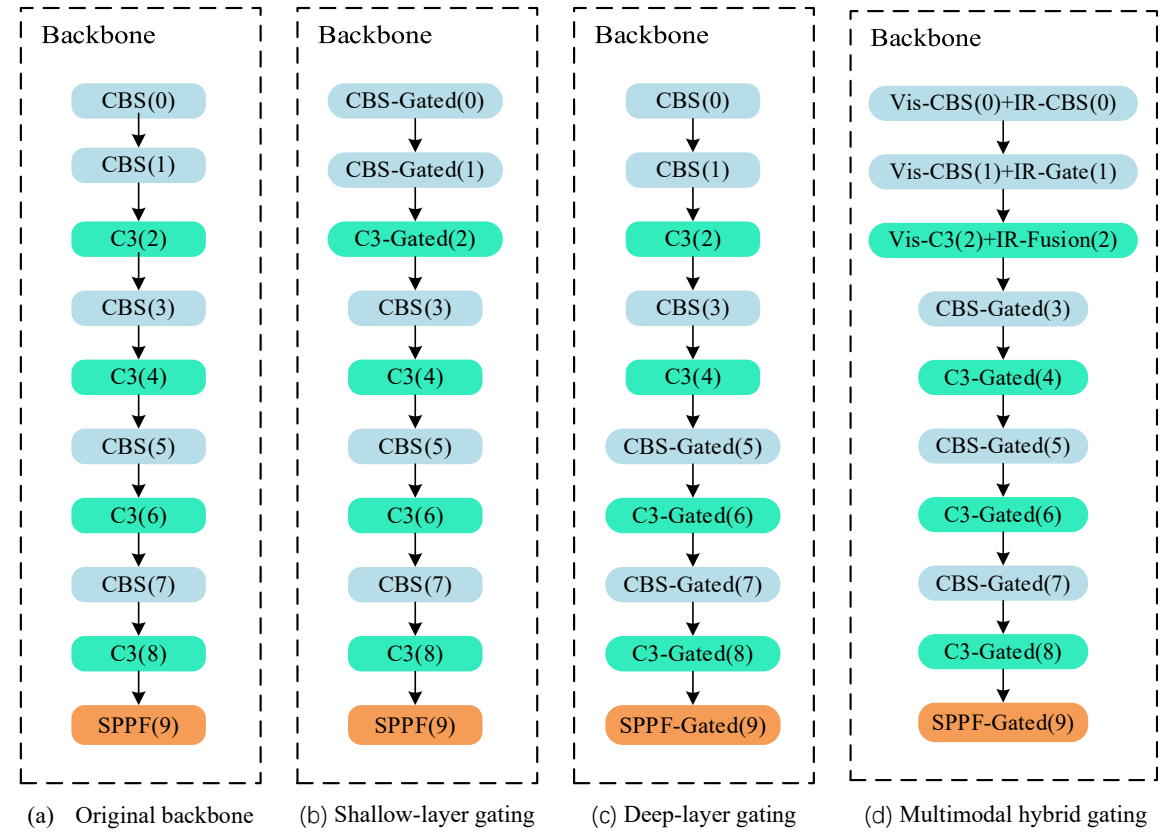


Fig.4 Vanilla versus Augmented Backbone Network

The improved backbone preserves the original architecture while introducing a gating mechanism to achieve dynamic weight allocation between infrared thermal imaging and visible-light features, thereby enhancing the multimodal fusion capability. The parameters and input dimensions of each module are listed in Table 1.

Tab.1 Cascaded Module Specifications in Backbone Network

Stage	Module	Parameter	Enter size
0	Vis-CBS	38.4 K	3×640×640
0	IR-CBS	24.6 K	1×640×640
1	Vis-CBS	73.7 K	64×320×320
1	IR-Gate	18.5 K	64×320×320
2	Vis-C3	186.2 K	128×160×160
2	IR-Funsion	65.5 K	128×160×160
3	CBS-Gated	295.0 K	256×160×160
4	C3-Gated	1.05 M	256×80×80
5	CBS-Gated	590.0 K	512×80×80
6	C3-Gated	4.19 M	512×40×40
7	CBS-Gated	2.36 M	1024×40×40
8	C3-Gated	16.76 M	1024×20×20
9	SPPF-Gated	17.89 M	2048×20×20



(a) Original backbone (b) Shallow-layer gating (c) Deep-layer gating (d) Multimodal hybrid gating

Fig.5 Gating Insertion Mechanism for Backbone Networks

2.3.2 Structure of the Infrared Gating Module

The infrared gating module first applies convolution to the visible-light and infrared features, respectively, so as to align their channel dimensions. The aligned features are then concatenated along the channel dimension to form a joint feature representation. A convolutional layer is further used to extract relevant information from the joint features, and a gating weight map is generated through the Sigmoid activation function. The gating weight is formulated as follows:

$$G = \sigma(\mathrm{Conv}(\mathrm{Concat}(F_{vis}, F_{ir}))) \quad (5)$$

Where G enotes the gating weight map, which reflects the importance of feature responses at different spatial locations, and σ denotes the activation function. F_{vis} and F_{ir} represent the feature maps extracted from the visible-light branch and the infrared branch, respectively.

