AIMS Mathematics

AIMS Mathematics, 9(11): 31295–31316.
DOI: 10.3934/math.20241509
Received: 24 July 2024
Revised: 28 September 2024
Accepted: 24 October 2024
Published: 04 November 2024

https://www.aimspress.com/journal/Math

Research article

FOA-BDNet: A behavior detection algorithm for elevator maintenance personnel based on first-order deep network architecture

Zengming Feng[1,2,*] and Tingwen Cao[2]

[1] School of Mechanical Engineering, Hubei University of Technology, Wuhan, China
[2] Wuhan Vocational College of Software and Engineering (Wuhan Open University), Wuhan, China

* **Correspondence:** Email: fengzengming6561@sina.com; Tel: +18627116561.

Abstract: The operation space of the vertical lift shaft is small, the components are complex, the occluding and different behavior space characteristics are similar, and the unsafe behavior is not easy to detect, which makes the operation safety of maintenance personnel in the elevator greatly threatened. This paper proposes an elevator maintenance personnel behavior detection algorithm based on the first-order deep network architecture (FOA-BDNet). First, a lightweight backbone feature extraction network is designed to meet the online real-time requirements of elevator maintenance environment monitoring video stream detection. Then, the feature fusion network structure of “far intersection and close connection” is proposed to fuse the fine-grained information with the coarse-grained information and to enhance the expression ability of deep semantic features. Finally, a first-order deep target detection algorithm adapted to the elevator scene is designed to identify and locate the behavior of maintenance personnel and to correctly detect unsafe behaviors. Experiments show that the detection accuracy rate on the self-built data set in this paper is 98.68%, which is 4.41% higher than that of the latest target detection model YOLOv8-s, and the reasoning speed reaches 69.51fps/s, which can be easily deployed in common edge devices and meet the real-time detection requirements for the unsafe behaviors of elevator scene maintenance personnel.

Keywords: behavior detection; target detection; deep learning; real-time detection; elevator safety
Mathematics Subject Classification: 68T07, 68T45, 74A08

1. Introduction

With the advancement of urbanization, elevators have become the main means of transport for modern buildings, and is the most widely used. As a special means of elevator transportation, its safety has always attracted public attention. To improve the safety of elevator operation and reduce the incidence of elevator failure, it is very important to carry out the regular maintenance of elevators [1]. As the guardian of elevator safety, elevator maintenance personnel need to regularly enter the elevator car top, elevator shaft, and other dangerous areas to replace parts, repair faults, clean guide rails and chains, and to perform lubrication operations. Due to the narrow space and poor light of the elevator shaft, the vision of maintenance personnel is limited, and there are many mechanical equipment arranged on the top of the elevator bridge. As a result, the top surface of the elevator car forms a complex spatial structure, and the elevator maintenance personnel are very prone to accidental falls, falls, electric shocks, or pincers on the elevator shaft and bridge top surface, which seriously affect the personal safety of the elevator maintenance personnel [2,3]. Therefore, it is of a great significance to monitor the behavior of the maintenance personnel in real time during elevator maintenance.

At present, elevator behavior detection work mainly focuses on the impact of passengers' uncivilized behavior on elevator safety [4−7], and few researchers have paid attention to the behavior detection of elevator maintenance personnel. Some other researchers have installed sensing devices in the elevator shaft to monitor the safe operation of the elevator and any illegal intrusion behaviors [8−10], though the crucial detection of the safety behavior of maintenance personnel was still not involved. Recently, some researchers have begun to study the detection of unsafe behaviors of personnel when the elevator runs in different scenarios. Wei et al. [11] developed a multi-scale cascaded feature fusion YOLOv5s model (AMCFF-YOLOv5s) based on the attention mechanism to estimate the number and category of miners and to identify unsafe behaviors, such as whether miners wear safety hats. However, only a single unsafe behavior, such as whether to wear a hard hat, was studied, and more serious behaviors such as falling and holding a wire cable were not studied. Kong et al. [12] used computer vision and long and short-term memory methods to predict unsafe behaviors in construction scenes. This method can automatically predict people's possible unsafe behaviors by monitoring the behaviors of construction workers in real-time. The long short-term memory (LSTM) algorithm is good at dealing with time series prediction, but lacks the generalization ability for real-time detection of high dimensional data such as images. Casini [13] et al. comprehensively reviewed the application of virtual reality (VR), augmented reality (AR), and mixed reality (MR) technologies in the safety maintenance of building elevators, however, extended reality (XR) technology is still immature at present, with large investment costs and unstable operations, and cannot adapt to the ordinary environment similar to video surveillance. D'Souza[14] et al. applied the Internet of Things (IoT) technology in the daily maintenance and safety inspection of elevators to improve the work efficiency and to save labor costs. However, the IoT technology requires the support of a large number of sensing equipment, the elevator still needs maintenance personnel when there is an anomaly, and the unsafe behaviors of on-site operators cannot be detected. The above research on elevator safety inspections did not involve the detection of unsafe behaviors of the maintenance personnel, or the detection of simple content. This paper will focus on the real-time detection of various behaviors of elevator maintenance personnel to monitor unsafe behaviors give timely warnings and to fill the lack of research in this field.

The recognition and detection of human behavior are mainly based on surveillance videos and wearable sensors [15]. This paper mainly studies the recognition and detection technologies of abnormal human behavior in elevator maintenance based on surveillance videos. The key to human behavior recognition and detection is to extract key information from video data to represent the behavior, and the quality of the extracted features directly affects the speed and accuracy of the abnormal behavior recognition and detection algorithm. Human behavior feature extraction methods include traditional methods such as human appearance and movement information [16], movement trajectory [17], space-time interest points [18], two-dimensional human bone information [19], and three-dimensional human bone information [20], as well as feature extraction methods based on deep learning. The feature extraction method based on deep learning uses deep neural networks to directly learn deep features from images. When used, the network structure needs to be designed according to the rules of feature extraction, and the network parameters need to be obtained through training and learning. Compared with traditional methods, the deep learning method does not need a manual feature design. In addition, it has a superior universality for problems such as light changes, occlusions, and perspective conversions in different video data. As long as the network is designed according to the object, and the convergence is trained in specific data samples, the end-to-end abnormal behavior identification can be realized without human interference in other cases. Liao [21] et al. use a 3D convolutional neural network (CNN) for action recognition, to introduce attention mechanism into the network framework, and to add values associated with the foreground region in the feature mapping to construct residual attention units to reduce the adverse impact of background motion on the recognition process. However, the number of 3D convolutional parameters is huge, and it is difficult to converge in model training. At the same time, the training resources incur high costs. Feichtenhofer et al. [22] proposed a spatial-temporal fusion structure based on time and space flows. After fusing the feature graphs of the two networks at different levels, the three-dimensional CNN was used to process the fused features, to better realize the interaction between the time network and the space network, and to effectively detect the action behaviors in the video stream. However, the two-flow structure can not realize the effective information exchange between the two convolutions, and alternatively introduces additional redundant information, which reduces the robustness of the model. Sudhakaran [23] et al. took the differences between adjacent frames as the input and encodes them using convolutional long and short term memory networks to realize an end-to-end violent behavior detection of the video stream data. The recurrent neural network can reflect the relationship between the time series data, however, when solving the problem of long series, the recurrent network is prone to the problem of gradient disappearance, and the lack of generalization ability to slow movement behavior in the video.

