Scientific Reports

https://doi.org/10.1038/s41598-026-47077-6

Article in Press

Research on identification method and application of unsafe behavior of coal mine personnel

Received: 17 July 2025

Accepted: 29 March 2026

Published online: 03 April 2026

Cite this article as: Juan L., Zhu Q., Jiang D. *et al.* Research on identification method and application of unsafe behavior of coal mine personnel. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-47077-6

Liang Juan, Quanjie Zhu, Dongsheng Jiang, Yan Liu, Shaojie Chen & Yingnan Hao

Article

Research on identification method and application of unsafe behavior of coal mine personnel

Liang Juan[1], ZHU Quanjie[2,*], JIANG Dongsheng[2], LIU Yan[3], CHEN Shaojie[2], HAO Yingnan[2]

[1] School of Humanities and Social Science, Institute of Disaster Prevention, Langfang Hebei 065201, China
[2] School of Safety Emergency Technology and Management, North China Institute of Science and Technology, Langfang Hebei 065201, China
[3] Shiyan Tobacco Company of Hubei Province, Hubei Shiyan 442099, China
* Correspondence: zhqj2016@ncist.edu.cn

Abstract: Accurate analysis and identification of underground monitoring videos in coal mines can prevent safety accidents caused by unsafe behaviors of underground personnel and protect their safety. In light of this context, this study enhances the traditional YOLOv11 algorithm for target detection and proposes a fast and effective method for identifying unsafe behaviors among underground personnel in the complex environment of coal mines. Firstly, a statistical analysis of the most common types of unsafe behaviors in current underground coal mines is conducted, exploring the classification of miners' unsafe behaviors into item-type, action-type, and area-type categories. Secondly, based on the characteristics of these unsafe behaviors, we propose dataset augmentation and denoising preprocessing techniques to enhance fine-grained feature extraction. Simultaneously, we introduce the parameter-free SimAM to improve the saliency mapping of miners' behaviors. Finally, we optimize the YOLOv11 algorithm by incorporating a function enhancement module and the K-means++ anchor frame, and we propose a dual-model recognition method for target detection that integrates the YOLOv11 algorithm with the YOLOv11-Pose algorithm. To validate the performance of our non-standard miners' behavior recognition method, we test it on a self-constructed dataset. The research results demonstrate that our method can quickly and effectively recognize unsafe behaviors among underground personnel. Compared to traditional methods, our approach significantly improves recognition accuracy on both the self-constructed dataset and the public dataset, achieving a mean Average Precision (mAP) of 95.7%, an accuracy rate of 95.3%, and a recall rate of 95.1%. These findings are significant for preventing underground safety accidents.

Keywords: mine safety; unsafe behavior; YOLOv11; denoising; object detection

1. Introduction

Coal miners face the challenge of executing complex tasks in confined spaces, which not only require a high level of professional knowledge and skill but also demand constant vigilance to avoid unsafe behaviors[1-2]. This vigilance is crucial for ensuring the safety of both themselves and their colleagues. However, the unique conditions of the coal mine environment, along with the significant pressures of the job, can lead to occasional unsafe behaviors that may result in accidents, serious injuries, or even fatalities. At present, the identification of unsafe behaviors among personnel in coal mines heavily relies on manual inspections and the analysis of surveillance footage. These approaches pose several challenges, including inefficiency, significant subjectivity, low accuracy, high costs, and difficulties in covering all operational areas[3-5]. Therefore, developing effective strategies to identify and prevent unsafe behaviors among underground workers, as well as improving the overall safety standards in coal mining operations, is an urgent issue that needs to be addressed.

Underground personnel's unsafe behaviour recognition is a multidisciplinary research field. With the continuous progress of machine vision and computer technology, the research on the use of machine vision technology to identify the unsafe behaviour of underground personnel has made some progress. The traditional methods of underground miners' behaviour recognition have been analyzed and some results have been achieved. Yao[2] constructed a model for predicting and evaluating miners' unsafe behaviour based on application, management and

behavioural errors; Li et al[3] proposed Yolov5 to identify miners' behaviour and underground equipment; Liang[4] et al improved YOLOv5 by integrating spatio-temporal feature information to workers' operating behaviour monitoring. Traditional miner behaviour recognition methods rely on manual feature extraction, with low recognition efficiency and large application limitations. Based on the above considerations, deep learning has been gradually applied to miners' behaviour recognition. Ren et al[7] applied transfer learning and residual networks to identify unsafe behaviours of miners such as falling and throwing; Khan et al[8] proposed the Inception-v3 network to monitor the severity of accidents of pedestrians; Kolar et al[9] applied convolutional neural networks for safety guardrail monitoring. The above methods give us some new insights into behaviour recognition, but they do not achieve the expected results in terms of generalisation and scale invariance of behaviour recognition in complex environments.

In recent years, unsafe behavior recognition methods have made significant progress in theory and application, and the specific progress covers several research directions based on machine learning, deep learning, sensor data fusion, and video analysis. Among the traditional machine learning methods, Support Vector Machine (SVM), Random Forest (RF), and other methods are widely used in early unsafe behavior recognition, which can classify and predict the behaviors of warehouse personnel to a certain extent, but suffer from the shortcomings of limited recognition accuracy in complex environments. In recent years, deep learning methods, especially Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs), have achieved excellent performance in video and time-series data analysis. For example, CNN-based image classification techniques were used in literature[1-5] to identify unsafe behaviors of warehouse personnel, and the results showed significant advantages in coping with problems such as occlusion and light changes. Literature[6-10] explored temporal models incorporating LSTM for identifying unsafe behaviors in continuous actions, which achieved better results in the recognition of dynamic behaviors. Other literatures[11-15] focus on the fusion of sensor data, which can further improve the accuracy and robustness of recognition by combining data from wearable devices (e.g., accelerometers and gyroscopes) with video data. The combination of multimodal data helps to solve the problem of a single data source in special cases (e.g., limited camera view angle or signal loss). In addition, the literature[16-20] presents real-time monitoring systems that leverage deep learning techniques for the detection and alarming of unsafe behaviors. These advanced systems enhance the accuracy of behavioral identification while simultaneously possessing the capability to respond in real time. Consequently, they offer significant technical support for safety management initiatives, thereby improving overall safety outcomes.

Recent trends in research indicate a significant transition in the recognition of unsafe behaviors, moving from the reliance on a singular data source and static analytical methods to the integration of multimodal fusion techniques and real-time intelligent warning systems. Future investigations are anticipated to concentrate on preserving high accuracy in unsafe behavior recognition within complex environments while concurrently reducing the system's dependence on computational resources. The exploration of these research avenues is expected to enhance safety management protocols in warehouse settings and provide valuable insights applicable to analogous industrial scenarios.

The author constructs a comprehensive dataset addressing the unsafe behaviors exhibited by underground personnel through a meticulous examination of the underlying causes and potential consequences of such behaviors. This dataset is generated from a systematic classification of unsafe behaviors observed among workers in coal mining environments. Following the preprocessing of the data, the author develops an identification model for unsafe behaviors by integrating an enhanced version of the YOLOv11 network with the YOLOv11-Pose network. This model seeks to refine the capabilities for recognizing and providing early warnings regarding unsafe personnel behaviors within the challenging conditions of underground environments, thereby enhancing both the accuracy of recognition and the timeliness of responses.

2. Definition and classification of unsafe acts of personnel

Unsafe behaviors in the workplace are defined as deviations from normative conduct exhibited by workers. In accordance with the Chinese national standard Classification Standards for Occupational Injuries and Accidents in Enterprises, unsafe behavior is characterized as human errors that have the potential to precipitate safety incidents. More broadly, these behaviors encompass any actions by employees that could directly or indirectly result in safety accidents. This includes violations of established protocols, unsafe actions that contribute to accidents, and noncompliance with safety procedures. Such behaviors pose significant risks not only to the individuals involved but also to the overall safety culture within

the organization. Understanding and mitigating these behaviors is essential for enhancing workplace safety and preventing incidents.

2.1. Definition and causes of unsafe behaviors

Unsafe acts are typically characterized as human errors that transpire during the production process, potentially resulting in accidents. These behaviors can be either intentional or unintentional and often lead directly to safety incidents. A clear distinction between various types of lapses and violations is crucial when defining unsafe behaviors. For example, a lapse may manifest as an unintentional act stemming from inattention or poor judgment, while a violation generally occurs as a result of a willful disregard for established safety protocols or procedures. Furthermore, it is essential to consider the potential level of harm associated with unsafe acts, which can vary from minor oversights to significant violations that have the capacity to induce major accidents.

In the context of coal mining, an unsafe act by an underground worker can be defined as any behavior that occurs within the mine operating environment and subsequently increases the risk of accidents or may result in injury. This definition encompasses not only overt violations, such as the failure to adhere to established safety protocols or the willful disregard of safety warnings, but also subtler, less obvious actions that can be equally hazardous, including lapses in attention or decision-making induced by fatigue. Within the specific context of coal mining operations, employees often encounter a variety of stressors and challenges that can significantly influence their behavioral choices and cognitive processes regarding safety.

The unsafe behaviors exhibited by underground personnel represent a multifaceted phenomenon influenced by a variety of factors, which span individual, organizational, environmental, and behavioral psychological dimensions. These unsafe behaviors are recognized as a principal contributor to mining accidents. To facilitate a more accurate identification and intervention regarding the dangerous behaviors of underground workers, it is imperative to consider both their work habits and the inherent specificity and diversity of underground operations. Accordingly, we categorize the unsafe behaviors of underground personnel into three distinct types: object-type, action-type, and area-type, as delineated in Table 1. Each of these categories encompasses specific manifestations of unsafe behavior relevant to personnel operating within underground environments.

Table 1. Example of unsafe behavior classification of underground personnel

Classification	Examples of unsafe behaviors	Consequences of behavior
Object-Type	Not wearing a helmet	There are risks of falling objects and collisions in the underground environment that can significantly increase the likelihood of head injuries.
	Improper storage of objects and tools	It may cause safety incidents such as falls and collisions, or make it difficult to quickly find the tools you need in an emergency.
Action-Type	slip and fall	When working on unstable or slippery surfaces, improper movement can lead to falls and subsequent injuries.
	Riding Belt	The use of informal methods of movement, such as conveyor belts, increases the risk of injury.
	Climbing Fence	Basic safety regulations are violated, increasing the risk of falls and injuries.
	Illegal train surfing	Climbing or hanging from a moving vehicle is highly susceptible to serious injuries such as falls and collisions.
Area-Type	Invasion of hazardous areas	Unauthorized entry into areas marked as dangerous or off-limits may expose you to unknown security risks.
	Unauthorized absence from work	It may lead to work interruptions or lack of security monitoring, increasing security risks.

Unsafe behaviors in underground work environments can be categorized into three primary types: item-type, action-type, and area-based unsafe behaviors. Item-type unsafe behaviors predominantly involve the improper utilization of personal protective equipment and inadequate storage of tools and items. These behaviors are often attributable to carelessness or a lack of safety awareness during daily operations. Although such negligence may seem trivial, it can lead to severe consequences in the inherently high-risk context of underground work. Action-type unsafe behaviors typically arise from a deficit in safety awareness and an egregious neglect of established underground work procedures. These actions frequently demonstrate a

blatant disregard for personal safety and the safety of colleagues, manifested through a series of reckless and potentially hazardous activities. Area-based unsafe behaviors often emerge from a worker's misplaced sense of invulnerability, accompanied by a lack of accountability and disregard for safety advisories. These behaviors are characterized by violations of site-specific safety regulations while engaged in underground tasks. Examples of area-based unsafe behaviors include unauthorized entry into zones designated as hazardous, utilizing unsafe transportation equipment such as conveyor belts, and abandoning assigned workstations without proper authorization. Understanding these categories of unsafe behaviors is crucial for developing effective safety interventions and promoting a culture of safety in underground operations.