The resulting weight map characterizes the importance distribution over different spatial positions. The weight map is then multiplied element-wise with the infrared features to obtain the enhanced infrared features, which are subsequently fed into the IR-Fusion module for further fusion. The infrared feature enhancement process is expressed as follows:

$$F_{ir'} = G \odot F_{ir} \quad (6)$$

Where $F_{ir'}$ denotes the enhanced infrared features after gating modulation.

The visible-light and infrared features are first jointly used to generate a gating weight map, which is then employed to enhance the infrared branch features. Therefore, the infrared feature signal multiplied element-wise by the gating weight map represents the retained branch features after gating reweighting, rather than a second infrared input signal, as shown in Fig.6.

This design enables the effective target information in the infrared branch to be adaptively enhanced according to the joint response of the visible-light and infrared modalities, while suppressing background thermal noise and irrelevant interference.

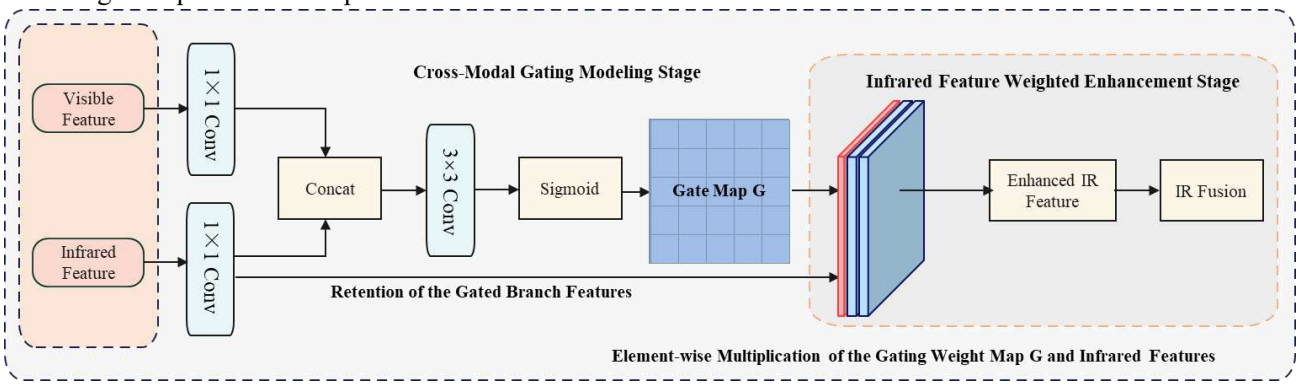
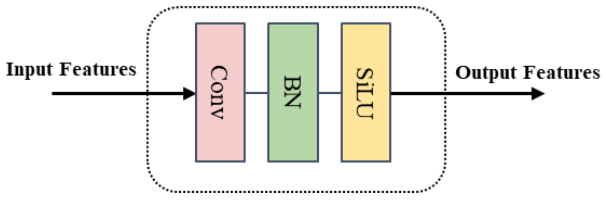


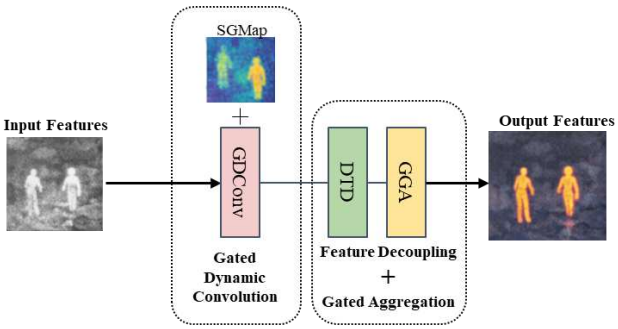
Fig.6 Structure of the IR-Gate Module

2.3.3 Improved Core Feature Extraction Module

The convolution-batch normalization-SiLU (CBS) block serves as the basic feature extraction unit. In the improved design, the original standard convolution is replaced by a processing module integrating gated perception, dynamic convolution, feature decoupling, and gated aggregation. This modification transforms the generic convolution operator into a more adaptive feature processing module, making it more suitable for complex underground environments characterized by small targets, low contrast, and strong noise interference.



(a) Original CBS Unit



(b) Improved Gated CBS Unit

Fig.7 CBS Block Enhance

The original CBS block before improvement is shown in Fig.7(a). The input channel consists of a single feature map, and the parameters of the fixed convolution kernel remain unchanged after training. Its mathematical expression is given as follows:

$$Y(i,j)=\sum_{c=1}^{C_{in}}\sum_{m=1}^{k}\sum_{n=1}^{k}W_{c,m,n}\ X_{c,i+m-1,j+n-1}+b \quad (7)$$

Where X denotes the input feature, Y 为 denotes the output feature, W denotes the fixed convolution kernel weight, which does not vary with the input after training, and b enotes the bias term. The representation capability of a fixed convolution kernel is inherently limited, because the convolution operation always uses the same set of weights. As a result, adaptability to different input scenarios can only be compensated for by subsequent layers. After fixed convolution-based feature processing, the activation mechanism relies solely on the standard SiLU activation function used in YOLO, without any dedicated optimization for thermal imaging. Therefore, its performance is limited when dealing with highly variable and complex tasks, such as infrared noise, environmental variations, and underground thermal imaging scenarios.