Through the above review and summary, it can be found that there are few studies on the behavior detection of elevator maintenance personnel at present; however, the injury incidents of maintenance personnel that occur during the regular maintenance work of elevators are not uncommon. Therefore, this paper conducts a real-time online detection of the behavior of elevator maintenance personnel based on advanced deep learning methods to make up for the lack of research in this area and to improve the accuracy and efficiency of unsafe behavior detection. Designing a deep neural network target detector suitable for elevator scenes is the key to accurately and efficiently detecting unsafe behaviors. In deep neural networks, 3D CNNs require a large amount of computation and are difficult to train, and recurrent neural networks are prone to problems such as gradient disappearance, and a long-term dependence on reduced accuracy, and are not suitable for processing image data. In this paper, a two-dimensional CNN is used to construct a deep target detection model. First, considering the real-time requirements of detection tasks, the CSPNet [24] structure is adopted in the backbone feature extraction network for lightweight transformation. The SPPFCSP [25] module is added at the end of the backbone network to improve the detection accuracy and speed, to adapt to the multi-scale changes of detection targets, and to improve the detection speed. Then, a feature fusion structure with a far intersection and close connection is designed to map the shallower feature layer to the deeper feature layer, and the sub-shallower feature layer is fused with the sub-deep feature layer to promote the fusion learning of fine-grained information and coarse-grained information, and to improve the robustness of the model for objects with similar colors and textures, large scale changes, and complex outlines. Finally, a first-order deep object detection network (FOA-BDNet) is constructed for the behavior detection of the environmental maintainers of elevator shafts. This network considers the lightweight transformation and feature fusion methods previously designed and adopts an Anchor-free model to classify regression targets in the decoding stage of the prediction head. The auxiliary discriminant algorithm of unsafe behavior is designed to improve the judgment accuracy of difficult processes and to distinguish falling and squatting behavior.

To sum up, the main contributions of this paper are as follows:

(1) This paper studies the real-time detection of unsafe behaviors in the process of elevator maintenance by the deep learning method, which makes up for the lack of research in this field;

(2) A first-order target detector (FOA-BDNet) adapted to the elevator shaft environment is designed to detect multi-scale targets in small, complex, and dim spaces;

(3) The algorithm to identify and detect unsafe behaviors, such as hard hat detection, falling, hand-holding steel cable, and standing beams, is designed to realize the high-precision real-time detection requirements of various unsafe behaviors.

2. Materials and methods

2.1. Introduction to the dataset

In this paper, field research and semi-structured interviews of elevator maintenance enterprises were conducted and combined with the needs of elevator companies. The definition of the elevator maintenance personnel's unsafe behavior contained elevator operation without helmets, hand grasping the wire rope, and standing on the beam Additionally, to increase the fall hazardous state, the first three kinds of unsafe behaviors can bring about personal safety hazards, give warning prompts, and the latter kind of hazardous state should be immediately alerted. Samples of the various behavioral states are shown in Figure 1.

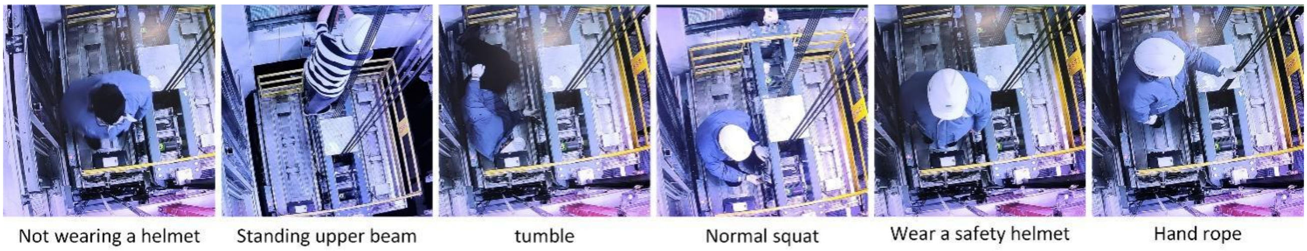


Figure 1. Sample behavioral states.

For the unsafe behaviors of not wearing a helmet, grasping the wire rope by hand, and standing on the beam, the model in this paper only needs to correctly detect the different behavioral states, and the training samples correspond to the three different behaviors; the last hazardous state is usually misdetected by the depth model because it is similar to the normal squatting behavior. It is usually difficult for depth target detection algorithms to differentiate between the two actions of squatting. This paper designs the auxiliary algorithms to differentiate between the two, and subsequently needs to detect the normal squatting behavior. Therefore, it is necessary to detect the normal squatting behavior to obtain the size and localization data, and then combine it with auxiliary algorithms to accurately identify the fall hazardous state. Each behavioral state training sample is maintained at about 50 images, to ensure that the different categories of training samples are roughly balanced, and to avoid having too few samples of a certain category because it is difficult to learn the real knowledge alongside the generalization of the performance deterioration.

2.2. *Deep feature layer fusion algorithm for "far and near connectivity"*

Deep CNNs can capture high-level, intermediate, and even low-level features due to different hierarchies and filter sizes, and can also compress information into a smaller size through pooling mechanisms. The scale invariant nature of CNN models allows for the capture of arbitrarily located target features, making them an ideal model for processing image data, especially for "feature extraction". The shallow feature map of an image contains more pixel information, which is fine-grained information that contains details of the image such as color, texture, edges, corners, etc., while the deeper feature map is coarse-grained information (i.e., more abstract semantic information). However, the accurate description of the target features requires both fine-grained detail features and coarse-grained global semantic features, while the network will lose a large amount of fine-grained information while deepening; the deeper the network, the more serious the loss of detail information. This paper proposes a before and after view of the depth of the feature layer fusion algorithm for the interaction between the detail information and the semantic information of the learning. The network structure is shown in Figure 2, and the feature fusion process can be used in the form of Eqs (1)−(3).

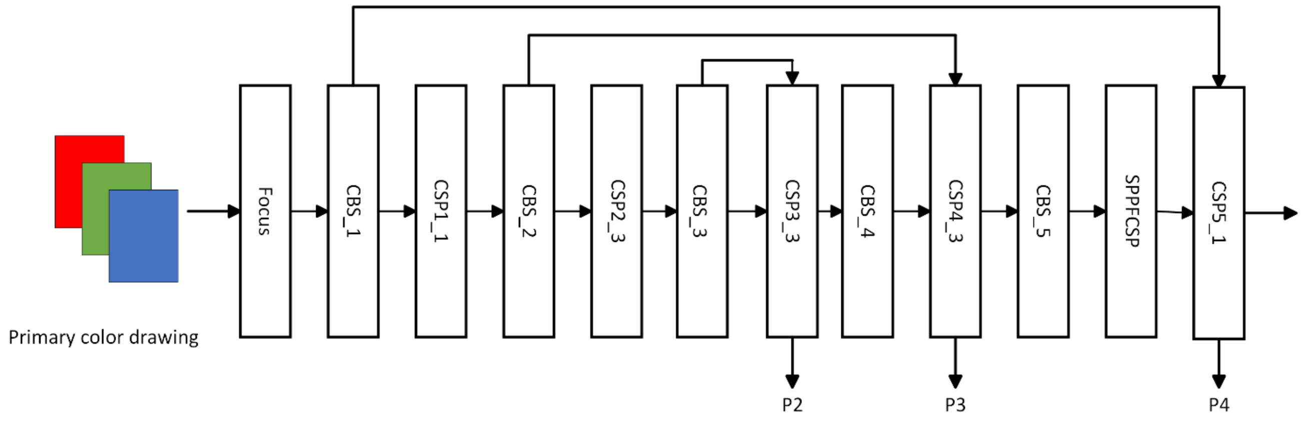


Figure 2. Converged network of "far-flung, near-connected" features.