2.2. Routine Unsafe Behavior Recognition Methods

In the domain of unsafe behavior recognition, algorithms are distinguished based on their applications and performance metrics. Traditional machine learning algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), decision trees, and random forests, are particularly effective for small datasets and scenarios where manual feature extraction is manageable, owing to their simplicity and strong interpretability. Nonetheless, these algorithms exhibit limitations in addressing complex behavioral patterns, especially in contexts involving multidimensional data. Future research may aim to integrate traditional algorithms with deep learning methodologies, thereby capitalizing on deep learning's robust feature extraction capabilities while preserving the interpretability intrinsic to conventional approaches. A comparative analysis of the advantages and disadvantages of these traditional methods is provided, with the findings detailed in Table 2.

Table 2. Comparison of various recognition algorithms

Algorithm Type	Typical Algorithms	Dominance	Defection
Traditional machine learning algorithms	SVM、KNN、Decision Trees、Random Forests	Algorithms are simple and interpretive	Feature engineering relies on manual labor, which is time-consuming and demanding[22].
Convolutional Neural Networks	YOLO、Faster R-CNN、SSD	Efficient extraction of image features	High-end hardware is required for training, and reasoning is costly[7, 23]
Recurrent Neural Network	LSTM、GRU	Capable of capturing timing information	Hyperparameter sensitivity and slow convergence[10]
Spatio-Temporal Graph Convolutional Networks	ST-GCN、2S-AGCN	Ability to integrate temporal and spatial features	Static adjacency matrices cannot adapt to topology changes[24]
Human Posture Estimation + Behavior Recognition	OpenPose + LSTM	Enables fine-grained behavioral recognition	Heavy multi-target occlusion and low light at night cause attitude estimation to be ineffective[5].
Deep Reinforcement Learning	DQN、DDPG	Adaptive to gradually optimize recognition strategies	Long training time[25]
Hybrid models (multimodal fusion)	CNN + LSTM、OpenPose + spatio-temporal modeling	Fusion of multiple information sources for improved recognition	Excessive complexity of model architecture[26]

The analysis presented in the preceding table indicates that deep learning Convolutional Neural Networks(CNNs) demonstrate significant efficacy in the domains of image feature extraction and behavioral recognition. Advanced algorithms such as You Only Look Once(YOLO) and Faster Region-based Convolutional Neural Networks(R-CNN) facilitate efficient target detection and recognition processes. These methodologies can be effectively implemented within warehouse environments to monitor and identify potentially hazardous behaviors, such as workers failing to wear helmets or deviating from designated walking paths. However, a notable limitation of deep learning techniques is their reliance on extensive datasets for training, as well as substantial computational resources, thereby posing challenges for widespread implementation. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), are effective architectures for the analysis of time-series data due to their inherent ability to model temporal dependencies. This characteristic makes them well-suited for applications in warehouse surveillance, where the identification of unsafe behaviors—such as prolonged occupancy in restricted areas—can be critical for maintaining safety and security. However, the deployment of these networks is not

without challenges; specifically, the issues of gradient vanishing and the high complexity associated with training processes present significant hurdles that must be addressed. Spatio-Temporal Graph Convolutional Networks (ST-GCN) have been identified as highly effective for the skeleton-based recognition of human behavior due to their ability to integrate temporal and spatial information. This approach enables the detection of complex lifting poses through the analysis of spatiotemporal variations in keypoints of the human body. While this methodology significantly enhances the handling of dynamic behaviors, it is associated with substantial computational intensity and necessitates specialized hardware support to implement it effectively. The integration of human pose estimation with behavioral recognition methodologies, such as the combination of OpenPose and Long Short-Term Memory (LSTM) networks, presents significant advantages for fine-grained behavioral detection. This approach leverages skeletal information to mitigate background interference, thereby enhancing the accuracy of behavioral assessments. In the context of warehousing environments, such methodologies can effectively identify specific operational postures of workers, including whether they are utilizing equipment in a manner consistent with proper guidelines and safety regulations. However, the practical application of these techniques encounters notable challenges, particularly regarding pose estimation accuracy, which is often compromised by factors such as occlusion and variable ambient lighting conditions. These factors necessitate ongoing research and development to improve the robustness and reliability of pose estimation in real-world settings. Deep reinforcement learning techniques, such as Deep Q-Networks (DQN) and Deep Deterministic Policy Gradient (DDPG), demonstrate an exceptional capability for the continuous enhancement and optimization of behavior recognition strategies. These approaches exhibit high adaptability, rendering them particularly suitable for applications in warehouse security management, where real-time decision-making is essential. However, a notable challenge associated with these methods is their significant requirement for interaction data, which often leads to extended training durations, particularly within the context of security behavior simulations. Hybrid models, such as those integrating Convolutional Neural Networks (CNN) with Long Short-Term Memory networks (LSTM) or merging pose estimation with spatio-temporal frameworks, have demonstrated superior performance in the realm of behavioral recognition. These models facilitate a more nuanced detection of unsafe behaviors by effectively synthesizing diverse information modalities. Their application is particularly advantageous in complex environments, such as warehousing, where the integration of visual data with spatial location data can enhance the ability to ascertain whether individuals are adhering to designated pathways.

From an overall perspective, the prevailing trend in the recognition of unsafe behavior is gravitating towards the integration of lightweight algorithms and multimodal information. In practical applications, the synergy between the efficient feature extraction capabilities of deep learning techniques and the inherent interpretability of traditional methodologies is poised to enhance the field significantly. Furthermore, advancing the robustness of models through the fusion of multimodal data represents a critical avenue for future research endeavors in this domain.

3. Data set construction

3.1. Data collection and preliminary processing

The heterogeneity of data sources necessitates the conversion of data into a standardized format to facilitate efficient processing and analysis. This requirement is particularly pertinent in the context of video data, where disparate formats—including AVI, MP4, and MPEG—exhibit varying storage specifications and compression methodologies. The disparate nature of these formats can pose significant challenges for machine learning models that rely on consistent input data. In the case of underground surveillance videos, which are often characterized by substantial file sizes, direct processing can result in excessive consumption of computational resources and diminished processing speeds. Therefore, converting these videos into a uniform format is essential for ensuring both consistency and efficiency during processing. Moreover, video format conversion not only standardizes the data but also allows for the application of more efficient compression techniques. By optimizing storage and enhancing processing capabilities, this conversion process is crucial for managing the demands posed by large-scale video datasets. Thus, the implementation of uniform data formats is integral to advancing the capabilities of machine learning applications in the domain of underground surveillance.

In the realm of video format conversion tools, the Open Source Computer Vision Library(OpenCV) serves as a prominent open-source resource that encompasses a diverse array of functions for image and video processing, including the capabilities for video reading, writing, and format conversion. Initially developed by Intel in 1999, OpenCV has since

established itself as a foundational library within the fields of computer vision research and practical applications. OpenCV facilitates the conversion of video formats and enables preprocessing tasks such as frame extraction, resizing, and color space conversion. This is achieved by representing images as arrays of pixels, allowing for direct programmatic manipulation and conversion of each pixel element. In the context of this study, OpenCV was employed to adjust the resolution and frame rate of a set of underground surveillance videos collected for analysis. Furthermore, when considering the specific environment of underground monitoring, OpenCV allows for the adaptation of parameter settings that are critical for optimal video processing. The relevant video parameter settings utilized in OpenCV are detailed in Table 3. It is also essential to account for the interval timing during image extraction; excessively long intervals may result in the omission of significant features, whereas overly short intervals may lead to redundant image captures, thereby increasing the overall workload. This careful consideration of parameter settings and extraction intervals underscores the importance of tailored video processing methodologies in enhancing the efficacy of surveillance video analysis[27].

Table 3. OpenCV video parameter settings

Parameter Name	Parameter Setting Algorithm
Resolution Width	cap.set(cv2.CAP_PROP_FRAME_WIDTH, 432)
Resolution Height	cap.set(cv2.CAP_PROP_FRAME_HEIGHT, 368)
Frame Rate	cap.set(cv2.CAP_PROP_FPS, 30)
Contrast	cap.set(cv2.CAP_PROP_CONTRAST, 60)
Saturation	cap.set(cv2.CAP_PROP_SATURATION_SCALE, 60)
Interval Frame Duration	cap.set(cv2.CAP_PROP_FRAME_SKIP, 0.01)

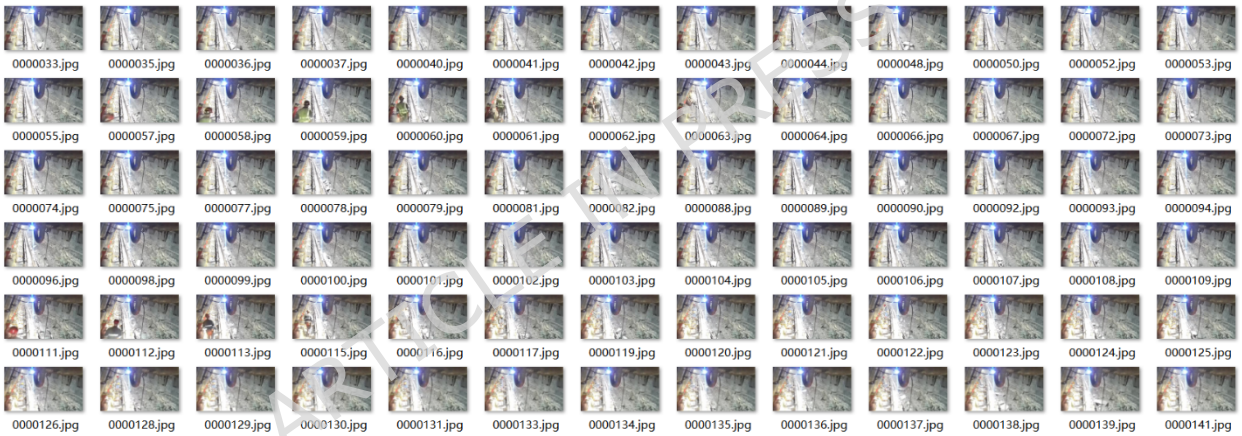


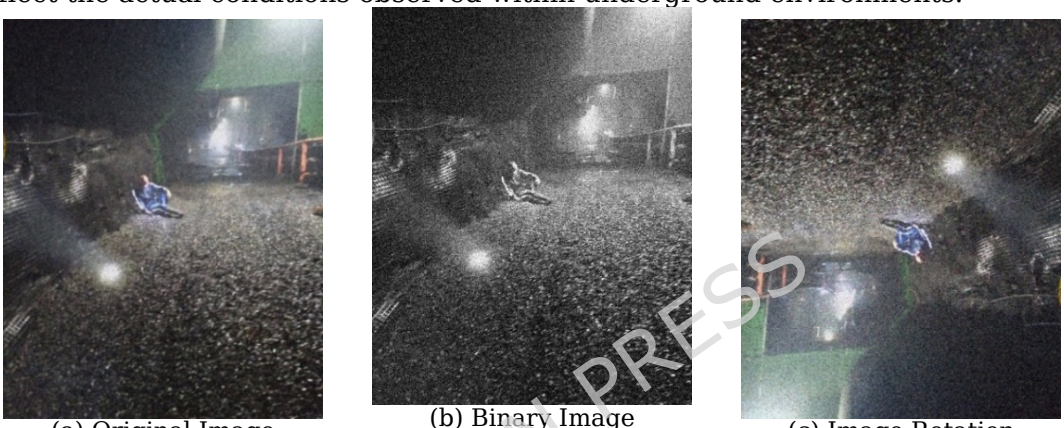
Figure 1. OpenCV format conversion processing video

In the realm of computer vision, the application of OpenCV techniques for the in-depth processing of surveillance video provides significant insights into the video content. The transformation of a continuous video stream into a sequence of images arranged sequentially, as depicted in Figure 1, facilitates a comprehensive analysis of the video material on a frame-by-frame basis. This method not only enhances the precision of the computer's analysis but also results in a substantial reduction in both resolution and frame rate of the derived images. The implications of diminishing resolution and frame rate are manifold, most notably leading to a marked decrease in the data volume of each individual image frame. This reduction effectively compresses the overall size of the video or image file. Such a compression strategy is particularly advantageous for extensive surveillance durations, as it significantly alleviates storage space requirements and augments data processing efficiency, all while preserving essential visual information. Moreover, the OpenCV processing workflow supports efficient data compression without compromising the recognizability of the video content. This preservation of critical visual elements holds considerable significance for the long-term storage and retrieval of video data, underscoring the practical benefits of employing OpenCV for surveillance applications.