The improved gated CBS unit is shown in Fig.7(b). Its input consists of two branches, namely the feature map and the SG spatial gating map. The dynamic convolution kernel is generated or reconfigured adaptively according to the input features. Specifically, the dynamic kernel is first defined and then used to produce the convolution output. Its mathematical expression is given as follows:

$$W_{dyn}=\Phi(X)$$
$$Y=W_{dyn}*X \quad (8)$$

Where $\emptyset(X)$ denotes the function used to generate the convolution kernel, X denotes the input feature, Y denotes the convolution output, and W denotes the dynamically generated convolution kernel. After dynamic convolution, a dual-branch decoupling operation is performed to separately organize the thermal features of high-temperature targets and the structural features of roadway contours. Subsequently, thermal imaging is optimized through a dual-threshold gating strategy, thereby suppressing high-frequency noise and enhancing weak thermal features.

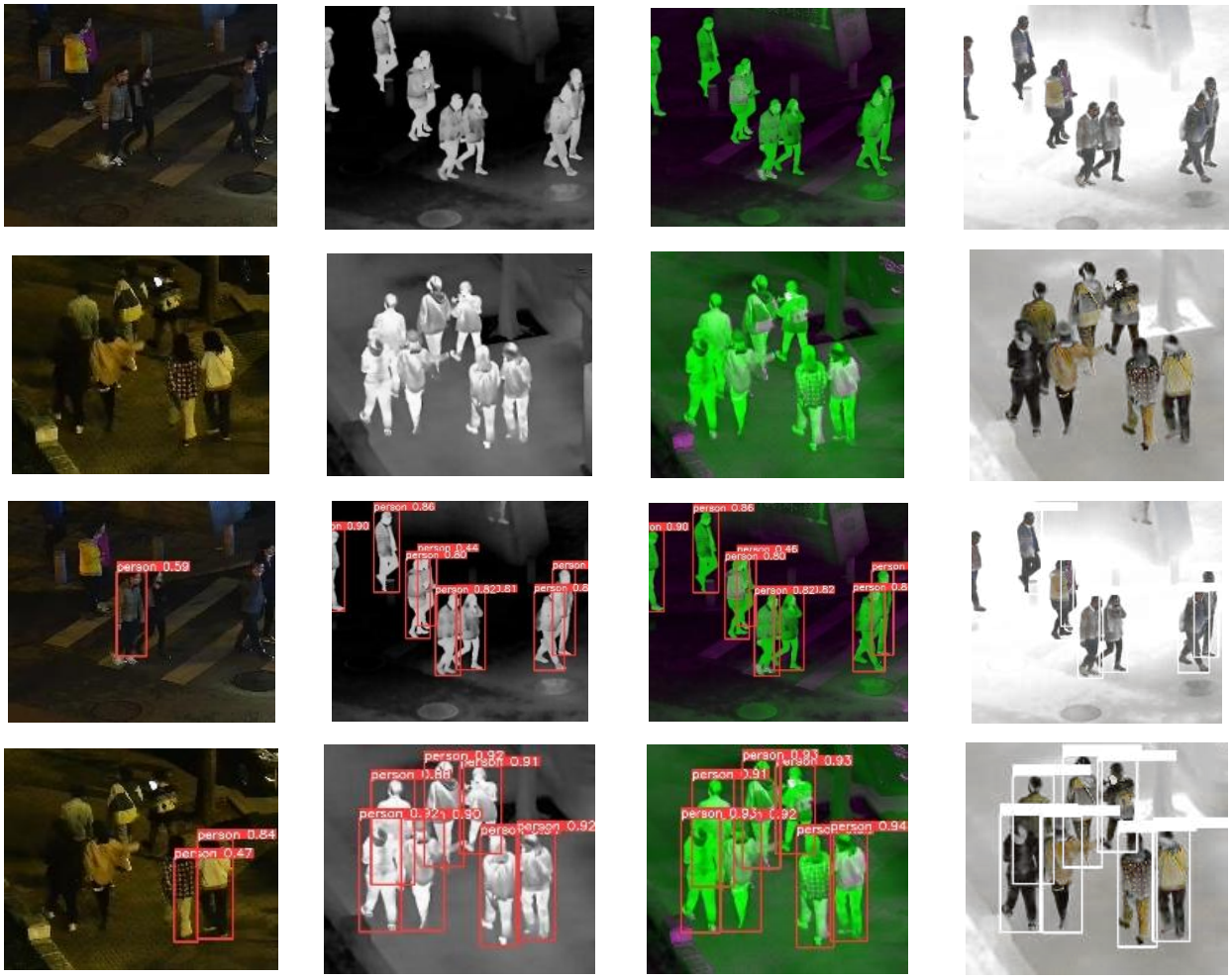
3 Experimental Results and Analysis

3.1 Experimental Setup

All experiments were conducted under the same software and hardware conditions. The hardware specifications and deep learning environment configurations are listed in Table 2. The models were trained using stochastic gradient descent (SGD) with a learning rate of 0.01. The main objective of the experiments was to verify the effectiveness of the improved algorithm for personnel detection in underground environments characterized by low illumination and high dust concentration. Particular emphasis was placed on demonstrating the contribution of the gating mechanism and bimodal fusion to the complementary exploitation of infrared and visible-light features, as well as on evaluating whether the inference speed of the improved model on edge computing devices can satisfy the requirements of underground safety monitoring.