$$T_1 = f_{3*3}\left(\delta\left(Concat\left[CBS\_1, CSP5\_1\right]\right)\right) \tag{1}$$

$$T_2 = f_{3*3}\left(\delta\left(Concat\left[CBS\_2, CSP4\_3\right]\right)\right) \tag{2}$$

$$T_3 = f_{3*3}\left(\delta\left(Concat\left[CBS\_3, CSP3\_3\right]\right)\right). \tag{3}$$

Equations (1)−(3) describe the network fusion operation: T_1,T_2,T_3 refers to the output of three different scales of feature maps, which provide different scales of target prediction information to the model prediction part; f_{3*3} is the convolution operation with a convolution kernel size of 3; $\delta(\cdot)$ denotes the Relu activation function; $Concat[\cdot]$denotes the superposition of the different feature layers in the channel dimensions; and $Cnotesotes, CBS_2, CBS_3, CSP5_1, CSP4_3$ and $CSP3\_3$ denote the different layers of the feature layers in Figure 2, respectively. Combined with Figure 2 and Eqs (1)−(3), the operation of the "far-from-near" deep feature layer fusion algorithm maps the shallower features from the original input image closer to the deeper feature layers, and the deeper feature layers will be fused to the more initial fine-grained color, edge, and corner features, which will help the model to correctly locate the target and distinguish between similar targets. Features extracted from different network layers represent varying levels of abstract information. In all versions of the YOLO (You Only Look Once) series, the fusion of shallow detail features with deep high-level semantic features has not been considered. This omission limits the model's ability to learn richer and more robust representations. In the elevator environment, slender small targets such as steel cables are prone to be missed, while safety shoes and gloves can be easily misidentified due to their reflective backgrounds. The 'distant interaction and close connection' algorithm integrates the contextual information during the backbone feature extraction phase, thus allowing the model to simultaneously focus on local details and global semantics. This enables the model to learn richer and more robust representations, thus enhancing its expressive capabilities.

2.3. First-order deep target detection network

In this paper, we propose a first-order deep target detection network applicable to the behavior detection of maintenance personnel in complex elevator shaft environments, named FOA-BDNet, which follows the classical first-order network architecture and consists of three main modules: the backbone feature extraction network, the feature enhancement network, and the target prediction network. The overall structure of the network is shown in Figure 3.

In Figure 3, the backbone feature extraction network (Backbone) part uses the well-known CSPDarknet [24] deep learning framework, whose lightweight feature makes it very efficient to run in embedded devices and resource-constrained environments, which greatly facilitates the deployment of the model on edge computing devices in elevator shafts. However, it should be noted that the CSPDarknet network structure adopted in this paper is not without any modification, but adopts a strategy of dynamically adjusting the network depth and width (DAN), which dynamically adjusts the network depth and width according to the complexity of input data to improve the computational efficiency and real-time performance. The pseudo-code for this strategy is shown as Algorithm 1.

Algorithm 1: Dynamically Adjusting Network Depth and Width Strategies

```
    function DAN( x );
Input: input complexity factor
Output: network depth and width
if x = 0 then
        return 0;
if x > threshold_high then
        current_depth = base_depth + 2;
        current_width = base_width + 32;
if x < threshold_low then
        current_depth = max( base_depth - 2, 1) ;
        current_width = max( base_width - 16, 1);
else
        current_depth = base_depth;
        current_width = base_width;
    return new layers;
    end
```

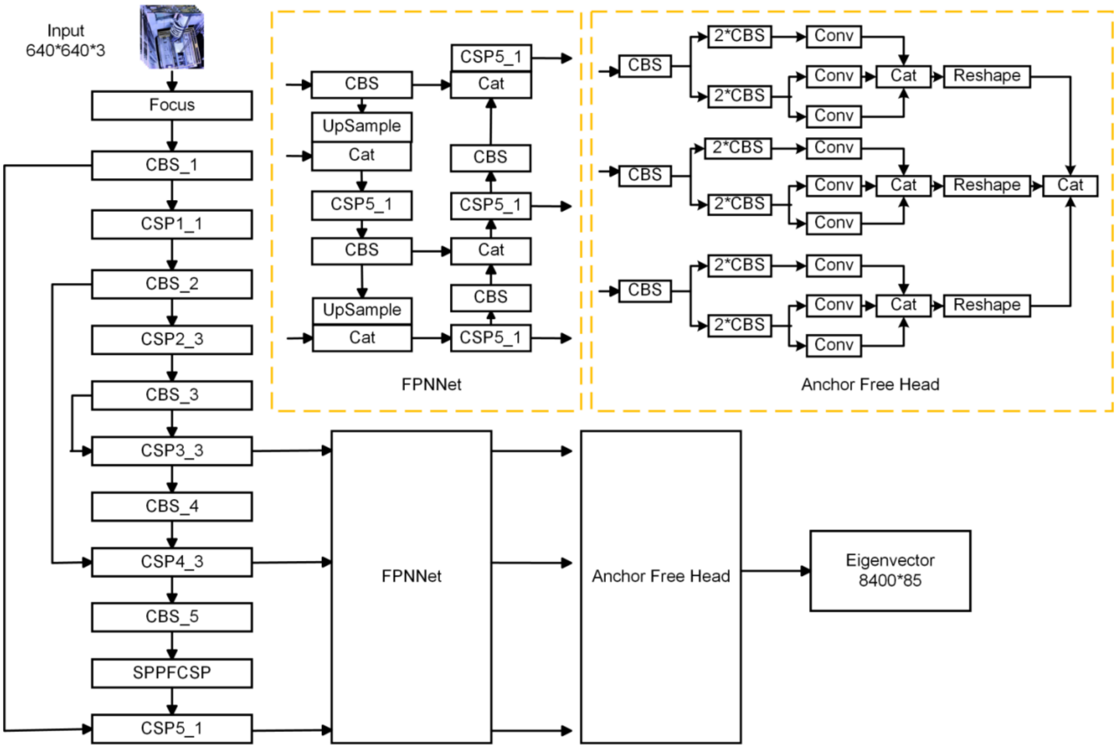


Figure 3. FOA-BDNet model network structure.

Algorithm 1 pseudocode implements a CSPDarknet network model that dynamically adjusts the depth (number of layers) and width (number of channels) of the network based on the input data complexity. In each forward propagation, the input complexity is calculated according to the space size of the input image data. Then, the depth and width of the current network are adjusted according to the set complexity threshold. Finally, the adjusted network structure is constructed through the new network depth and width, and forward propagation is performed. This allows for a dynamic balance between the lightweight design and performance by reducing the computation when dealing with simple inputs, and increasing the network power when dealing with complex inputs.