3.2. Expansion of datasets

In the development of a dataset aimed at analyzing unsafe behaviors of individuals in underground mining environments, we employed a range of data augmentation techniques to

effectively simulate the intricate and dynamic conditions characteristic of such settings. This process entailed the extraction of relevant information pertaining to hazardous behaviors from an array of data sources, followed by the integration and standardization of this information to ensure the dataset's consistency and usability.To enhance the adaptability and accuracy of the model in actual mining scenarios, we implemented a series of augmentation and generalization operations on both the original underground dataset and a laboratory-simulated dataset. For instance, we applied black-and-white processing (illustrated in Fig. 2(b)) to the original images, altering light intensity to mimic the shift from color to monochrome as observed in varying lighting conditions encountered underground. This adjustment is intended to bolster the model's capacity to recognize unsafe behaviors across diverse lighting environments.Furthermore, we employed a data flipping technique, resulting in a 180-degree rotation of the images (depicted in Fig. 2(c)), to train the model to identify unsafe behaviors from multiple perspectives, thereby simulating various viewpoints. The utilization of these data augmentation techniques not only enhances the dataset's diversity but also increases the model's adaptability to the complexities inherent in real-world mining situations. Ultimately, these methodologies provide a more robust training foundation for the model, allowing it to better reflect the actual conditions observed within underground environments.



(a) Original Image (b) Binary Image Processing (c) Image Rotation

Figure 2. Example of mine dataset augmentation

Following the data augmentation process, the dataset undergoes a rigorous screening procedure to eliminate samples of low quality or those inconsistent with the research objectives. This step is essential to ensure the overall quality and reliability of the dataset, which is critical for subsequent stages of image denoising and dataset labeling. Ultimately, a comprehensive set of custom datasets characterized by diverse scene attributes is constructed through meticulous screening, integration, and augmentation procedures. As detailed in Table 4, the cumulative number of images from downhole and laboratory samples reached 31,000, comprising 9,700 images categorized as items, 15,800 categorized as actions, and 5,500 categorized as areas. The dataset encompasses a wide array of underground work environments, including confined spaces, varied working surfaces, and different equipment zones. Furthermore, it incorporates numerous instances of unsafe personnel behaviors across various scenarios. This robust dataset provides a solid foundational resource for subsequent image processing and model training endeavors.

Table 4. Sample size of the homemade personnel unsafe behavior dataset

Classification	Unsafe behaviors	Number of downhole samples	Number of laboratory samples	Total number of samples
Object-Type	Not wearing a helmet	3200	2900	6100
	Improper storage of objects and tools	2100	1500	3600
Action-Type	slip and fall	1500	1600	3100
	Riding belt	3100	2500	5600
	Climbing fence	2600	2100	4700
	Illegal train surfing	1300	1100	2400
Area-Type	Invasion of hazardous areas	1500	1300	2800
	Unauthorized absence from work	1500	1200	2700
Total		16800	14200	31000

3.3. Denoising the dataset

In the context of underground coal mining environments, the transmission of video images from surveillance systems to operational terminals is frequently subjected to a distinct set of noise challenges. These challenges primarily stem from the unique working conditions and environmental factors prevalent in subterranean mining sites. If left unaddressed, such noise issues can significantly compromise the accuracy and stability of models designed for recognizing unsafe behaviors among personnel. The various types of noise encountered in these environments include, but are not limited to, low-light noise, interference from dust particles, and blurring induced by equipment vibrations. Low illumination noise typically arises from inadequate lighting conditions common in underground settings, leading to granular noise that adversely affects image clarity. Dust particle interference results from the pervasive underground dust, which causes image loss and introduces random, non-uniform distortions. Furthermore, blurring due to equipment vibrations occurs when surveillance cameras are mounted on machinery that experiences oscillations, resulting in dynamic blurring and geometric distortion of captured images. Consequently, it is imperative to implement effective denoising techniques for video images to enhance clarity and contrast. Such improvements are essential for bolstering the efficacy of recognition models designed to detect unsafe behaviors among personnel in underground coal mining operations.

Image denoising algorithms can be categorized into two primary classes: traditional methods and deep learning methods. Traditional approaches primarily rely on the statistical characteristics or a priori knowledge inherent to the image. Examples of such methods include mean filtering, median filtering, and wavelet transforms. In contrast, deep learning methods leverage neural networks to learn high-level features from images, thereby achieving superior denoising performance. Notable architectures within this framework include self-encoders and convolutional neural networks. In this context, after a comprehensive evaluation and consideration of the operational environment, the image denoising algorithm based on the Deep Learning Convolutional Neural Network(DnCNN) has been selected. DnCNN is an advanced deep learning framework specifically designed for denoising images and enhancing sound signals. Its efficacy in removing random noise from seismic data and its applications in sound signal enhancement have been well-documented in the literature. The fundamental principle of the DnCNN algorithm is rooted in deep convolutional neural networks, which are capable of effectively discerning and eliminating noise from images by learning the intricate relationships between pairs of noisy and noise-free images. Unlike traditional denoising techniques that directly output the denoised image, DnCNN employs a residual learning strategy. Specifically, it first predicts the residuals—defined as the noise components—between the noisy and clean images, subsequently subtracting these residuals from the noisy input to yield the denoised output. This approach benefits from the typically simpler statistical properties of the residuals, facilitating a more efficient training process for the network model.

The architecture of DnCNN comprises a sequence of multiple convolutional layers, accompanied by batch normalization and activation function layers. The convolutional layers play a crucial role in feature extraction from images, while the batch normalization layers effectively address the internal covariate shift, thereby enhancing both the speed and stability of the training process. Additionally, the incorporation of the ReLU activation function introduces non-linearities, enabling the network to capture more intricate data patterns. As illustrated in Figure 3, the DnCNN network contains a total of seventeen layers. The initial layer consists of 64 convolutional kernels, each with a size of 3×3, coupled with the ReLU activation function. In the subsequent layers, from the second to the sixteenth, there is an increased number of batch normalization layers relative to the first layer, which significantly contributes to the efficiency and robustness of the model training. Finally, the concluding layer comprises a convolutional layer whose primary function is to output the residual image, thereby enhancing the overall image denoising performance of the network.

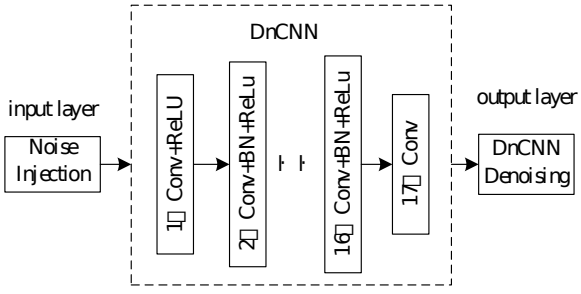


Figure 3. DnCNN network structure diagram

To train a DnCNN network, the difference between the network's predicted residual maps and the true residual maps is usually calculated using the Mean Square Error (MSE) as a loss function. Given a set of training samples, the loss function can be expressed as:

$$L(q) = \frac{1}{2N} \sum_{i=1}^{N} R(y_i; q) - (y_i - x_i)^2 \quad (1)$$

Where N is the number of training samples; y_i is the noisy image; x_i is the corresponding clean image; $R(y_i; q)$ is the residuals predicted by the network, and q is the network parameters.

The implementation of DnCNN facilitates the optimization of data processing in various applications. Within underground environments, real-time video surveillance is paramount, necessitating the deployment of fast and precise denoising algorithms. Despite the complexity inherent in the deep network architecture of DnCNN, its superior denoising capabilities and efficacy in handling real-time data render it particularly suitable for the processing of surveillance video in subterranean settings.

To evaluate the efficacy of the DnCNN image denoising algorithm on surveillance images, a self-constructed dataset focusing on insecure behaviors was utilized for model training. Through an iterative prediction process comprising multiple rounds, the model weights were continuously adjusted and optimized, leading to a gradual enhancement in the quality of the final output images, aligning more closely with the characteristics of the original noise-free images. Figure 4 illustrates both the unprocessed image and the image processed by the DnCNN algorithm. A comparative analysis reveals a significant improvement in clarity and detail in the DnCNN-processed images. The unprocessed imagery, influenced by the challenging environmental conditions prevailing in underground settings, exhibited various forms of noise, such as blurred contours, grainy backgrounds, and shadows resulting from uneven illumination, which obscured critical details and features in the images. Conversely, the images processed by the DnCNN algorithm displayed sharper edges, smoother textures, and more uniform lighting conditions, thereby facilitating the identification of unsafe behaviors, such as individuals not wearing helmets and engaging in illegal operations. This highlights the potential of the DnCNN algorithm in enhancing the interpretability of surveillance footage, particularly in complex environments.

(a) Noise Injection (b) DnCNN Denoising

Figure 4. Effect after DnCNN denoising

3.4. Labeling of data sets

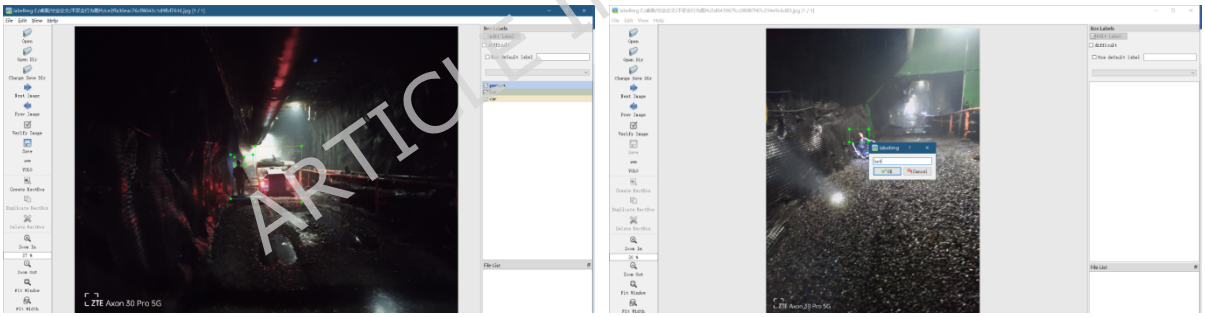
In the context of varying detection targets and network models, it is imperative to construct distinct datasets tailored to specific objectives, such as those encompassing miners, safety helmets, and human skeletal data. The creation of these target recognition datasets necessitates the utilization of authentic underground personnel imagery to ensure the validity and reliability of the experimental outcomes. The experiment began by compiling a dataset of 300 images. To expand the dataset, contrast was increased, brightness was reduced, and images were flipped, resulting in a total of 780 images. These were then divided into a test set of 163 images, a training set of 546 images, and a validation set of 71 images.

Dataset labeling serves as a fundamental component in the domain of machine learning. Specifically, in the development of a machine vision-based dataset aimed at recognizing unsafe behaviors of personnel in underground environments, the significance of precise dataset

annotation cannot be overstated. The primary objective of this annotation process is to accurately identify and label the various unsafe behaviors within the dataset, thereby furnishing the machine learning model with high-quality, labeled training data. For a dataset focused on personnel unsafe behaviors, the explicit labeling of each behavior category—such as object-related, action-related, or area-specific behaviors—exerts a direct influence on both the accuracy and the efficiency of the model's ability to discern unsafe behaviors. Consequently, this annotation process represents a critical factor in achieving high-quality outcomes in machine vision applications, ultimately enhancing the model's performance in real-world scenarios.