Tab.2 Experimental Setup

System	Windows11
CPU	Intel(R)Core(TM)i9-14900HX (2.20 GHz)
GPU	NVIDIA Geforce RTX4060
RAM	16.0GB
DL Framework	Pytorch 2.5.1+cu121
Python Version	3.9.16
Cuda Version	12.9



(a)RGB Dataset (b)IR Dataset (c)3ch Dataset (d)4ch Dataset

Fig.8 Dataset

3.2 Data Preparation

The experimental dataset used in this study was the publicly available low-light visible-infrared paired dataset LLVIP (TANG et al. 2023), which contains synchronized visible-light and infrared thermal images and includes a total of 30,976 single-class target images, corresponding to 15,488 paired samples. The dataset mainly consists of pedestrian images captured under low-light conditions, which are relatively close to underground mine environments. The image resolution is 512×640. To more accurately evaluate the generalization ability of the proposed model in real underground scenarios, the dataset was divided into training, validation, and test sets at a ratio of 7:2:1.

All images were annotated using LabelImg , and the annotations were automatically converted into the YOLO format, with corresponding txt files stored in the label folder. Each annotation file contains the image category and coordinate information and corresponds to one image sample. The numbers of samples in the training, test, and validation sets were 10,841, 3,097, and 1,550, respectively. Under the experimental settings, the inference speed of the graphics processing unit at an input resolution of 640 ×640 was approximately 8.1 ms per image. Although single-class target detection usually converges rapidly, 100 epochs and a batch size of 16 were adopted in order to improve generalization under low-light conditions and complex backgrounds. Automatic mixed precision training was enabled to reduce memory consumption and accelerate training.

It should also be noted that the same infrared and visible-light data may exhibit different characteristics under different channel formats. The three-channel output format (JPG) contains only three channels, and the infrared information is nonlinearly compressed

into the red-green-blue color space. In contrast, the four-channel output format (TIFF) preserves the absolute radiometric values of the infrared modality, which can be used for temperature inversion and is therefore more suitable for radiometric measurement requirements, while retaining complete radiative information.

3.3 Evaluation Metrics

To effectively evaluate the detection results, precision(P) was adopted as the core metric for accuracy assessment, representing the proportion of correctly predicted targets among all predicted targets. Recall(R) was used as the key indicator for completeness evaluation, reflecting the ability of the model to identify all real targets. Mean average precision was taken as the core metric for object detection, representing the overall stability and performance of the trained model (LI et al. 2025). Processing time refers to the end-to-end time required to complete single-frame inference and serves as a key quantitative indicator of real-time performance. Computational cost directly affects power consumption and latency, and represents the theoretical number of floating-point operations required to process a single frame. Model weight size corresponds to the physical storage space occupied by the trained parameters and is therefore a key parameter related to deployment cost.