The feature enhancement network (Neck) not only adopts the classic PAN structure to solve the deficiency of feature pyramid network in multi-scale detection tasks in target detection tasks, but also adds the SPPFCSP [25] module to speed up the network operation speed and to improve the detection accuracy. More importantly, the front and back deep feature layer fusion algorithm proposed in Section 2.2 of this paper is used to strengthen the fusion of the Neck partial detail features and high-level semantic features, and generate feature maps with multi-scale information to improve the accuracy of target detection. In addition, the Anchor-free mode is adopted in the Head part of the prediction network. The Anchor-free mode eliminates the process of generating preset boxes, and does not need to tune Anchor-related hyperparameters, thereby avoiding a large number of IOU calculations, reducing memory usage, and speeding up the model convergence and reasoning speeds.

2.4. Algorithm for assisted calculation of fall status

In the elevator shaft, the overhead view monitoring, fall, and normal squatting states are more similar, and the CNN network has displacement, scale, and deformation invariants to the image features. The target detector is usually difficult to directly distinguish between the two states. The design of the auxiliary algorithm shown in Figure 4, and the expression is in the form of (4)−(6).

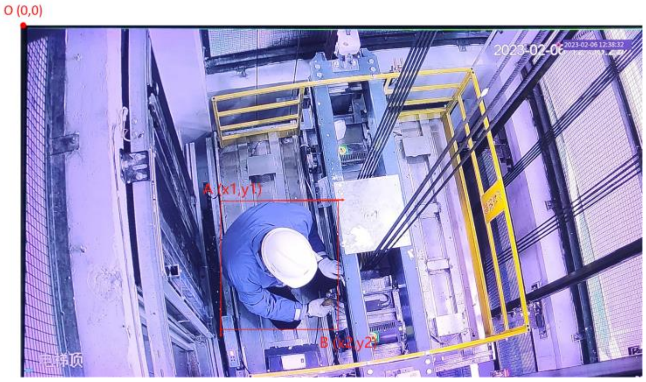


Figure 4. Predictive framing positioning objectives.

$$\mu = \frac{1}{i}\sum_{i=1}^{i} max_i\left(|x1 - y1|, |x2 - y2|\right) \tag{4}$$

$$L = max\left(|x1 - x2|, |y1 - y2|\right) \tag{5}$$

$$\beta = \frac{max\left(|x1 - x2|, |y1 - y2|\right)}{min\left(|x1 - x2|, |y1 - y2|\right)}. \tag{6}$$

The coordinates of two points $A(x_1, y_1), B(x_2, y_2)$ in Figure 4 are the coordinates of the upper-left and lower-right corners of the target predicted by the FOA-BDNet algorithm , respectively which are relative to the image origin $O(0)$, and the width and height of the target frame can be calculated by subtracting the value of the lower-right corner from the value of the upper-left corner. The average value μ of the longest side of i squatting behaviors is calculated according to Eq (4), the longest side m of the target frame is calculated according to Eq (5), and the predicted frame width-to-height ratio β is calculated according to Eq (6). If the FOA-BDNet algorithm predicts the detection result of a fall, then the longest side L of the target frame of the result is obtained. The state is considered to be a fall state if the longest side L is greater than 3/2μ and β is greater than 4/3; if it is less than 3/2μ or β is less than 4/3, then it is considered as a normal squatting behavior. This auxiliary algorithm is designed to improve the accuracy of FOA-BDNet when detecting the fall state and to avoid confusion with squatting or sitting behaviors that may lead to false alarms.

3. Experimental design and analysis

3.1. Experimental platform and parameter settings

In this paper, the experimental hardware environment for the model training test is as follows: the CPU of the computer is 12th Gen Intel® Core™ i5-12600KF 3.70GHz; the system memory is 16G; and the graphics card is NVIDIA GeForce RTX 3070 GPU, with 8GB of video memory capacity. The software environment as follows: Windows 10 operating system; Pycharm compilation environment; PyTorch1.12 deep learning framework; cuda11.6 accelerated computing platform; Anaconda3.0 environment manager; and the programming language is Python3.8. After many experimental explorations, utilizing the parameters set in Table 1, the FOA-BDNet model proposed in this paper to detect insecure behaviors by elevator maintenance personnel achieved a stable and reliable performance.

Table 1. Experimental setup.

Set item	Parameter
Iteration	300
Batch size	32
Initial learning rate	1e-2
Min learning rate	1e-4
Optimizer	SGD
Momentum	0.937
Weight decay	5e-4
Learning rate decay type	COS
Thread	4

In Table 1, the parameter settings follow different principles. The ‘Iteration’ is typically determined based on the size of the dataset, model complexity, and the required accuracy. In this paper, the model converges after 300 iterations. The principle for selecting ‘Batch size’ is that a larger batch size can improve the training speed and reduce the training time, though it may require more memory. Considering the GPU and CPU memory of the experimental platform, a batch size of 32 was chosen to ensure the training speed while accommodating the hardware capabilities. The ‘Initial learning rate’ is usually set at a high value to allow for rapid convergence in the early stages of training. The ‘Min learning rate’ is typically set to 0.01 times the initial learning rate, which is an empirical value to prevent the learning rate from decreasing too low, which would subsequently cause the model to stop learning. The ‘Optimizer’ is chosen based on the task characteristics and the training data. Adam is usually used for more complex models, while SGD performs well in some image tasks; therefore, this paper chose the SGD optimizer. ‘Momentum’ is used to accelerate the SGD optimization process and reduce the oscillation, typically set between 0.9 and 0.999. ‘Weight decay’ is a regularization method to prevent model overfitting, and is usually set to a small value (like 0.0001 or 0.0005); this model chose a value of 0.0005. The 'Learning rate decay type' selects an appropriate learning rate decay strategy (such as step decay, cosine decay, etc.) to optimize the learning rate changes during training. In this model, the 'COS' strategy was chosen for its stability. Finally, the 'Thread' is determined by the number of CPU cores; this paper selected 4 threads to meet the experimental platform's hardware environment.

3.2. Analysis of experimental results

3.2.1. Data set setup and labeling

In Section 2.1, the main content of the dataset has already been introduced. This section focuses on the setup of the dataset in the model training and testing phases. In this work, we designed a first-order deep object detection model to train on the elevator maintenance personnel behavior dataset, thereby enabling the model to have a generalization ability and robustness in efficiently and accurately recognizing and locating different behavioral patterns. The behaviors to be trained include the following five actions: not wearing a safety helmet, grasping the steel cable, standing on the beam, falling, and crouching. The risk factor was high since the collected images were obtained from non-professional maintenance personnel that simulated unsafe behaviors in the elevator shaft. To quickly gather images of the various behaviors, approximately 50 training images and 15 test images were collected for each behavior before swiftly exiting the worksite. Although this is a small dataset, it is sufficient to prevent underfitting due to insufficient training samples while maintaining a relative balance among the different types of samples. The test images are independent of the training images, and the ratio of training to test images is kept within a reasonable range of 7:3 to 8:2. Finally, labels were created using the professional labeling software Labelme; Figure 5 shows an example of the labeling process.