In the domain of machine vision, dataset annotation tools are pivotal for facilitating effective and precise data annotation. One notable open-source tool that has gained prominence in this field is LabelImg. This tool is esteemed for its user-friendly interface combined with robust functionalities. LabelImg accommodates a diverse array of image formats, including widely utilized types such as JPEG and PNG. Additionally, it offers the capability to export various annotation file formats, including XML (Pascal VOC format) and TXT (YOLO format), thereby catering to the requirements of different machine learning frameworks. The user interface of LabelImg is characterized by its intuitive design, allowing users to effortlessly select regions of interest within images and assign appropriate labels through straightforward mouse interactions. This design not only mitigates the learning curve associated with data annotation but also significantly enhances labeling efficiency. Furthermore, LabelImg incorporates keyboard shortcut functionalities, which serve to expedite the annotation process, particularly when engaged with large-scale datasets. Overall, LabelImg represents a valuable asset in the annotation of datasets for machine vision applications.

Prior to initiating the labeling process, the computer environment was systematically configured to ensure the optimal functionality of the LabelImg tool. During the data annotation phase, each instance of unsafe behavior depicted in the images was meticulously marked, allowing for precise correspondence between the labels and the respective objects or behaviors. Upon completing the configuration of the annotation tool, the command 'LabelImg' was entered to launch the interface, as illustrated in Figure 5. The central region of the interface is designated as the annotation area. Given that the dataset format requirements stipulated by the model are specialized, the TXT markup file conforming to the YOLO format was utilized for the annotation format selection.



(a) Equipment labelling (b) Personnel labelling

Figure 5. Example of dataset labelling

The initial step involves importing the relevant image for annotation via the 'Open' function. Subsequently, as illustrated in Figure 5, the image undergoes a systematic labeling and classification process. This process categorizes the images based on various types of unsafe behaviors, such as designating 'no-hat' for instances of helmet non-use, and 'hat' for situations where a helmet is being worn. Other forms of unsafe behaviors are similarly annotated. This methodology contributes to the development of a robust and precise dataset that enhances the identification and characterization of unsafe behaviors within the studied context.

4. Modelling improvement

The Underground Personnel Unsafe Behaviour Recognition Model(UBIRM) represents a significant advancement in the monitoring and prevention of incidents involving personnel operating in underground environments. This technological framework employs advanced video analysis techniques to assess various aspects of personnel behaviour, including posture, position, and movement. By doing so, it effectively identifies unsafe practices, which may include non-compliance with safety protocols such as the absence of personal protective equipment (e.g., helmets), failure to utilize safety restraints (e.g., seatbelts), or unauthorized access to restricted areas. The architecture of the UBIRM comprises two primary components:

the YOLOv11 target detection model and the YOLOv11-pose behavioural detection model. The YOLOv11 algorithm is heralded as a prominent advancement in deep learning and computer vision, characterized by its high efficiency and accuracy in real-time object detection. Concurrently, the YOLOv11-pose model leverages deep learning methodologies to facilitate two-dimensional pose estimation for multiple individuals, thus addressing the complexities associated with single-person pose recognition and extending its applicability to recognize multiple persons in intricate scenarios. Through targeted enhancements of both models and their subsequent fusion, the UBIRM achieves precise and timely identification of unsafe behaviours across three critical categories: object, action, and geographic area. This multifaceted approach not only amplifies the effectiveness of safety warnings but also contributes to the overarching objective of mitigating risks and enhancing the safety of underground personnel.

4.1. YOLOV11

YOLOv11 represents a significant advancement in the field of target detection algorithms, characterized by enhancements in both detection accuracy and inference speed. This is achieved through an improved network architecture, the integration of an attention mechanism, and a lightweight design approach. In comparison to its predecessor, YOLOv11 exhibits superior performance in the detection of small targets and is adept at operating within complex scenes. Notably, it is capable of facilitating real-time target detection on edge devices, positioning itself as a rapid, efficient, and resource-conserving solution for target detection tasks. The architecture of this network is illustrated in Figure 6.

This study employs a dual-model collaborative recognition framework, based on YOLOv11 and YOLOv11-pose, to achieve comprehensive identification and early warning of multiple types of safety hazards in coal mine production scenarios. The YOLOv11 model is first employed for object detection tasks, such as identifying workers and key equipment. Based on these detection results, and by combining spatial relationships and attribute information between objects and personnel, object-related hazards, such as failure to wear safety helmets, can be identified. Building on this, the YOLOv11-pose model is introduced to detect and model human keypoints (skeletal points). By analyzing human posture features and their temporal variations, behavior recognition is achieved, enabling the identification of action-related hazards, such as falls and unsafe behaviors like riding on moving machinery.

For region-related safety risks, a collaborative mechanism is constructed that integrates both models. First, YOLOv11 is used to detect mechanical equipment in coal mine operations, and hazardous regions are dynamically delineated based on the spatial locations of the detected equipment. Subsequently, YOLOv11-pose is utilized to analyze personnel behavior and position in real time. When personnel are detected entering hazardous areas or performing unsafe actions, the system can promptly trigger warning alerts. Additionally, for critical equipment that requires dedicated supervision, continuous monitoring of personnel presence is conducted to automatically detect and alert against absenteeism.

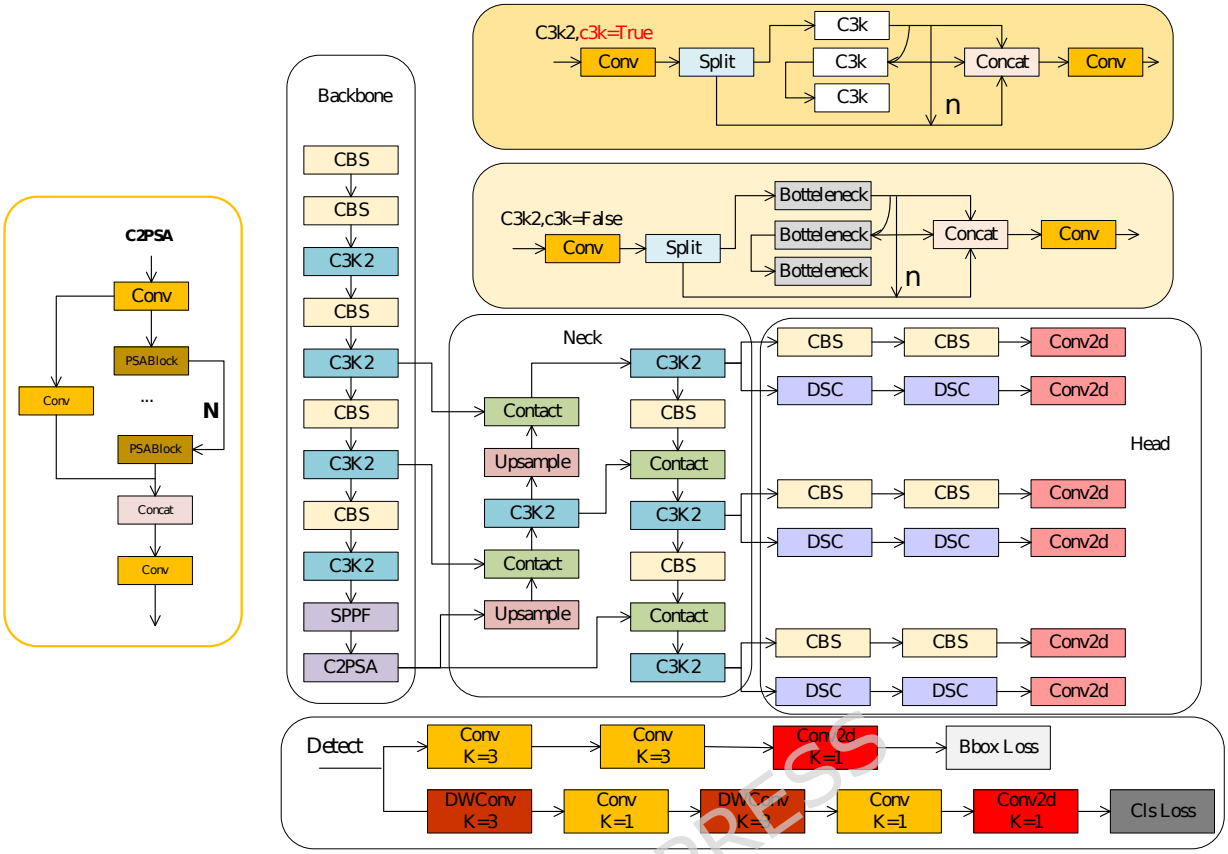


Figure 6. YOLOv11 network structure diagram

4.2. Algorithmic improvements

4.2.1 Feature Enhance Module

The inherent complexity and variability of the downhole environment present significant challenges for image recognition tasks. While straightforward interferences, such as variations in lighting conditions and dust contamination, can often be mitigated through dataset preprocessing techniques, more complex background interferences—such as occlusion and long-distance targets—pose ongoing difficulties. To address these challenges, the recognition capabilities in critical areas or within intricate backgrounds have been enhanced through the integration of the Feature Enhancement Module(FEM) into the YOLOv11 model[18]. The architectural framework of the Feature Enhancement Module is depicted in Figure 7. The traditional YOLOv11 algorithm employs a series of 'concat' operations to fuse features from multiple layers, thereby maximizing information utilization. Consequently, the Feature Enhancement Module is strategically positioned subsequent to the concatenation operations. This module is designed to filter out features that contain extraneous or invalid information, thereby indirectly augmenting the utilization of pertinent information. As a result, this enhancement leads to improved predictive performance of the algorithm.

FEM

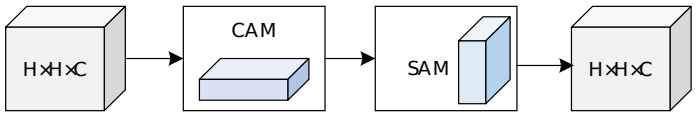


Figure 7. Function Enhancement Module Structure

The complete functional enhancement module comprises a channel activation module and a spatial attention module that work in tandem. The channel activation module's primary function is to eliminate channels that contain a significant amount of invalid information, thereby indirectly increasing the proportion of valid information while reducing redundancy. The spatial attention module complements the channel activation module by on the most meaningful features, specifically local features. This module employs max-pooling and average-pooling

techniques for feature maps of size H×H×C, subsequently combining the two resulting feature maps. It then applies convolution and nonlinear activation to derive the weight coefficients for each component. Finally, these weight coefficients are multiplied by the input feature image to produce a new feature map after scaling. The calculation formula is:

$$Y = SAM(CAM(X)) \tag{2}$$

Where X represents the input feature map and Y denotes the enhanced output feature map. CAM refers to the computation of the Channel Activation Module, which emphasizes significant features while diminishing less important ones by assigning weights to each channel of the input feature map. SAM signifies the computation of the Spatial Attention Module, which focuses on important spatial regions.

4.2.2 K-means++

In the investigation of detecting and identifying unsafe behaviors of individuals in underground coal mines utilizing the YOLOv11 algorithm, the optimization of anchor boxes is a critical component for enhancing the detection accuracy of the model. An anchor box is defined as a set of predefined bounding boxes established based on prior knowledge, designed to facilitate the prediction of actual object locations within an image. In the specific context of underground coal mines, various factors—including the posture of personnel, the distance of the subjects from the camera, and the angular orientation of the camera—can significantly influence detection efficacy. Consequently, it is imperative to appropriately target and optimize anchor boxes to align with the unique characteristics of objects in this challenging environment.

The K-means++ algorithm represents an optimized version of the K-means clustering method, enhancing the selection process for initial centroids, thereby increasing the diversity of initialization. This improvement effectively reduces the likelihood of the algorithm converging to a local optimum, allowing for better adaptation to the varying shapes and sizes of objects within a given dataset. Consequently, the K-means++ algorithm enhances both the efficiency and accuracy of detection algorithms. The application of this algorithm in optimizing target anchor frames encompasses several critical steps: data preparation, initialization of centroids, clustering iterations, and anchor frame updates.

(1)Data Preparation: Extract the bounding box dimensions of all target objects from the underground coal mine dataset, and normalize these dimensions to eliminate the influence of image size. After normalization, all samples are mapped to the range of [0, 1], ensuring the comparability of data across different scales, and finally forming a normalized sample dataset.