3.4 Results and Analysis

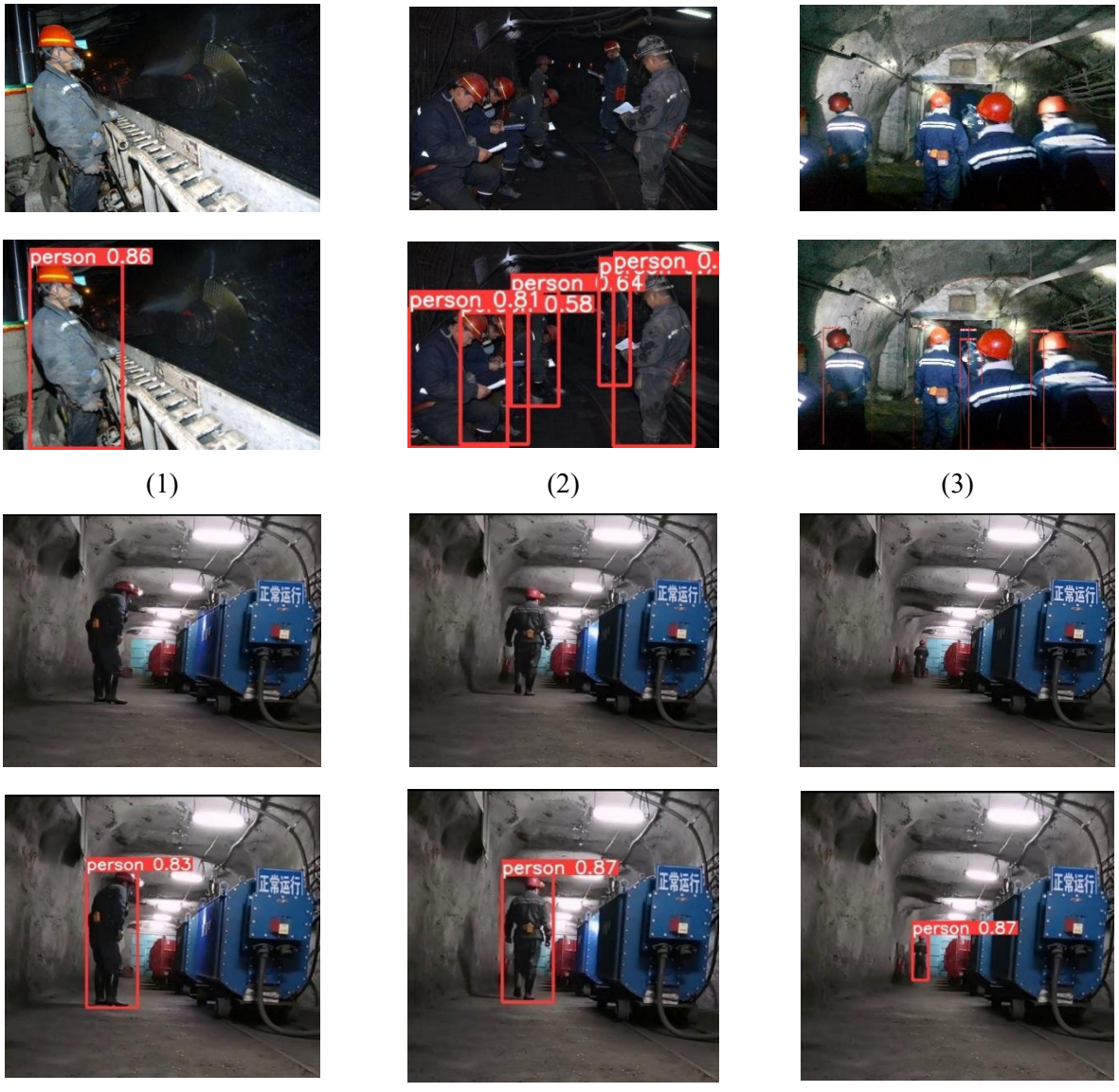
3.4.1 Detection Execution

The improved algorithm was tested on an underground dataset characterized by extremely poor illumination, dust interference, and complex human postures, as shown in Fig.9. The results demonstrate that the improved method achieves stable detection for single-person targets and can correctly separate multiple targets in crowded scenes. Even under occlusion or posture variation, the detection accuracy can still be maintained. Six representative real underground images were selected for visual analysis, as summarized in Table 3, covering typical operating conditions such as single-person close-range scenes, long-distance small targets, near-equipment background interference, dense multi-person occlusion, and motion blur. These results indicate that the proposed algorithm exhibits strong adaptability to low-light environments, satisfactory multi-person detection capability, and reliable discrimination between personnel and equipment under posture variation and complex background conditions, thereby demonstrating good adaptability to underground working environments.。

Figure	Background	Confidence	FPS	Result Description
1	Close-range equipment in a dark scene	0.86	114.15FPS	Dust noise and local overexposure do not degrade the detection results, indicating that the model has a certain tolerance to local abnormal noise and is suitable for underground environments with dust interference and unstable illumination.
2	Crowded multi-person occlusion	0.58-0.84	188.3FPS	Multiple people are clustered together with obvious mutual occlusion. The model can still respond to multiple persons under severe occlusion, indicating that it maintains a certain sensitivity to local human cues.
3	Motion blur with crowded people	0.45-0.92	155.01FPS	The scene contains multiple people, occlusion, and obvious motion blur. Although these factors cause unclear human contour boundaries and missing local features, no missed detections or false detections occur.
4	Single person at	0.83	156.54FPS	This result indicates that under conditions of limited occlusion and

	medium-close range			discernible target edges, the model can produce a strong response to human appearance structure. The large device on the right does not trigger a false detection, showing that the model has a certain suppression capability for underground equipment backgrounds.
5	Single person at medium-long range	0.87	142.12FPS	This result shows that the model still has good recall ability for persons at medium to long distances, indicating that its multi-scale feature representation can cover small targets within a certain range.
6	Long-distance small target	0.87	117.25FPS	The target is located farther away in the scene, and the proportion of small-scale features is even lower. The model still produces a high-confidence detection result, indicating that it has a certain detectability for small-scale targets.

Tab.3 Qualitative Analysis of Detection Targets



(1) (2) (3)

(4) (5) (6)

Fig.9 Underground Mine Commissioning Test

3.4.2 Ablation Study

To verify the effectiveness of the multimodal hybrid gating mechanism in improving the single-class target detection performance of the YOLOv5s baseline, an ablation study was conducted on the LLVIP dataset, and the detailed results are presented in Table 4.

As can be observed from the table, more sufficient fusion generally leads to more stable feature representation and higher mean average precision. The introduction of the gating mechanism results in a slight increase in computational cost, while the improvement in detection accuracy is accompanied by an increase in model weight size, which is consistent with the general characteristics of multimodal fusion. Under the specific conditions simulated by the dataset, the infrared unimodal model was more effective than the visible-light unimodal model, which also indicates the rationality of the dataset selection. In addition, the processing time of the four-channel early fusion model was approximately doubled compared with that of the previous model, mainly because the number of input channels increased to four, which in turn raised the computational burden of the first convolutional layer.