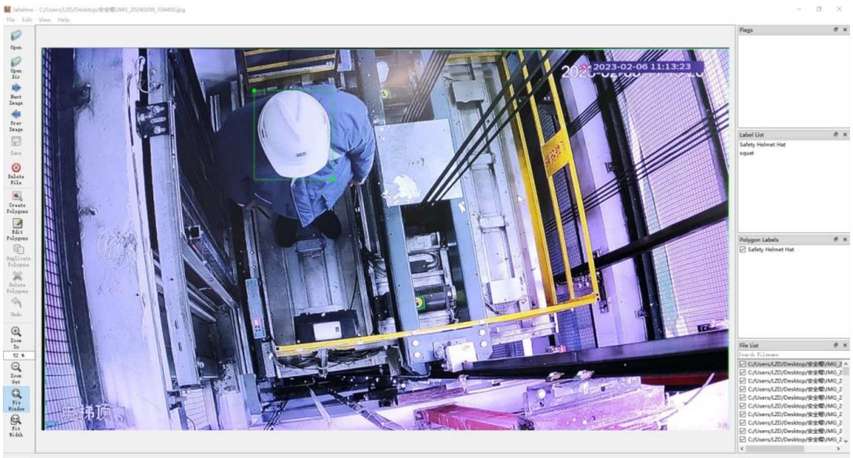

labelme - C:/Users/LZD/Desktop/安全帽/IMG_20240206_104450.jpg
File
Edit
View
Help
Open
Open Dir
Next Image
Prev Image
Save
Delete File
Create Polygons
Edit Polygons
Duplicate Polygons
Delete Polygons
Undo
Zoom In
52 %
Zoom Out
Fit Window
Fit Width
2023-02-06 11:13:23
电梯顶
Flags
Label List
Safety Helmet Hat
squat
Polygon Labels
☑ Safety Helmet Hat
File List
Search Filename
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2
☑ C:/Users/LZD/Desktop/安全帽/IMG_2

Figure 5. Labelme label maker tool labeling samples.

3.2.2. Evaluation indicators

Unlike traditional classification algorithms, the relevant metrics to evaluate the performance of a deep CNN target detection model are mainly Average Precision (AP) and Frames Per Second (FPS). AP is the area under the Precision-Recall curve: the larger the value of AP, the higher the average accuracy of the model. The mean accuracy (mAP) is the average accuracy of all categories of labels. The recall, also known as the recall rate or the check all rate is the ratio of correctly recognized objects to the total number of objects: a larger recall rate indicates that the model detects more correct targets. FPS, namely the number of images that can be processed in one second, can measure the speed of the model to process the data, and it is an important index to considering the real-time performance, usually under the same hardware conditions to be compared mFPS is the average obtained from multiple tests. The F1-score, also known as the F1 score, is a measure of classification problems, and is often used as the final indicator of the multi-classification problem. It is the average of the reconciliation of the precision and the recall, and points out that recall and precision are equally important. The F1-score takes the value range of [0,1], in which 1 represents the perfect classifier and 0 represents the worst classifier.

3.2.3. FOA-BDNet performance evaluation

To validate the effectiveness of the FOA-BDNet model, the performance of the model was tested in two dimensions: Objective metrics and field validation. The objective metrics adopt the target detection evaluation method mentioned above, including the class mAP, the AP, the Recall, and the F1 score. These four objective evaluation metrics evaluate the performance of the FOA-BDNet model in detecting the targets in four different dimensions, among which the “standing beam” behavior is more similar to the “normal standing” behavior, which is more difficult to distinguish; the performance of the model to detect difficult targets is more similar to the “normal standing” behavior. The behavior of the “standing beam” is similar to the behavior of the “normal standing”, which is more difficult to distinguish; the performance of the model to identify the detection target is more indicative of the generalization and robustness of the model. Therefore, this behavior is used as a reference standard and is compared with the famous first-order deep target detection model YOLOv8-s, which was recently released. The results of the comparison are shown in Figures 6–9.

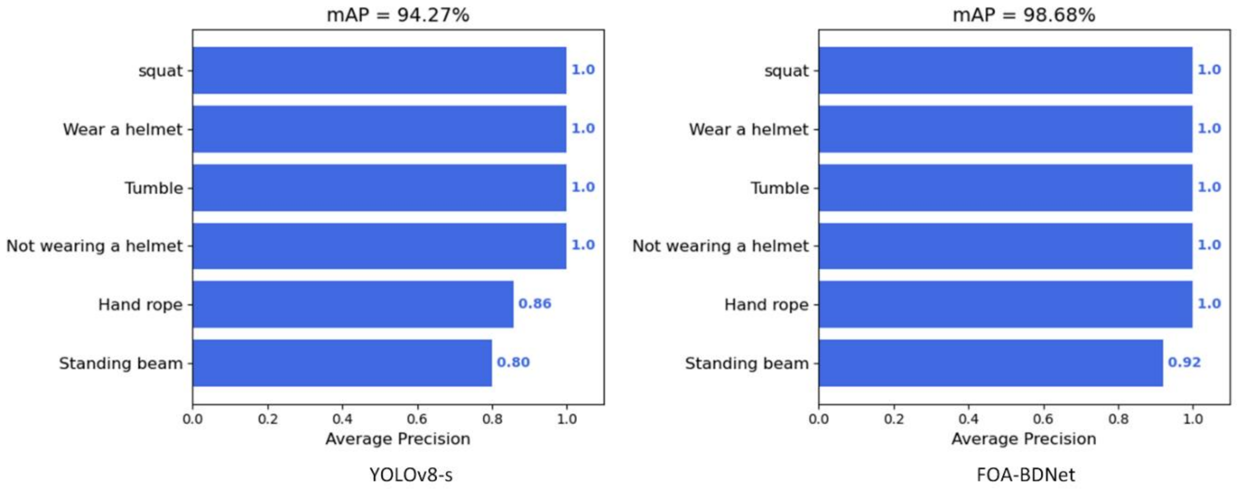
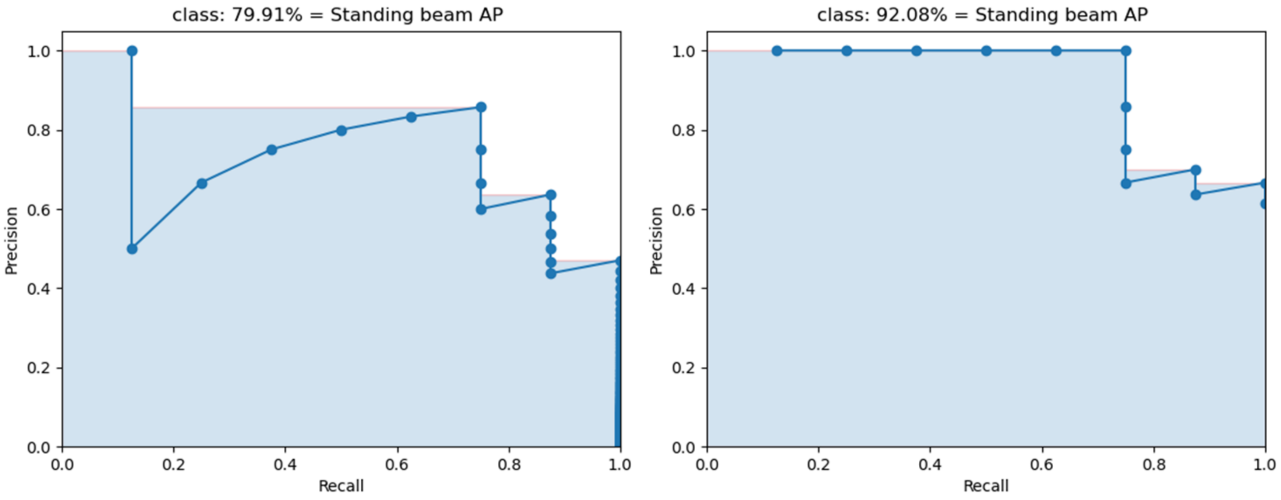
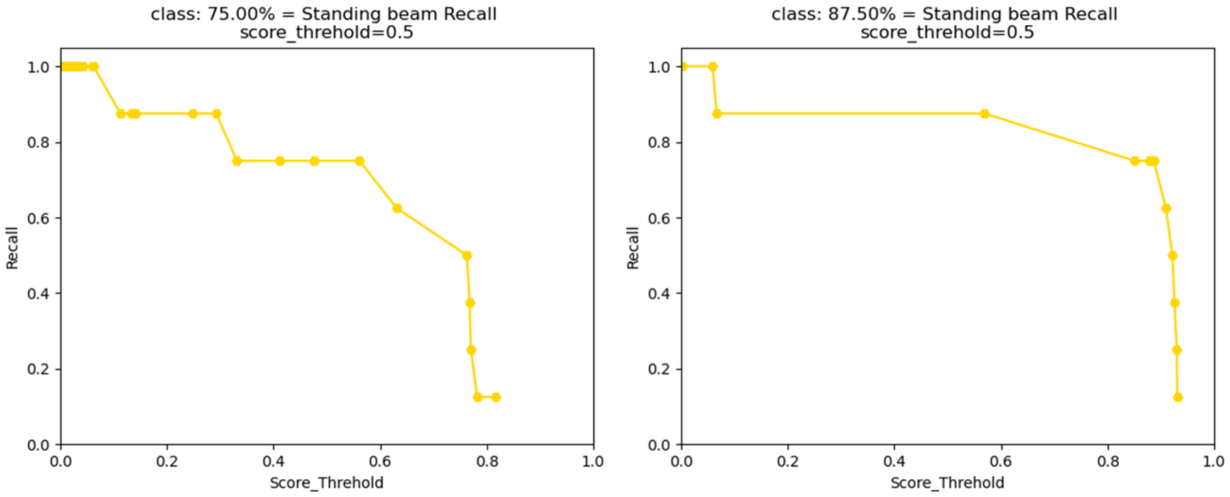