(2)Initialization of Centroids: The K-means++ method commences by randomly selecting one bounding box as the first centroid. Subsequently, additional centroids are selected sequentially based on a probability distribution that is squared relative to the distance from each data point to the existing centroids.

(3)Clustering Iteration: In this phase, the Euclidean distance metric is employed to assign each bounding box to the nearest centroid. The position of each centroid is then recalibrated to reflect the mean position of all data points allocated to the corresponding cluster. This iterative process continues until convergence is achieved, defined as either the change in centroid positions falling below a predetermined threshold or a specified number of iterations being completed.

(4)Anchor Frame Update: After the clustering converges, KKK cluster centers are obtained, which correspond to the anchor box sizes in the normalized space. The clustering results are first de-normalized and mapped back to the model input scale. Finally, the optimized anchor box set is obtained and applied as the new optimized anchors in the YOLOv11 model. The optimization of target anchor frames through the K-means++ algorithm not only enhances the model's adaptability and accuracy in detecting unsafe behaviors within underground coal mines but also significantly improves the efficiency of both training and inference processes.

4.3. Construction and training of recognition model

The fusion modeling process articulated in this paper not only automates the safety monitoring of underground operations but also augments the cognitive capabilities of the system through continuous learning and optimization. Such advancements are critical for ensuring the safety of miners and for sustaining operational productivity within the mining sector.

4.3.1 Construction of the model

YOLOv11 represents a cutting-edge advancement in real-time object detection, effectively amalgamating the processes of detection and classification into a unified regression framework. The development of this model necessitates the selection of an appropriate deep learning

framework, such as PyTorch, along with the design of a suitable network architecture. The architecture of YOLOv11 is characterized by a series of convolutional layers, pooling layers, and activation functions meticulously designed to extract relevant features from images and to produce corresponding bounding boxes. In the course of model construction, YOLOv11 incorporates innovative activation functions, feature pyramid networks(FPNs), and sophisticated optimization strategies for anchor boxes, all aimed at enhancing both detection accuracy and computational efficiency. The methodology and implementation of the coal mine personnel unsafe behavior recognition model are outlined in the accompanying flowchart, as shown in Figure 8. This comprehensive approach underscores the model's potential efficacy in complex real-world applications.

The design of network architecture is a fundamental aspect of constructing the YOLOv11 model. YOLOv11 utilizes a modified convolutional neural network(CNN) architecture that integrates deeply separable convolution with feature pyramid networks(FPNs) to facilitate efficient feature extraction across varying scales. The model's architecture typically encompasses multiple convolutional layers, a batch normalization layer, and an activation function, such as Leaky ReLU, which collectively enhance the efficiency and accuracy of feature extraction. Furthermore, the incorporation of residual connectivity enables the model to effectively train deeper networks, thus mitigating the issues associated with gradient vanishing and ultimately enhancing the overall performance of object detection.

In the context of model parameter settings, YOLOv11 is characterized by a set of hyperparameters, including the learning rate, batch size, and anchor frame size. The learning rate is typically subject to dynamic adjustment, facilitating rapid convergence during the initial phases of training, while subsequently being reduced incrementally to promote stabilization in the later stages of the training process. The selection of batch size serves as a critical factor in balancing memory consumption with training speed, with common values set at 16 or 32. In terms of anchor frame size, an analysis of the training data is conducted utilizing the K-means clustering algorithm, aimed at determining the optimal anchor sizes conducive to effective target detection. This approach enhances the model’s effectiveness in identifying objects across a spectrum of sizes. Moreover, the model incorporates an array of data augmentation strategies, including random cropping, rotation, and color transformation. These strategies are intended to bolster the model's robustness and generalization capabilities, thereby ensuring efficient target detection across diverse environmental conditions.

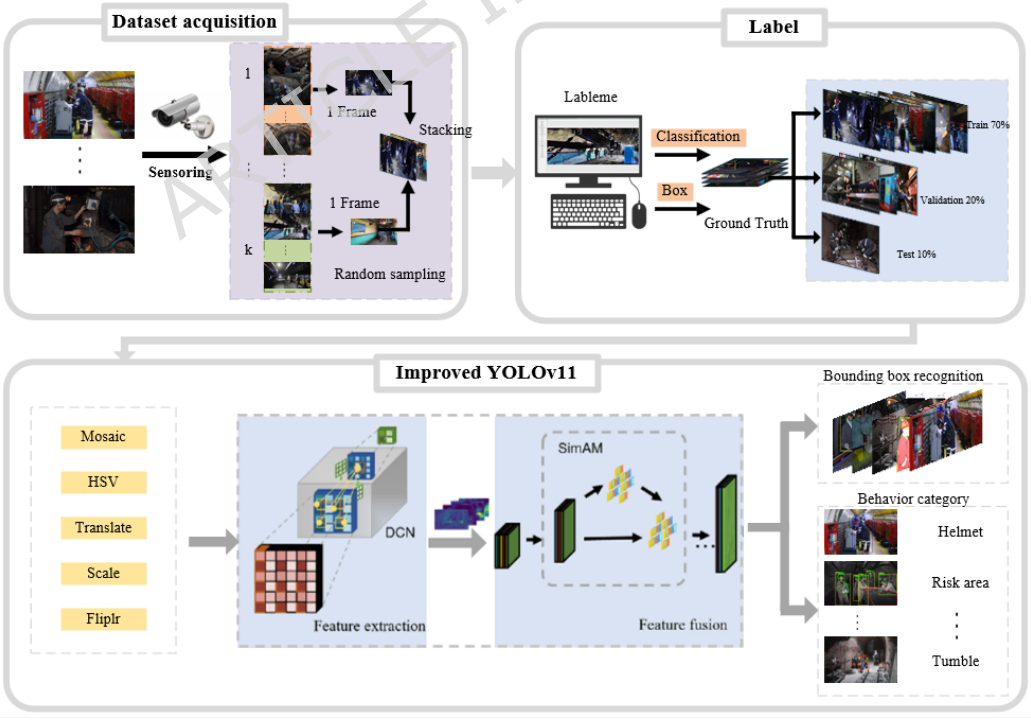


Figure 8. The schematic diagram of miner behavior discrimination

In this case, the parameters in terms of network architecture are set:

(1) Input layer: the input image size is 640 x 640 pixels, and standardised preprocessing is used.

(2) Backbone network: CSPDarknet is used as the feature extraction network to enhance the feature extraction capability.

(3) Feature Pyramid Network: using the FPN structure to combine different levels of features to improve the detection accuracy of small objects.

(4) Detection header: output the detection header including category prediction and bounding box regression, and set 3 anchor frames to adapt to different-sized targets.

The model parameter setting aspects are:

(1) Anchor frames: set up 3 anchor frames with dimensions (10, 13), (16, 30), (33, 23) based on the dataset statistics.

(2) Loss function: YOLO loss function is used, which consists of localisation loss, classification loss and confidence loss.

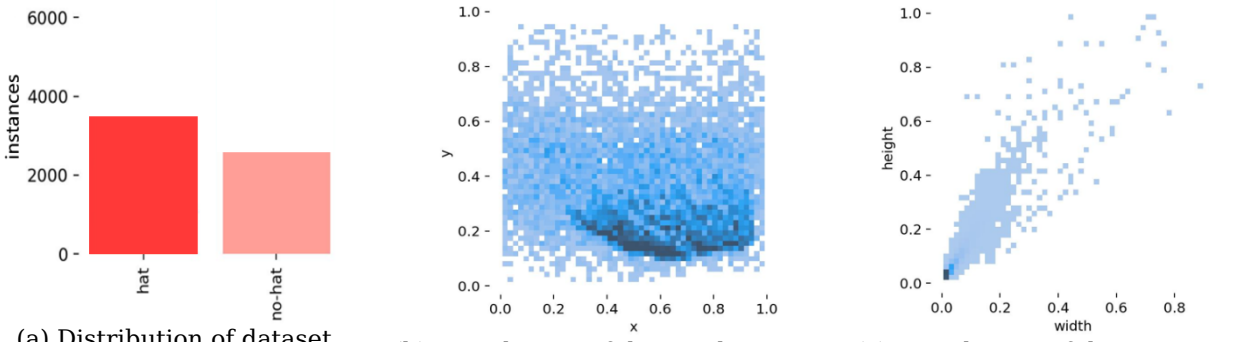
(3) Optimiser: The Adam optimiser was used with the initial learning rate set to 0.001.

4.3.2 Training model

The training process for the YOLOv11 model commences with the meticulous preparation of a well-annotated dataset, which must encompass images representing multiple categories along with their respective labels. Typically, an annotation tool, such as LabelImg, is employed to generate a label file that adheres to the YOLO format. This format includes the coordinates of the bounding boxes and the corresponding category information. In the context of training, the YOLOv11 model leverages transfer learning, often initiating with pretrained weights to facilitate faster convergence and enhance predictive accuracy. Central to the model training is the application of a cross-entropy loss function, which serves to assess the effectiveness of classification, in conjunction with a smoothed L1 loss function, primarily utilized to optimize the regression of bounding boxes. This dual approach is instrumental in ensuring robust performance in object detection tasks.

During the training phase, the implementation of data enhancement techniques is essential for improving the generalization capabilities of the model. Common techniques employed include random cropping, rotation, scaling, and color transformation. These strategies are crucial in enabling the model to maintain robust recognition performance across varied environmental conditions. The training process typically utilizes optimization algorithms such as Stochastic Gradient Descent(SGD) or the Adam optimizer, with the learning rate being systematically adjusted across multiple iterations to facilitate improved convergence. Following the training phase, the model's efficacy is assessed using a validation dataset, which serves to evaluate its performance on previously unseen data. Key evaluation metrics, including mean average precision(mAP) and recall, are employed to ensure the model demonstrates satisfactory detection performance. Upon completion of the training, the generated weight files are preserved for deployment in real-world applications. This structured building and training methodology enables YOLOv11 to achieve efficient real-time object detection across a range of application scenarios. In terms of dataset preparation, the process is divided into two primary components: dataset partitioning and data enhancement. The dataset is partitioned into 70% for the training set, 20% for validation, and 10% for testing. Regarding data enhancement, techniques applied include random cropping, horizontal flipping, and color dithering, among other operations, all aimed at augmenting the dataset's variability and improving model robustness.

In the context of model training, optimal hyperparameter configurations, such as learning rate and batch size, are identified through techniques such as grid search or Bayesian optimization. The specific training configuration employed in this study includes a batch size of 16, with the number of training rounds set to 50. Each training round comprises approximately 1,000 iterations. To facilitate effective training monitoring, TensorBoard is utilized to track both loss and accuracy metrics throughout the training process. For hyperparameter tuning, a learning rate decay strategy is implemented, whereby the learning rate is halved after every 10 epochs. Additionally, the early stopping criterion is established to halt training when the loss on the validation set exhibits no improvement over 5 consecutive epochs. The multi-task loss function associated with YOLOv11 is employed to optimize both bounding box regression and classification tasks. Furthermore, the Adam optimizer is utilized for the training process, contributing to the overall efficacy of the model performance.



(a) Distribution of dataset categories (b) Distribution of dataset locations (c) Distribution of dataset sizes

Figure 9. Analysis of the dataset

In this study, we conducted 200 rounds of training on the helmet dataset utilizing the enhanced YOLOv11 network model. The resultant detailed informative analysis of the helmet dataset is presented in Figure 9. Specifically, Figure 9(a) illustrates the helmet dataset subsequent to format conversion and data augmentation in the preliminary phase, comprising a total of 6,100 samples, with 3,500 representing worn helmets and 2,600 corresponding to unworn helmets. Figure 9(b) depicts the position coordinates of the center points of the anchor frames, where the horizontal and vertical axes represent these coordinates (X, Y). The intensity of color in this figure indicates the concentration of the center points at specific coordinates; a darker hue signifies a higher concentration. The analysis reveals a relatively uniform distribution of safety helmets within the training dataset, although a notable concern persists with the skewing of some samples towards the lower right corner. Furthermore, Figure 9(c) illustrates the dimensions of the detected objects in the dataset, specifically the width and height expressed as a percentage of the overall size of the targets. The representation indicates that darker shades are associated with smaller volume detected targets, suggesting that these smaller objects constitute a significant percentage of the detection targets. This observation aligns with the practical scenarios of scaling and long-distance object detection.