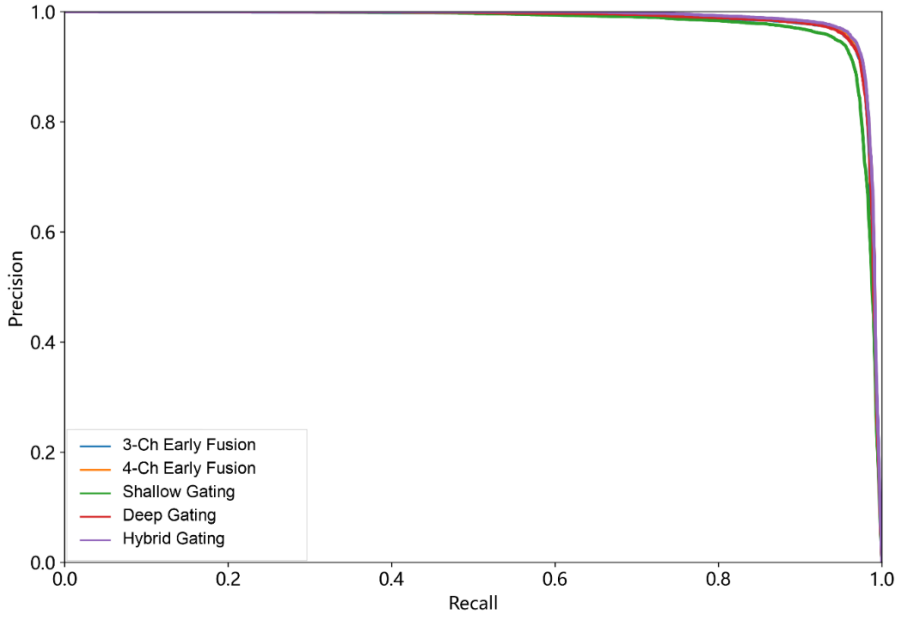


Fig.10 PR Curve Comparison Across Algorithms

GFLOPs are determined by the network architecture and input size. Therefore, the baseline network and the early fusion model exhibit the same computational cost when the overall architecture remains unchanged. After the introduction of the gating mechanism, the network structure is modified, which leads to a corresponding change in GFLOPs. However, the computational overhead introduced by the gating operation is negligible compared with that of convolutional layers, since its scale is several orders of magnitude smaller. The reported GFLOPs values were calculated with a precision of 0.1.

Regarding model size, the three-channel early fusion model had a measured weight size of 56,754,677 bytes (approximately 54.11 MB) before slimming, which was nearly identical to those of the RGB and IR baselines. Compared with the split-stream four-channel early fusion model, the difference in weight size was extremely small and remained below 2 KB. To facilitate deployment and distribution, slimmed weight files were adopted, retaining only the parameters required for inference. Overall, a moderate increase in model size was accompanied by a noticeable improvement in detection performance, indicating that the proposed design achieves a reasonable trade-off between model compactness and accuracy.

Tab.4 Ablation Results

Model	PrecisionP/%	RecallR/%	mAP@0.5/%	Computational Cost GFLOPs/F	Compressed Weight Size
YOLOv5s（RGB baseline）	94.33	90.95	95.72	16.2	13.74 MB
YOLOv5s（IR baseline）	96.11	95.31	98.33	16.2	13.74 MB
YOLOv5s three-channel early fusion	96.47	94.83	98.17	16.2	13.74 MB
YOLOv5s four-channel early fusion	96.16	95.60	98.41	16.2	13.74 MB
YOLOv5s+four-channel early fusion + shallow backbone gating	94.41	95.11	97.70	16.3	13.75 MB
YOLOv5s+four-channel early fusion + deep backbone gating	96.09	95.15	98.28	16.3	13.79 MB
YOLOv5s+four-channel early fusion + multimodal hybrid backbone gating (ours)	95.96	95.28	98.49	16.9	14.13 MB

Tab.5 Comparative experimental results

Baseline	Model	mAP@0.5/%	mAP(0.5:0.95)/%	RAM	FPS
YOLOv5l	RGB	90.8%	51.9%	1.9 GB	85.34
YOLOv5l	IR	94.6%	61.9%	1.9 GB	84.92
YOLOv8	IR	93.9%	58.6%	2.1 GB	78.45
Faster R-CNN	IR	94.6%	54.5%	2.9 GB	22.18
CRT-YOLO	IR	95.1%	59.7%	2.2 GB	74.61
IC-Fusion	RGB+IR	97.9%	70.3%	3.8 GB	35.26
DenseFuse+YOLOv5l	RGB+IR	88.2%	48.7%	2.6 GB	42.15
SeAFusion+YOLOv5l	RGB+IR	91.3%	52.3%	2.8 GB	38.42
Ours	RGB+IR	98.5%	65.6%	2.1 GB	68.83