Figure 6. Comparison of FOA-BDNet and YOLOv8-s in mAP metrics.



(a) (b)

Figure 7. Comparison of FOA-BDNet and YOLOv8-s in Precision metrics, (a) YOLOv8-s metrics (b) FOA-BDNet metrics.



(a) (b)

Figure 8. Comparison of FOA-BDNet and YOLOv8-s in Recall metrics, (a) for YOLOv8-s metrics (b) for FOA-BDNet metrics.

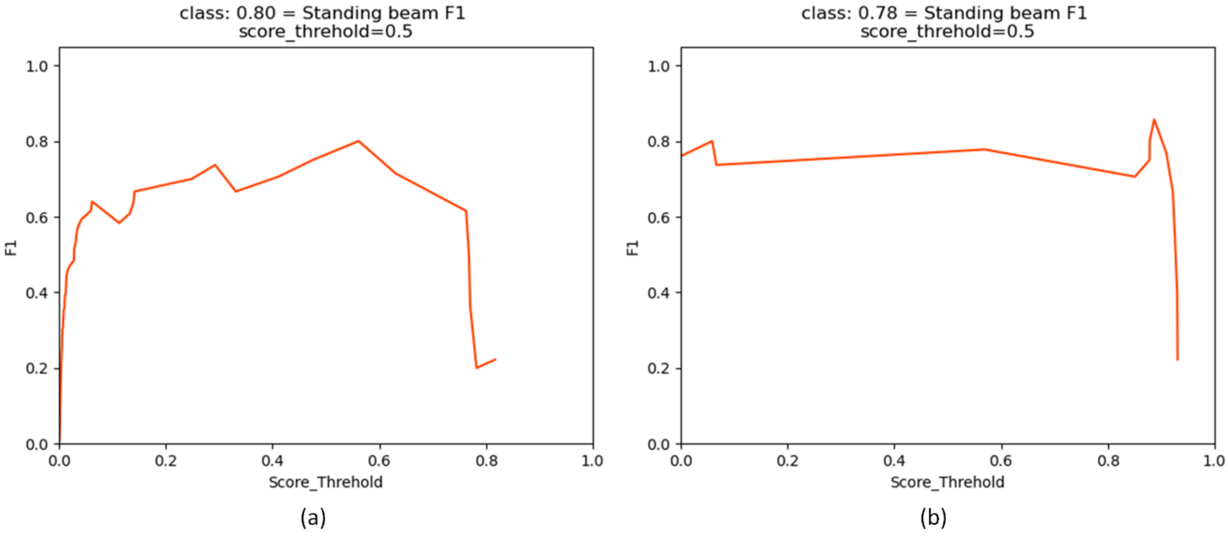


Figure 9. Comparison of FOA-BDNet and YOLOv8-s in F1-score metrics, (a) for YOLOv8-s metrics (b) for FOA-BDNet metrics.

Figure 6 displays the bar graphs of 6 kinds of behavior detection accuracy rates. The FOA-BDNet and YOLOv8-s models performed well in four kinds of behaviors: “Wear a helmet”, “Not wearing a helmet”, “Squat”, and “Hand rope”. However, on the confusing behaviors of “Tumble” and “Standing beam”, the FOA-BDNet model was higher than the AP value tested by YOLOv8-s, thus indicating that the FOA-BDNet algorithm is more suitable for the detection scenarios of elevator maintenance personnel behaviors and has a better resolution performance for behaviors with similar features. Figures 7–9 utilize take the difficult behavior of the standing beam as a reference for a performance comparison. According to Figure 7, the larger the area surrounded by the Precision-Recall curve, the better the robustness of the model. The area surrounded by the proposed algorithm was 12.17% higher than that of YOLOv8-s. The AP index of this behavior is significantly higher than that of the YOLOv8-s model, thus indicating that the model in this paper has a better generalization performance in hard-to-detect behaviors, and shows better robustness as a whole. As shown in Figure 8, when the threshold value of Recall is 0.5, FOA-BDNet will be 12.5% higher than YOLOv8-s. Recall refers to the missed detection rate of the model, which indicates that FOA-BDNet has a much better ability to detect confusing behaviors than the YOLOv8-s model. As shown in Figure 9, although FOA-BDNet is 2% lower than YOLOv8-s at the threshold value of 0.5 on the F1 score indicator, the F1 score of the proposed algorithm is higher than that of the YOLOv8-s model, thus showing a good stability. The greater the F1 value, the better the model generalization performance. The proposed algorithm has a good accuracy and a low false detection rate in detecting the behavior of the elevator maintenance personnel.