4.4. Assessment of model validity

The experiments conducted in this study employ Mean Average Precision(mAP) and detection rate as key metrics to assess the performance of the target detection model. mAP serves as an aggregate measure of the Average Precision(AP) across all target classes. The computation of AP integrates both the precision rate(P) and the recall rate(R). In evaluating the model's performance, the experiments utilize frame rate, mAP, and detection rate, with mAP providing a comprehensive representation of the model's average precision across various target classes. Frame rate, defined as the quantity of frames processed per second, is a crucial parameter, with a higher frame rate indicating better performance. mAP is particularly informative as it encapsulates the overall detection efficacy of the model across all categories, while simultaneously accounting for both precision and recall metrics. The evaluation parameters are derived as follows: precision (P) is calculated according to the established formula:

$$P = \frac{TP}{TP + FP} \tag{3}$$

Where P denotes the precision rate; True Positives(TP) represents the number of positive samples (target samples) that are correctly identified as positive; False Positives(FP) indicates the number of negative samples (non-target samples) that are incorrectly identified as positive; and TP + FP is the total number of positive and negative samples detected by the model.

Recall (R) is calculated using the following formula:

$$R = \frac{TP}{TP + FN} \tag{4}$$

Where R denotes the recall rate, TP represents the number of positive samples that are correctly identified as positive, while False Negatives(FN) indicates the number of positive samples that are incorrectly identified as negative.

The average precision (AP) is calculated as follows:

$$AP = \int_0^1 P \times R dR \tag{5}$$

Where P represents precision and R denotes recall.

For each category c, calculate its AP value, i.e., the size of the area under the precision and recall curve for that category. In the case of discretisation, the AP can be calculated in the following way:

$$AP_c = \sum_{n=1}^{N} (R_n - R_{n-1}) P_n \tag{6}$$

where P_n and R_n are the precision and recall at the nth threshold, respectively, and N is the number of thresholds. R_n-R_{n-1} is the increment of recall, and AP_c is the corresponding precision.

In summary, the formula for the average AP value for all categories C, i.e., mAP, is derived:

$$mAP = \frac{1}{|C|} \sum_{c} AP_c \tag{7}$$

where |C| is the total number of categories. For multi-category target detection tasks, mAP provides a measure of overall performance that reflects the average effect of the model across all categories.

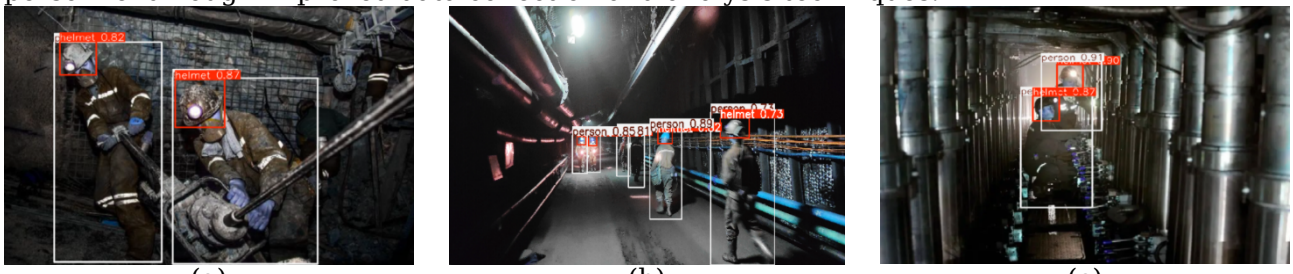
5. Experiments and Results

5.1. Typical case studies

Real-time unsafe behaviour detection on edge devices using the trained YOLOv11 model. Limited to the length of the article, the following focuses on helmet and action class recognition as an example.

5.1.1 Recognizing the wearing of helmets

The initial phase of dataset construction involves the identification of relevant data sources and a comprehensive analysis of the dataset. In the investigation of unsafe behaviors exhibited by underground personnel, primary data were collected from critical locations, including the open-off cut working face and the travelling roadway within a mine located in Yulin. To document the unsafe behaviors of personnel underground, data were captured utilizing explosion-proof mobile phones, supplemented by the review of surveillance footage from various areas of the mine, including the working face and travelling roadway. Additionally, the collaboration and assistance of underground staff were instrumental in obtaining this data.The examination of unsafe behaviors among underground personnel remains in its infancy within the realm of coal mining, resulting in a significant lack of dedicated datasets addressing such behaviors. To mitigate the deficiency of datasets documenting unsafe practices in this sector, the current study incorporated publicly available datasets, including COCO, SHWD, the Hard Workers Dataset, Safety Helmet Detection, and the Helmet Dataset, which pertain to non-mining contexts. Furthermore, a controlled laboratory environment was utilized to simulate underground scenarios, thereby generating an auxiliary dataset for analysis. The training regimen for the enhanced YOLOv11 network consisted of 200 epochs, with a training batch size established at 50 images per iteration. The Intersection over Union (IOU) threshold was set at 0.6, indicating that a predicted bounding box must overlap with the ground truth box by 60% or more to be considered a successful detection of the target object. This methodological framework aims to advance the understanding of unsafe behaviors among underground mine personnel through improved data collection and analysis techniques.



(a) (b) (c)

Figure 10. Effectiveness of Wearing Safety Helmets

After preprocessing and training the measured dataset alongside the public dataset, we take helmet-wearing detection as an example. The helmets are initially labeled with the tags “helmet” and “no-helmet,” as defined by the labeling tool “LabelImg.” Specifically, wearing a helmet is labeled as “helmet,” while not wearing a helmet is labeled as “no-helmet.” The detection results are illustrated in Figure 10. These results demonstrate that the improved YOLOv11x model presented in this paper can effectively detect helmet usage, regardless of varying illumination conditions—whether excessively bright or dim—or in complex backgrounds. Additionally, the model performs well even at long distances and with small-sized individuals. Furthermore, the enhanced model meets the requirements for multi-person detection, yielding accurate results.

5.1.2 Recognizing character movement

Human Skeletal Keypoint Estimation constitutes a fundamental aspect in the domain of computer vision and human pose analysis, involving the precise identification of anatomical landmarks of the human body from visual input, such as images. This process encompasses not only the localization of these keypoints—predominantly situated in the arms, legs, head, and other significant body regions—but also the accurate interconnection of these points to synthesize a coherent skeletal representation of the human form. The keypoint data model presented in this study adheres to the widely recognized COCO dataset format, encompassing a total of eighteen keypoints representing the human skeleton, designated from 0 to 17 sequentially. The corresponding body keypoint information and the resultant processing outcomes are illustrated in Figure 11 of the paper.

The algorithm for action class recognition presents a notable level of complexity, wherein the posture estimation module serves as a pivotal component of the YOLOV11-pose model. The core concept involves the extraction of the original feature map via the YOLOV11 feature extraction network to acquire pertinent information regarding the key points of the human skeleton. This process subsequently leads to the generation of a human skeleton map following appropriate information processing. To establish a framework for classification, specific judgment criteria are defined. These criteria involve the computation of various metrics, such as the distance, height, and angles among the joints within the human skeleton map. This calculated index information is then juxtaposed with threshold values predetermined by the trained model, facilitating a comparative analysis. Ultimately, this method culminates in the classification of unsafe behaviors based on the identified skeletal characteristics and their alignment with the established thresholds.

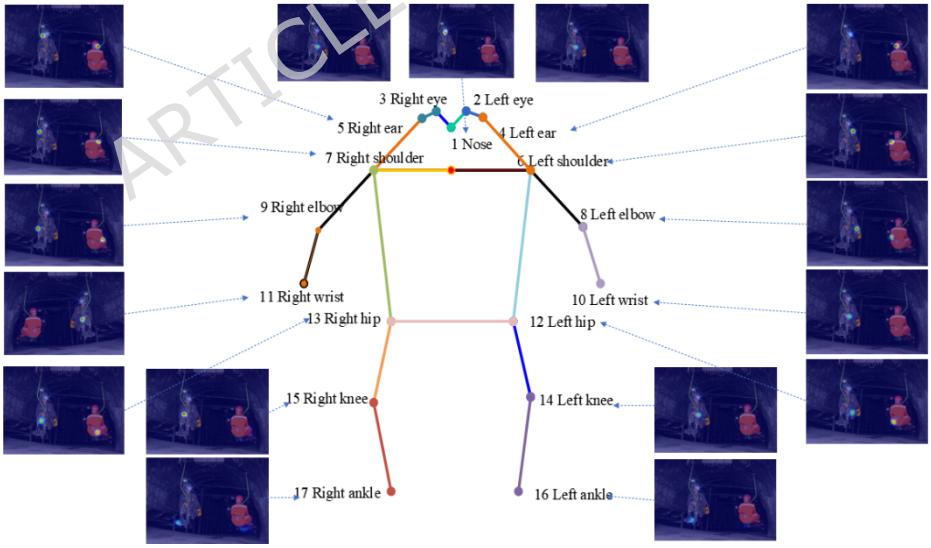


Figure 11. Map of key human body parts and their visual attention hotspots

This image illustrates the application of YOLOv11 for the identification of key skeletal points in miners. Through a detailed analysis of video frames, YOLOv11 accurately detects and labels various key anatomical landmarks on the miner's body, including the nose, eyes, shoulders, elbows, palms, knees, and feet. The capability to detect these key points facilitates the real-time capture of postural changes in miners. By analyzing this posture-related information, one can effectively evaluate the miner's behavioral state, thereby identifying instances of unsafe behavior. Figure 12 presents the results derived from YOLOv11's analysis of miners' skeletal postures, yielding extensive behavioral data from an image analysis perspective. The frame-by-frame examination enables accurate labeling of each key point, such

as the nose, shoulders, and palms, allowing for the continuous tracking of the miner's motion trajectory. This information on postural changes supports a multi-perspective timing analysis, enabling the identification of behavioral patterns. By examining the temporal variations at each key point, it becomes feasible to ascertain potential unsafe behaviors, including improper bending and rapid movements, which may elevate safety risks. Furthermore, the data related to postural changes serves as a fundamental basis for subsequent behavioral feature extraction, ultimately contributing to a more thorough assessment of workplace risks and informing the implementation of appropriate safety measures.

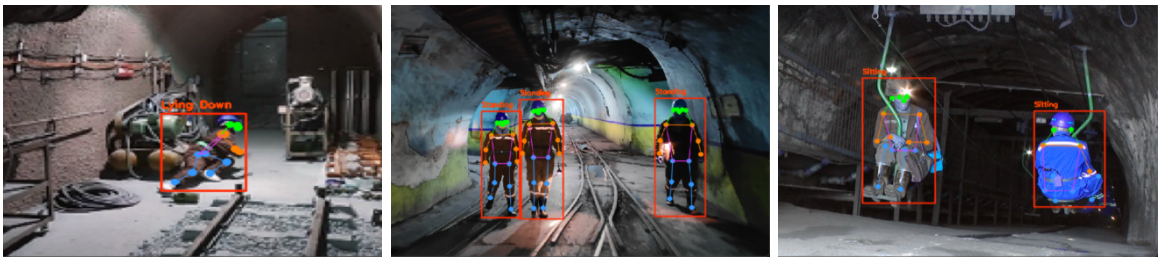
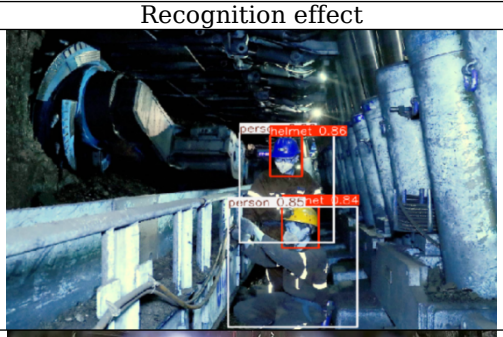


Figure 12. Skeletal recognition results of workers in walking and working conditions

5.2. Anomaly Recognition Cases

The abnormal operating behaviors exhibited by coal miners encompass unsafe equipment usage, improper postures of personnel, and erroneous identification of personnel's positions. Table 5 delineates the identification results of typical safety concerns within the mining environment, assessed through the application of machine vision technology. These concerns primarily include unauthorized access of personnel to hazardous areas, irregular operating behaviors, and safety monitoring within dangerous zones. Through the identification and monitoring of these behaviors, it is possible to issue timely warnings and avert accidents, thereby enhancing safety management practices in the mining context and mitigating risks associated with non-compliant operations.