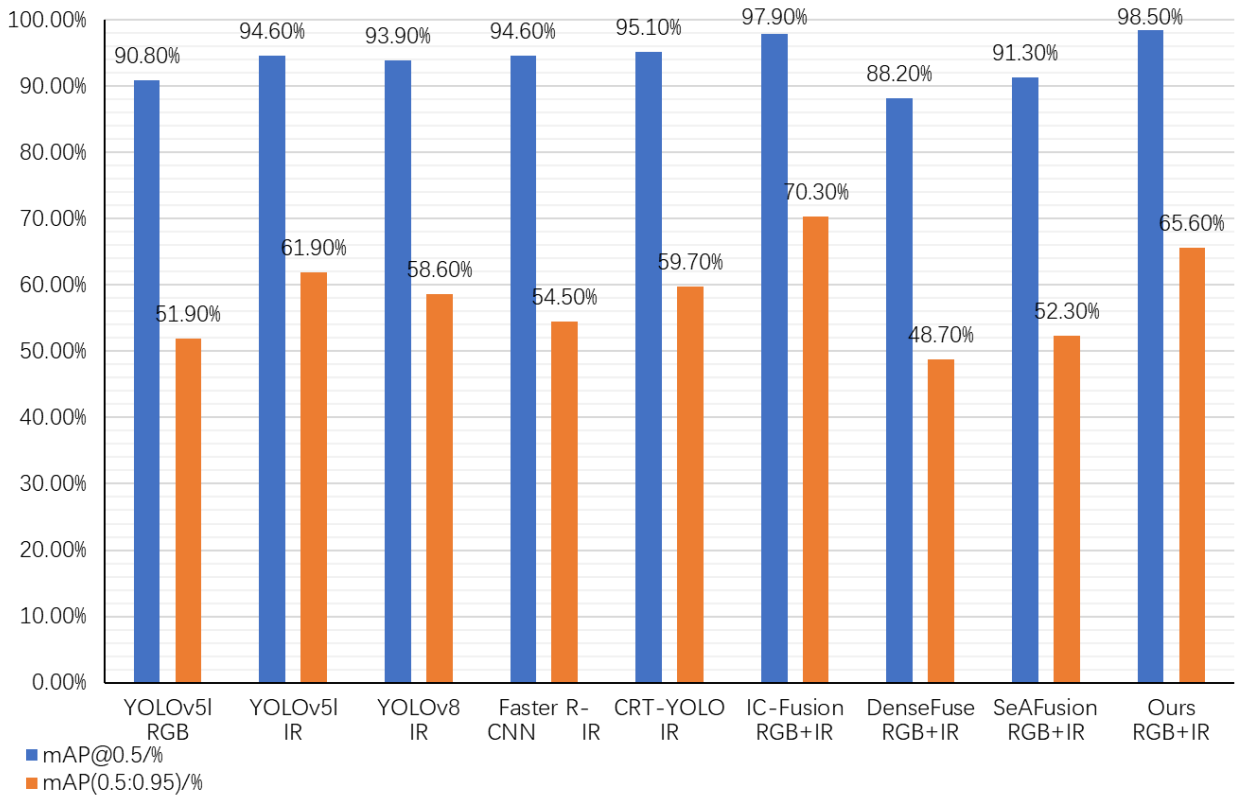


Fig.11 Detection Performance Comparison of the LLVIP Method

3.4.3 Comparative Experiments

To verify the superiority of the improved multimodal hybrid gating algorithm for single-class target recognition in underground environments, comparative experiments were conducted on the LLVIP dataset against several mainstream detection algorithms. Specifically, the single-stage detector YOLOv8 (Jia et al. 2022) and the two-stage detector Faster R-CNN (Li et al. 2025) were selected as representative comparison methods, and the results are presented in Table 5.

As shown in Fig. 11, the multimodal hybrid gating method achieved a clear improvement over the unimodal YOLOv5 baselines. Among the unimodal settings, YOLOv5 based on pure infrared input significantly outperformed that based on visible-light input, confirming that infrared data are more reliable under low-light conditions. In addition, the visible-light and infrared fusion scheme outperformed all unimodal YOLOv5 variants, which further demonstrates the complementary advantage of bimodal information.

Image-level fusion methods, such as DenseFuse +YOLOv5l and SeAFusion+YOLOv5l, also introduce bimodal information; however, their serial paradigm of fusion followed by detection leads to considerable computational redundancy (WANG et al. 2023; SMITH and DOE 2024). Compared with IC-Fusion (Hwang et al. 2025), the proposed model achieved a lower mAP@0.5:0.95 (65.6% versus 70.3%), indicating that its fine localization capability under stricter intersection-over-union thresholds remains insufficient. This limitation may be attributed to several factors. First, the proposed design mainly focuses on improving target saliency and detection stability in complex scenarios through gated enhancement, which contributes more substantially to target discoverability than to precise bounding-box regression. Second, infrared images often suffer from blurred human contours and thermal diffusion around object boundaries. Even though bimodal fusion enhances target responses, such characteristics may still reduce localization accuracy under high intersection-over-union (IoU) evaluation criteria. Third, to balance model complexity and deployment requirements, the proposed method was developed

based on YOLOv5s with an emphasis on lightweight deployment while maintaining high detection accuracy. Although IC-Fusion improves accuracy through dual-modal fusion, its dual-backbone architecture increases memory consumption to 3.8 GB and reduces the inference speed to approximately 35 frames per second, resulting in a substantially higher deployment cost.

4 Conclusions

To address the problem of optimizing the performance and efficiency of multimodal object detection in complex underground environments, systematic research and experimental validation were carried out from three aspects, namely input fusion strategy, feature regulation mechanism, and overall model design. On this basis, an improved framework was established to balance detection accuracy and lightweight deployment. Through multidimensional comparison and ablation analysis, the contribution of the gating module to overall model performance was comprehensively evaluated, and the accuracy and practical applicability of the split-stream four-channel early fusion algorithm with a multimodal gating mechanism under extreme illumination and complex background conditions were verified.

(1) By adopting a dual-stream early fusion strategy, two independent feature extraction branches were constructed. This design exhibited stronger target feature extraction capability than the single-stream input under complex underground conditions. The visible-light and infrared early fusion scheme outperformed all single-stream YOLOv5 variants, and the mean average precision of the visible-light and infrared fusion model was 4.17% higher than that of the infrared-only model, thereby demonstrating the complementary advantage of dual-stream input.

(2) The proposed hierarchical adaptive hybrid gating mechanism achieved a significant performance gain with only a 4.3% increase in computational cost, showing a substantially better trade-off than shallow or deep gating alone. In infrared-dominated detection tasks, its average precision was higher than that of the infrared baseline model, while the weight size of the hybrid gating model increased by only 2.8% compared with the infrared baseline.