The detection results shown in Figure 10 are from the actual tests of this algorithm compared to the YOLOv8-s algorithm in a real-world environment. When observing Figures 10(a) and 10(e), it is evident that both FOA-BDNet and YOLOv8-s performed exceptionally well in the tasks of ‘Not wearing a helmet’ and ‘Squat’, thereby successfully detecting all instances of these behaviors. This success can be attributed to the clear characteristics of these behaviors, the large size of the targets, and the absence of obstructions. However, the test results in Figure 10 also reveal performance gaps in the detection of three other behaviors, specifically those in (b)–(d). For instance, the steel wire rope is generally thin, making it easy to lose the target during the feature extraction process, and the gloves are often white, which can easily blend in with a white background or reflective objects. In Figure 10(b), the FOA-BDNet model successfully detected all instances of the ‘Hand rope’ behavior in the tested images, whereas the YOLOv8-s model missed detections in two of the test images. This indicates that FOA-BDNet has a better generalization ability for small targets and objects with similar colors. Looking at Figure 10(c), while both FOA-BDNet and YOLOv8-s successfully detected the ‘Standing beam’ behavior, FOA-BDNet’s detection results were more comprehensive, thus successfully identifying instances where both feet were standing on the beam. In contrast, YOLOv8-s exhibited target loss in recognizing the behavior of both feet on the beam. The proposed algorithm can discern subtle color differences between the beam and the bridge box surface, thus allowing for a more complete identification of the targets. In Figure 10(d), both FOA-BDNet and YOLOv8-s exhibited significant errors in detection. Without support from auxiliary algorithms, the FOA-BDNet model was prone to missed detections, while YOLOv8-s confusedv the ‘Tumble’ behavior with the ‘Squat’ behavior, thus leading to false detections. This highlights the difficulty of CNNs in distinguishing similar contour targets through simple classification and localization capabilities. Additionally, the features extracted through convolution and pooling operations struggled to differentiate between the targets of different scales. The results from the various behavior tests in Figure 10 demonstrate that the proposed algorithm performs better in detecting small targets and distinguishing objects with similar colors. However, when relying on size as the primary feature for distinguishing the targets, deep CNNs often exhibit poor generalization. Therefore, additional criteria are needed to differentiate targets that are dissimilar size-wise but similar in other characteristics.

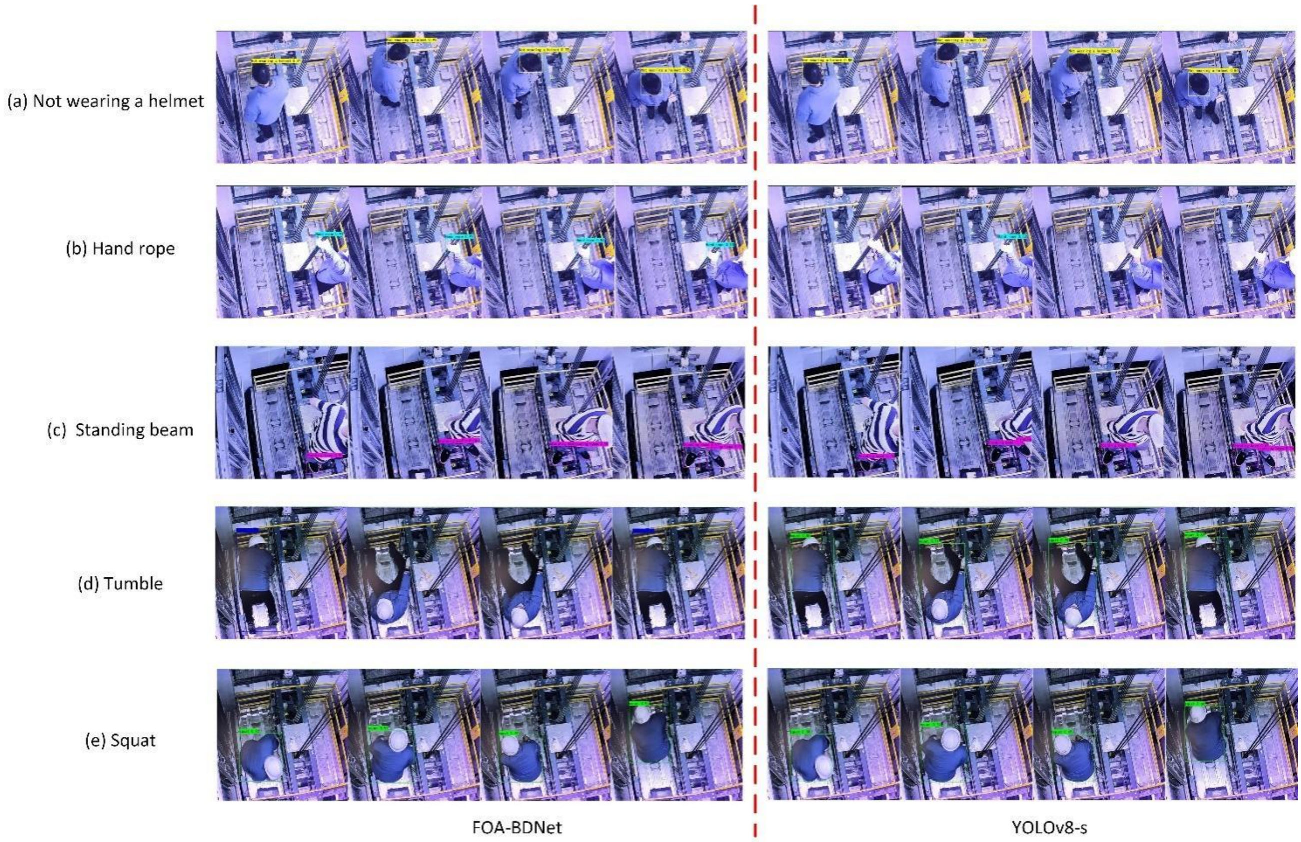


Figure 10. Comparison of the effectiveness of FOA-BDNet and YOLOv8-s in detecting different behaviors.

3.2.4. Fall behavior detection

In the previous section, we tested the omission and false detection cases between the proposed algorithm and the advanced YOLOv8-s algorithm in the "Tumble" behavior detection, which may be due to the similarity between the appearance of the falling and normal squatting behaviors in an overlooking angle larger than 45^0. It is difficult for a CNN to distinguish the difference between the two by extracting the edge, size, color, texture, and other features of the behavior. In addition, due to the translation invariance of a CNN, it is difficult to distinguish different behaviors by the change of position. Therefore, it is necessary to design an auxiliary behavior discrimination algorithm and combine the target size predicted by the positioning frame of the deep target detection model to assist in judging the attributes of the behavior. The specific methods have been introduced in detail in Section 2.4. The effect of this paper's FOA-BDNet model on detecting fall behaviors before and after adding the auxiliary algorithm is shown in Figure 11.

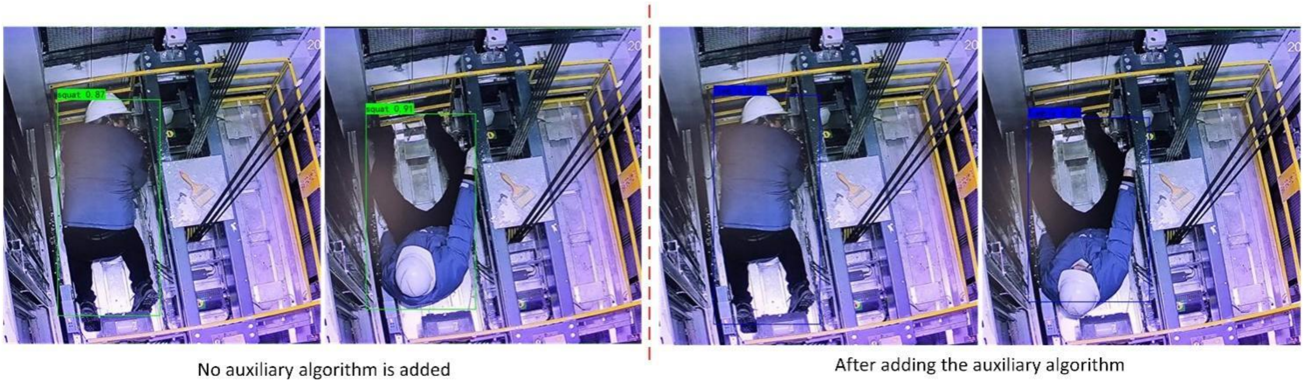

squat 0.87
squat 0.91
No auxiliary algorithm is added
After adding the auxiliary algorithm

Figure 11. Comparison of detection effect of FOA-BDNet model before and after incorporating fall assist algorithm.