Table 5. Typical case analysis and identification results

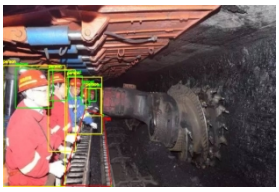
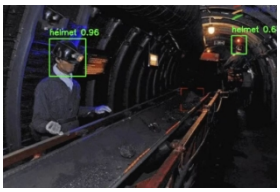
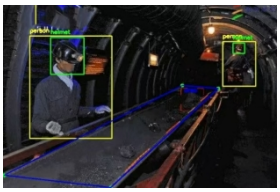
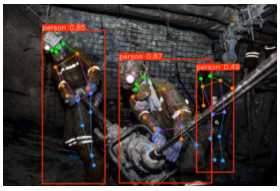
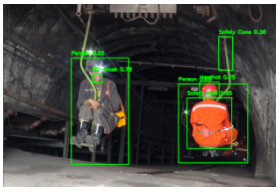
Case Name	Basis for case judgement	Recognition effect
Unusual entry identification of personnel	The machine vision model is used to identify whether there is a person entering the danger zone of the scraper conveyor in the comprehensive mining face, such as identifying that there is a person entering the danger zone when the transfer machine and the roadway belt conveyor are running.	
Identification of abnormal work by personnel	The machine vision model is used to identify whether personnel are working in the hazardous area of the conveyor belt, e.g., when the belt conveyor is recognised as being in a stopped and unlocked state, personnel are working in the hazardous area.	
Identification of foreign objects on belts	When the machine vision model is used to identify the large lumps of coal piled up and large lumps of coal dropped at the overlap of the unloading port of the transfer machine and the conveyor belt, the alarm reminds the operator to deal with the large lumps of coal piled up and the large lumps of coal dropped in a timely manner.	

6. Applications and Discussions

6.1. Comprehensive identification analysis

To assess the visualization effectiveness of the proposed method in the recognition of nine categories of miners' behavior, specific parameters were established. The confidence threshold was set at 0.25, while the Intersection over Union (IoU) threshold was set at 0.35. Original images were selected from the test dataset to represent five behavioral categories: falling, not wearing a helmet, riding illegally, traversing equipment, and entering hazardous areas, as detailed in Table 6. This methodological approach aims to provide a comprehensive analysis of miner behavior recognition within the defined categories. Table 6 illustrates the efficacy of three distinct methodologies employed for the video recognition of miners' behaviors within mining environments. The observed behaviors encompass helmet utilization, identification of hazardous areas, crossing safety barriers, drilling activities, ambulation, and usage of mine overhead passenger transport devices. A comparative analysis reveals variations among the methodologies concerning accuracy, localization of detection frames, and the processing of complex backgrounds. Notably, the methodologies do not exhibit uniform effectiveness in the identification of hazardous areas and the detection of multiple individuals. Furthermore, certain methods demonstrate limited performance when confronted with intricate background conditions. Nevertheless, the overall effectiveness of these approaches in recognizing miners' behaviors is evident. This comparative evaluation serves as a significant reference point for the ongoing enhancement of safety monitoring practices in mining operations.

Table 6. Comparative analysis of identification effects of multiple methods

Type	image	YOLOv5	YOLOv8	YOLOv11
Helmet				
Regional risk				
Across the belt				
Rilling				
Mine overhead passenger transport devices				

The analysis presented in Table 6 elucidates the comparative performance of various detection methodologies applied to the mining environment. The original image serves as a reference for the unprocessed scene, while 'Method 1', 'Method 2', and 'Method 3' represent the outcomes yielded by each respective detection technique. (1) Helmet Wearing Detection: All

three methods successfully identify whether a miner is wearing a helmet; however, subtle discrepancies arise in terms of accuracy and the delineation of bounding boxes. (2) Hazardous Area Identification: Notable variances are observed in the results for hazardous area identification among the three methods. Specifically, differences manifest in the selection and labeling of boundary lines, reflecting distinct interpretations of the geographical extent of hazardous areas. (3) Detection of Behavioral Actions: Each method demonstrates proficiency in detecting various behavioral actions, such as crossing belts, drilling, walking, and the use of mine overhead passenger transport systems. Nevertheless, differences in precise positioning, detection frame accuracy, and personnel identification are evident. Certain methodologies exhibited superior performance in scenarios involving multiple miners, while others encountered challenges in detecting activities against complex backgrounds. In summary, the findings encapsulated in this table underscore the applicability, strengths, and limitations inherent in diverse detection techniques within mining contexts, elucidating the nuanced differences across various methodologies.

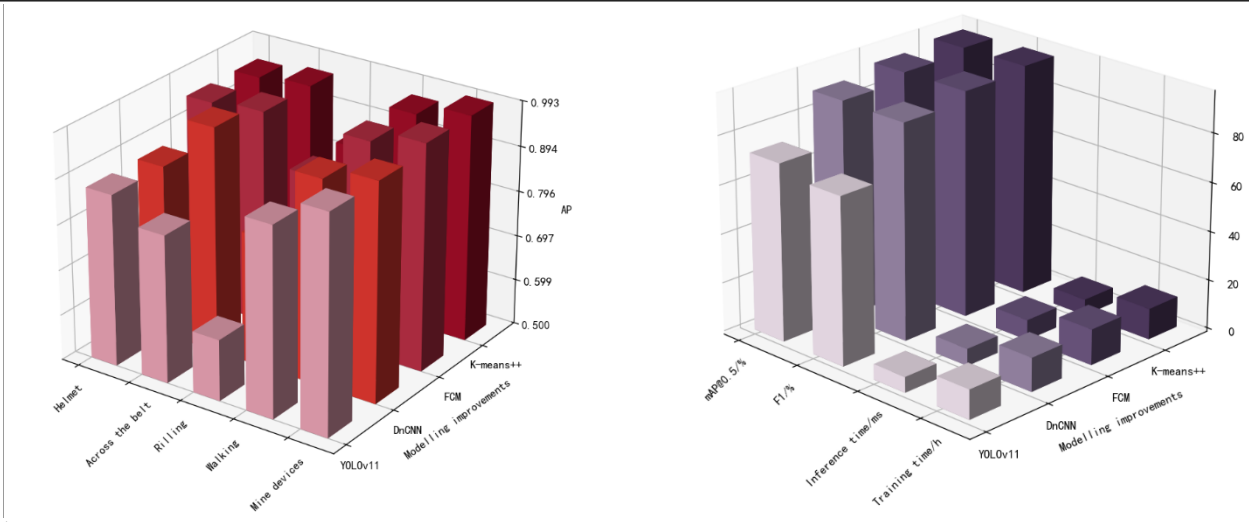
6.2. Analyses and tests

To assess the effectiveness of the integrated modules in YOLOv11, we conducted a comprehensive ablation study consisting of 100 iterations utilizing the Miner-Action dataset. YOLOv11 served as the baseline network, to which we sequentially incorporated DnCNN denoising, the Fuzzy C-Means (FCM) module, and K-means++ optimization techniques. The evaluation metrics employed in this ablation study included precision, recall, inference time, and training duration. We performed a comparative analysis of various model configurations based on metrics such as mean Average Precision (mAP), precision, recall, inference time, and training time. The results were systematically visualized through tables and graphs to facilitate an effective comparison of the algorithms' performance across the different evaluation metrics. Specifically, Table 7 and Figure 13 present the findings of this comparative analysis. Table 7 details the performance comparison among the different model configurations, thereby substantiating the efficacy of the proposed DnCNN denoising, the FCM module, and the K-means++ anchor frame optimization in enhancing the performance of mining worker behavior detection.

Table 7. Table of ablation comparison

Model	DnCNN	FCM	K-means++	mAP@0.5/%	F1/%	Inference time /ms	Training time /h
YOLOv11				73.1	69	6.8	12.5
YOLOv11	√			89.5%	89	7.4	14.2
YOLOv11	√	√		92.7%	93	8.7	15
YOLOv11	√	√	√	94.7	95	7.0	13.04

Figure 13 (a) illustrates the performance evaluation of Average Precision (AP) across five distinct miner behavioral categories, taking into account the implementation of various ablation modules. The data presented in Figure 13 (a) reveals that the AP for the traversing equipment behavioral category attains an impressive 95.4%. This high performance can be attributed to the reduced similarity in the range of strenuous movements compared to other behavioral categories, thereby facilitating more accurate identification. However, in complex mining environments where the activities of miners may be constrained, certain unsafe behaviors exhibit a significant correlation with the surrounding environment. Notably, as depicted in Figure 13 (a), while the precision and recall metrics for the drilling category are comparatively low, it nonetheless meets the anticipated performance outcomes.



(a) Comparison of Behavioural Category Recognition Accuracy

(b) Model Performance Comparison

Figure 13. The performance evaluation of miner behavior dataset by ablation study

The results depicted in Figure 14 illustrate the superior performance of the enhanced YOLO model employed in this study when compared to its original counterpart in the context of downhole environments. This improvement can be attributed to the DnCNN algorithm's adeptness at mitigating noise generated by the distinctive conditions encountered in downhole settings, which often include blurred contours, grainy backgrounds, and shadows resulting from uneven lighting. Such advancements facilitate a more accurate recognition of critical details and features within the images analyzed. Furthermore, the integration of the Fuzzy C-Means (FCM) module enhances the model's capacity to identify significant regions within complex background environments. The comparative experiments conducted validate the efficacy of the proposed DnCNN denoising technique, the FCM module, and the K-means++ anchor frame optimization in augmenting the performance of mining worker behavior detection. Through a series of ablation experiments, noteworthy findings emerged: the amalgamation of DnCNN and the FCM module resulted in a mean Average Precision (mAP) of 92.7%, accompanied by substantial improvements in precision and recall metrics. Nevertheless, it is important to note that these enhancements were accompanied by a slight increase in both inference and training times. The subsequent integration of DnCNN, the FCM module, and K-means++ anchor frame optimization further elevated the mAP to 94.7%, with an inference time of 7.0 ms and a training duration of 13.04 hours, thus achieving a commendable equilibrium between accuracy and computational efficiency.