(3) The ablation study and comparative experiments on the target dataset demonstrate that the split-stream four-channel early fusion gated YOLO algorithm can effectively cope with extreme illumination conditions. Among all methods evaluated on the LLVIP dataset, the proposed method achieved the highest mean average precision at an intersection-over-union threshold of 0.5, reaching 98.49%, while increasing the model size by only 0.39 MB (+2.8%) and the computational cost by 0.7 giga floating-point operations (+4.3%). Compared with models such as IC-Fusion, the proposed method is more lightweight and achieves a better balance between detection accuracy and structural complexity.

References

[1] CHENG Deqiang，SUN Fengqian，HAO Dashuai，et al. Review of object detection and hazard identification methods in mine video images[J]. Coal Science and Technology,2025,53(11):20−38

[2] CHENG Jian, LI Hao, MA Kun, et al. Architecture and key technologies of coalmine underground vision computing[J]. Coal Science and Technology, 2023, 51(9):202−218.

[3] Wenjuan Yang, Yanqun Wang, Xuhui Zhang. et al. MSS-YOLO: Multi-Scale Edge-Enhanced Lightweight Network for Personnel Detection and Location in Coal Mines[J]. Applied Sciences, 2025, 15(6): 3238.

[4] Diana Novak, Yuriy Kozhubaev, Vyacheslav Potekhin et al. Asymmetric Object Recognition Process for Miners' Safety Based on Improved YOLOv10 Technology[J]. Symmetry, 2025, 17(9): 1435.

[5] HWANG S, PARK J, KIM N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1037-1045.

[6] LI C, SONG D, TONG R, et al. Illumination-aware Faster R-CNN for robust multispectral pedestrian detection[J]. Pattern Recognition, 2019, 85: 161-171.

[7] LIANG Song, YUANYong, QIAN Jiansheng, et al. Real-Time Multi-Target Detection and Tracking Method Based on Dynamic Gating for Underground Coal Mines[J/OL]. Journal of China Coal Society.2025.1022

[8] Jin Xudong. Non-destructive Testing Method for Underground Mechanical Equipment in Coal Mines Baes-d on Infrared Temperature Measurement Technology[J]. Colliery Mechanical and Electrical Technology,2025,46(2):61-65.

[9] Zhang Jiewen1, Ma Ji1, Jiang Dong, et al. Personnel Detection Algorithm in Underground Coal Mine Based on Improved YOLOv7[J]. Coal Mine Machinery,2025,46(08):213-217.

[10] NAN Jingjing, PAN Hongguang, JIANG Ze, et al. Underground personnel recognition based on low-light enhancement of infrared and visible light image fusion[J]. Journal of Mine Automation,2025,51(4):107-113,145.

[11] LIAO L F, SONG C, WU S L ,et al. A novel YOLOv10 based algorithm for accurate steel surface defect detection[J]. Sensors,2025,25(3):769.

[12] WANG Kai, LU Feng, WANG Jishuo, et al. Lightweight algorithm for early recognition of mine exogenous fire based on YOLO[J]. Journal of China Coal Society, 2025, 50(9):4194−4206.

[13] WU Z S, XUE R, LI H. Real-time video fire detection modified YOLOv5 network model[J]. Fire Tec-hnology,2022,58(4):2377−2403.

[14] Kaiming H, Xiangyu Z, Shaoqing R, et al. Spatial pyramid pooling in deep convolution networks for visual recognition. [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37 (9): 1904-16.

[15] SHI P, LU J, WANG Q, et al. An efficient forest fire detection algorithm using improved YOLOv5[J]. Forests, 2023, 14(12):2440.

[16] HWANG S, PARK J, KIM N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1037-1045.

[17] LI Hao, ZHANG Haiming, NIE Shan, et al. Multimodal recognition method for crop diseases and pests integrating knowledge graph and dynamic gating mechanism[J/OL].Microelectronics and Computer. 1-10[2026-05-15]

[18] TANG Linfeng, XIANGXinyu, ZHANG Hao, et al. DIV Fusion: darkness-free infrared and visible image fusion[J]. Information Fus-ion,2023,91:477-493.

[19] LI Yang, ZHANG Jie, LI Qiangqiang. Infrared Image Object Detection Algorithm Based on Improved YOLOv5[J].Infrared Technology, 025,47(08):1009-1017.

[20] Jia X, Wei Y, Zhao H, et al. LLVIP: A visible-infrared p-aired dataset for low-light vision[J]. Pattern Recognition, 2022, 123:108398.

[21] Li Jinke, Wu Yue, Yang Xiaoyan. A high performance thermal Infrared object detection framework with centralized regulation[R/OL]. arXiv, 2025.

[22] WANG B, CHEN Y, LI Z, et al. Feature level fusion detection algorithm of visible and infrared images based on improved YOLOv5[J]. Optical Journal, 2023, 43(12): 1215002.

[23] SMITH J, DOE A. Multispectral RGB-LWIR Fusion with YOLO for Autonomous Object Detection in Low-Cost Mobile Robotics under Variable Lighting Conditions[J]. Journal of Student Scientific Research, 2024, 15(2): 45-58.

[24] Hwang S, Han D, Jeon M. Multispectral detection transformer with infrared-centric sensor fusion[R/OL].2025.arXiv:2505.15137.