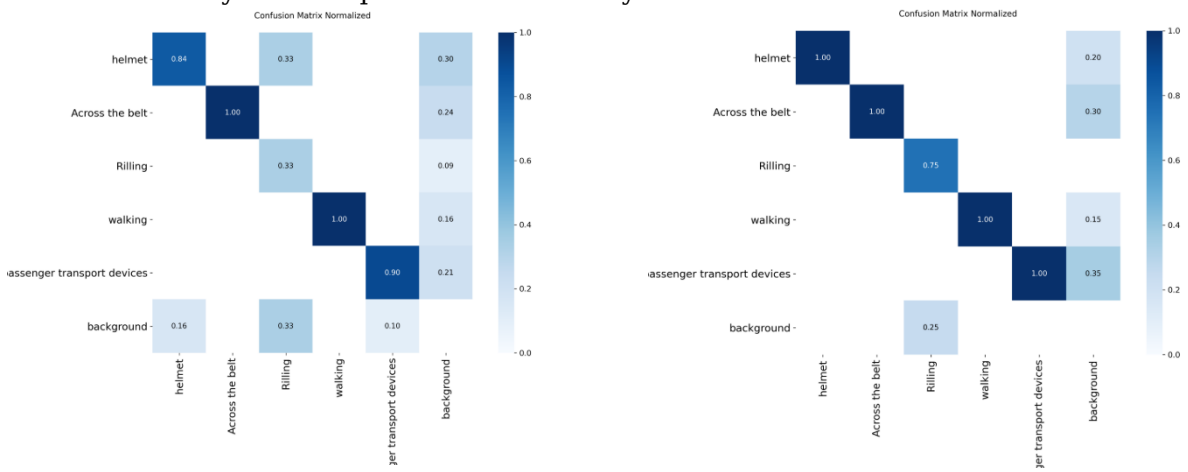


Figure 14. Model Effect Comparison

6.3. Comparative analysis

In the context of miner behavior recognition within intricate underground operating environments, conventional recognition methodologies often fall short of the requisite performance standards regarding computational efficiency and recognition accuracy. This

study proposes an enhanced collaborative architecture predicated on YOLOv11, specifically designed for the precise detection of miners' behaviors, and conducts an empirical evaluation utilizing the Miner-Action dataset. The approach incorporates video segmentation sampling, which facilitates the extraction of key frames depicting non-standard actions performed by miners in video footage. Significantly, this architecture integrates deformable convolution into the network backbone, thereby enhancing the extraction of pivotal features. To further elevate network efficiency, we introduce an augmented collaborative loss function that accelerates convergence during training. Empirical analyses underscore the proposed method's advantages, including a relatively modest parameter count, abbreviated training duration, and elevated detection accuracy. This method achieves a commendable balance between computational speed and recognition accuracy, enabling the identification of non-standard behaviors among miners—a factor of paramount importance for ensuring the safety of personnel in underground environments. To substantiate the effectiveness of the proposed methodology, a comparative analysis is executed against conventional methods, with findings detailed in Table 8 and illustrated in Fig. 15.

Table 8. Table of ablation comparison

Model	Input size	mAP@0.5/%	Parameters /m	Training time /h
YOLOv5	640×640	59.6	7.07	11.15
YOLOv8	640×640	67.5	7	13
YOLOv11	640×640	73.1	6.8	12.5
Improved YOLO	640×640	95.7	7.0	13.04

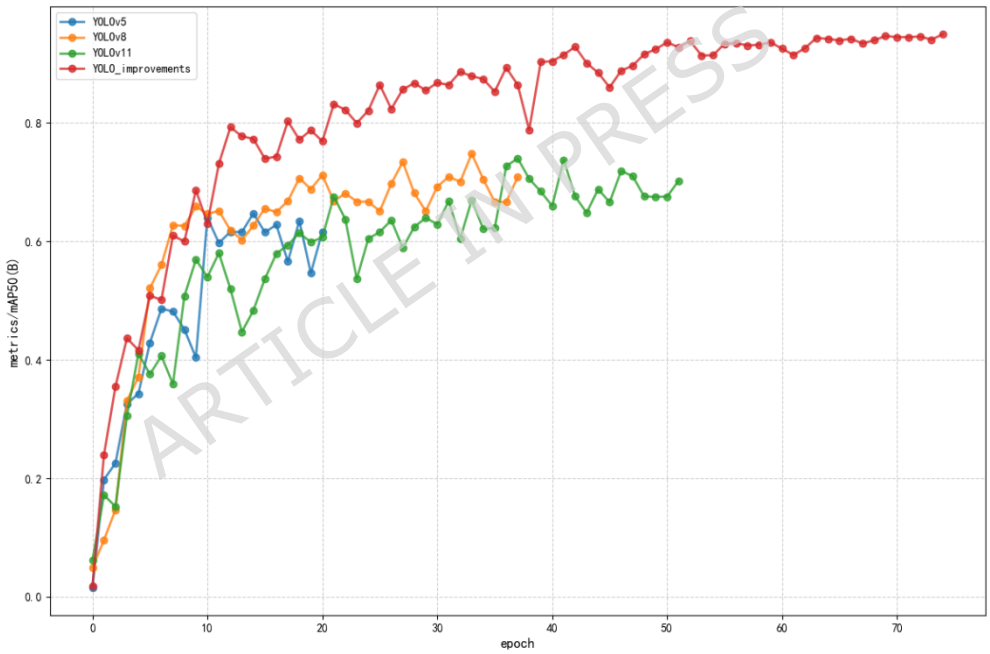


Figure 15. The mAP of different methods

In future research endeavors, there is a pressing need to enhance the recognition of miners' behaviors, decrease the false recognition rate, and bolster the resilience of complex systems against interference. Specifically, the improvement of anti-interference capabilities in intricate underground environments is crucial. Currently, the proposed method for classifying miners' behavior remains underdeveloped, particularly regarding the classification of behaviors under the shielding of mining equipment. Furthermore, the real-time application of the structural model within operational mines is still under investigation. Subsequent research should prioritize exploring the diversity of non-standard behavioral categories exhibited by miners, as well as the real-time implementation of structural models pertaining to these behaviors in mining contexts. The examination of this diversity, coupled with the deployment of network models in terminal devices within underground mines, promises to advance the effectiveness and reliability of behavioral recognition systems in such challenging environments.

7. Conclusion

1) In the context of coal mine safety, the behavioral characteristics associated with the unsafe practices of underground personnel can be systematically categorized. Based on an analysis of prevalent habits and contributory factors affecting the work of underground personnel, unsafe behaviors are classified into three primary categories: object-type unsafe behaviors, action-type unsafe behaviors, and area-type unsafe behaviors. This classification framework serves to facilitate a more efficient analysis and intervention process regarding the unsafe behaviors exhibited by personnel in coal mining environments. By adopting this structured approach, it is possible to enhance the understanding and management of safety risks associated with underground operations.

2) This study enhances the traditional YOLOv11 and YOLO-Pose target detection algorithms by incorporating a functional DnCNN denoising enhancement module along with a K-means++ algorithm for target anchor frame optimization. These advancements significantly augment the efficiency and accuracy of the algorithms in detecting three specific categories of unsafe behaviors exhibited by underground personnel.

3) Utilizing advancements in the YOLOv11 and YOLOv11-Pose target detection algorithms, this study presents a comprehensive early warning model designed to identify unsafe behaviors among underground personnel. Additionally, a systematic network architecture facilitating communication between underground and surface environments is proposed. Experimental validation demonstrates that the methodology outlined in this paper significantly enhances recognition precision and accuracy when compared to traditional approaches.

4) The research predominantly examines the definitions and underlying causes of unsafe behaviors exhibited by underground personnel. It also emphasizes the development of swift and effective methodologies for the identification of such behaviors. While the study engages in a preliminary discourse regarding the locational and attribute-related information of underground workers, it recognizes the necessity for future in-depth investigations into the spatial positioning of personnel as well as the operational dynamics of equipment and apparatus utilized in underground environments.

Author Contributions: Conceptualization, Juan Liang and Quanjie Zhu; Data curation, Juan Liang, Quanjie Zhu and Dongsheng Jiang; Formal analysis, Juan Liang, Quanjie Zhu and Dongsheng Jiang; Funding acquisition, Quanjie Zhu and Yan Liu; Investigation, Juan Liang, Quanjie Zhu and Shaojie Chen; Methodology, Juan Liang, Quanjie Zhu, Dongsheng Jiang and Yan Liu; Project administration, Juan Liang and Quanjie Zhu; Resources, Shaojie Chen, Yan Liu and Quanjie Zhu; Software, Quanjie Zhu, Dongsheng Jiang and Yingnan Hao; Supervision, Juan Liang and Quanjie Zhu; Validation, Dongsheng Jiang, Yingnan Hao and Shaojie Chen; Visualization, Quanjie Zhu, Dongsheng Jiang and Yingnan Hao; Writing – original draft, Juan Liang, Quanjie Zhu and Dongsheng Jiang; Writing – review & editing, Quanjie Zhu and Dongsheng Jiang.

Data availability: The datasets analysed during the current study are not publicly available due to privacy and commercial restrictions but are available from the corresponding author on reasonable request.

Funding: This work was supported and financed from the Langfang Science and Technology Program (grant number: 2024013001), the Hebei Natural Science Foundation (grant number: E2023508021), the Fundamental Research Funds for the Central Universities (grant number: 3142021002, 3142024005).

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

References

1. Aslanyan M. On mobile pose estimation and action recognition design and implementation. Pattern Recognition and Image Analysis, 2024, 34(1): 126-136.
2. Bishop C M, Nasrabadi N M. Pattern recognition and machine learning. New York: Springer, 2006.
3. Cao X, Zhang C, Wang P, et al. Unsafe mining behavior identification method based on an improved ST-GCN. Sustainability, 2023, 15(2): 1041.
4. Chen C, Xiang H, Huang S, et al. Study on Human Hazardous Behavior Recognition and Monitoring System in Slide Facilities Based on Improved HRNet Network. International Journal of Advanced Computer Science & Applications, 2025, 16(3).

5. Fiedler M A, Rapczyński M, Al-Hamadi A. Fusion-based approach for respiratory rate recognition from facial video images. IEEE Access, 2020, 8: 130036-130047.
6. Gao H, Wang X, Liu Z, et al. Healthcare System from Multisensor Collaboration and Human Action Recognition. Sensors & Materials, 2024, 36.
7. Graves A, Graves A. Long short-term memory. Supervised sequence labelling with recurrent neural networks, 2012: 37-45.
8. Guo Y, Meng W, Fan Y, et al. Wearable sensor data based human behavior recognition: a method of data feature extraction. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(8): 1246-1253.
9. Khan M N, Das S, Liu J. Predicting pedestrian-involved crash severity using inception-v3 deep learning model. Accident Analysis & Prevention, 2024, 197: 107457.
10. Kolar Z, Chen H, Luo X. Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images. Automation in Construction, 2018, 89: 58-70.
11. Lee T, Mihailidis A. An intelligent emergency response system: preliminary development and testing of automated fall detection. Journal of Telemedicine and Telecare, 2005, 11(4): 194-198.
12. Liang F. Construction site safety helmet wearing detection method based on improved YOLOv5. Journal of Physics: Conference Series, 2023, 2560(1): 012042.
13. Lijuan L, Peng Z, Shipin Y, et al. YOLOv5-SFE: An algorithm fusing spatio-temporal features for detecting and recognizing workers' operating behaviors. Advanced Engineering Informatics, 2023, 56: 101988.
14. Liu Y, Yu Y, Lu Y L, et al. Real-time human activity recognition based on time-domain features of multi-sensor. Zhongguo Guanxing Jishu Xuebao (Journal of Chinese Inertial Technology), 2017, 25(4): 455-460.
15. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533.
16. Özyer T, Ak D S, Alhajj R. Human action recognition approaches with video datasets—A survey. Knowledge-Based Systems, 2021, 222: 106995.
17. Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.
18. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.
19. Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms. International Conference on Machine Learning. PMLR, 2015: 843-852.
20. Standard for Classification of Casualties in Enterprise Employees (GB 6441-1986). State Bureau of Standards, 1986.
21. Vukicevic A M, Petrovic M N, Knezevic N M, et al. Deep learning-based recognition of unsafe acts in manufacturing industry. IEEE Access, 2023, 11: 103406-103418.
22. Wang C, Zhang H, Zhai Z. Real time dangerous action warning system based on graph convolution neural network. Academic Journal of Computing & Information Science, 2022, 5(6): 89-94.
23. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).
24. Yang X, Zhang X, Ding Y, et al. Indoor activity and vital sign monitoring for moving people with multiple radar data fusion. Remote Sensing, 2021, 13(18): 3791.
25. Yao W, Wang A, Nie Y, et al. Study on the recognition of coal miners’ unsafe behavior and status in the hoist cage based on machine vision. Sensors, 2023, 23(21): 8794.
26. Zhe C, Gines M H, Tomas S, et al. OpenPose: Realtime multi-person 2D pose estimation using part

affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(1): 172-186.
27. Cheng T. Research and implementation of intelligent monitoring system for training workshop based on OpenPose. Kunming University of Science and Technology, 2020.