As seen in Figure 11, the FOA-BDNet model significantly improves the generalization and robustness of fall behavior detection after adding the auxiliary algorithm for the squatting and falling behaviors. Before adding the auxiliary algorithm, the FOA-BDNet model will misdetect the fall behavior as the squatting behavior. After adding the auxiliary algorithm, the same test image can accurately detect the fall behavior. When the maintenance personnel falls, the size of the human body in the video image will become longer while the surveillance video equipment remains relatively stationary, and the size of the human body during the normal squatting behavior is smaller than that during the falling behavior. Therefore, the target detection algorithm must first detect the human body, and correctly select the longest and shortest boundaries of the human body. Then, according to the obtained predicted length of the box, which is combined with the auxiliary algorithm, the normal squatting and falling behaviors can be correctly distinguished.

3.2.5. Ablation experiment

Focusing on the task of behavior the detection for elevator maintenance personnel, this paper proposes three main innovative designs based on the YOLO architecture. First, a lightweight strategy for the backbone feature extraction network, which is referred to as Model a, is introduced to meet the real-time requirements of ordinary edge devices. This design aims to enhance the model’s operational efficiency without significantly reducing the detection accuracy, thus enabling timely monitoring and alerts driven by the underlying algorithms. Second, a ‘distant interaction and close connection’ feature fusion strategy is employed at different feature levels, referred to as Model b, which combines shallow granular information such as color, edges, and textures with deep high-level semantic information. This enhances the model's ability to distinguish the subtle detail features. Finally, to effectively differentiate the easily confused behaviors of ‘Squat’ and ‘Tumble’, a state-assisted calculation algorithm for these two behaviors has been designed. This method has been validated as effective in Section 3.2.4, so the ablation study will only need to verify the effectiveness of Models a and b. YOLOv8-s is currently the most advanced first-order deep object detector, and in the ablation study, YOLOv8-s will serve as the baseline model. The performance of the models will be trained and tested on the experimental platform described in Section 3.1, thereby following the hyperparameter settings in Table 1. To comprehensively discuss the time complexity and detection performance of the proposed algorithm, metrics such as mAP, params, FLOPs, and mFPS will be calculated for the different algorithms. The model's computational load (FLOPs) and parameter count (Params) are two important indicators for measuring the complexity of deep learning algorithms, which can be used to evaluate a model’s performance and practicality. Building upon the YOLOv8-s architecture, the innovative designs of Models a and b, along with the auxiliary algorithm optimization, contribute to the improved FOA-BDNet model, which is labeled as Ours. The results of the ablation study are presented in Table 2.

Table 2. FOA-BDNet network model ablation experiments.

Model	mAP（%）	mFPS	Params(M)	FLOPs(B)
Standard	94.27	42.39	11.2	28.6
a	93.86	70.29	9.3	24.8
b	96.16	41.69	13.1	29.2
Ours	98.68	69.51	11.2	27.8

By analyzing the experimental results in Table 2, it is evident that the baseline model YOLOv8-s performs exceptionally well in the task of detecting maintenance personnel behaviors, thereby achieving an mAP of 94.27% and an mFPS of 42.39 fps. When the backbone feature extraction part is replaced with Model a, there is a slight decrease in the average detection accuracy, though there is a significant increase for the average mFPS. Additionally, the reductions in both params and FLOPs indicate a decrease in the model's computational complexity, thus confirming that the lightweight effect of Model a has taken effect. Although the model's accuracy is slightly compromised, the increased operational speed is beneficial for deployment on ordinary edge devices. When Model b is added to the baseline model, there is a notable improvement in the average detection accuracy (mAP), with detection accuracies for behaviors such as ‘Not wearing a helmet’, ‘Squat’, and ‘Grasping the steel cable’ even reaching 100%. However, the params and FLOPs metrics increased, indicating that the feature fusion algorithm added complexity to the model, thus leading to a slight decrease in FPS of only 0.7 fps. This suggests that Model b’s feature fusion algorithm plays a positive role in extracting similar detail features, even though it also increases the computational load. The model ‘Ours’ represents the FOA-BDNet model proposed in this paper. Compared to the baseline model (YOLOv8-s), both the mAP and mFPS metrics improved, while the scale of params remained roughly the same and FLOPs showed a slight decrease. This indicates that the lightweight modifications and feature fusion structure can mutually enhance each other, thereby reducing the model's computational complexity while simultaneously improving the robustness and the generalization performance. FOA-BDNet outperforms YOLOv8-s in both accuracy and operational efficiency for the task of detecting the elevator maintenance personnel behaviors, thus suggesting that the proposed FOA-BDNet model is more suitable for this task.

3.2.6. Performance comparison of FOA-BDNet and human abnormal behavior detection

Human abnormal behavior detection technologies are an important branch of computer vision. For different application scenarios, the distribution of positive and negative samples of abnormal behavior data may be biased. At the same time, the results of different methods for the determination of the same abnormal behavior in the same scene are not the same. However, the general objective performance indexes obtained from algorithm testing on common datasets can evaluate the performance of most behavior recognition and detection algorithms. In the field of behavioral anomaly detection, the commonly used evaluation metrics such as AP are not sufficient to reflect the model performance, while the Receiver Operating Characteristic (ROC) curve is more robust to the unbalanced data, and can better evaluate the classification ability of the model. The area under the curve (AUC) is defined as the area under the ROC curve, and the value of the AUC ranges from 0 to 1. The closer the AUC is to 1, the better the performance of the model is under different thresholds. Therefore, the objective AUC index is used to measure the performance of different models. USCD [26] is the most commonly used classical human behavior detection dataset, which contains a total of 98 videos, and is constructed according to the movement direction of the crowd and scene disturbing factors in the video frame. Different directions and scene interference factors are constructed as USCD Pes1 and USCD Ped2. Table 3 records the performance of some different model algorithms on the typical abnormal behavior detection dataset.

According to the recorded data in Table 3, the AUC values tested by the proposed algorithm on the classical USCD Pes1 and USCD Ped2 abnormal behavior data sets reached 94.3% and 95.3%. respectively; however, it did not exceed the best AUC values recorded by other mainstream abnormal behavior detection algorithms. However, it is very close to the 97.4% achieved by “GAN” in the USCD Pes1 dataset and 97.8% achieved by “Object-centric AE” in the USCD Ped2 dataset, thereby considering that the deep model performance is also intrinsically related to hyperparameter settings. The performance of this model on the classical data set of human abnormal behavior detection has been in the ranks of good algorithms. Because FOA-BDNet is based on the YOLO architecture, the recently released YOLO series models YOLOv8-s have reached maturity. According to the test results of USCD Pes1 and USCD Ped2 data sets, the AUC index of the proposed algorithm slightly increased by 0.5% and 0.7%, respectively, compared with that of the YOLOv8-s model. Although the difference is not big, it also shows that the algorithm in this paper has a relatively good generalization performance in the field of human abnormal behavior detection.

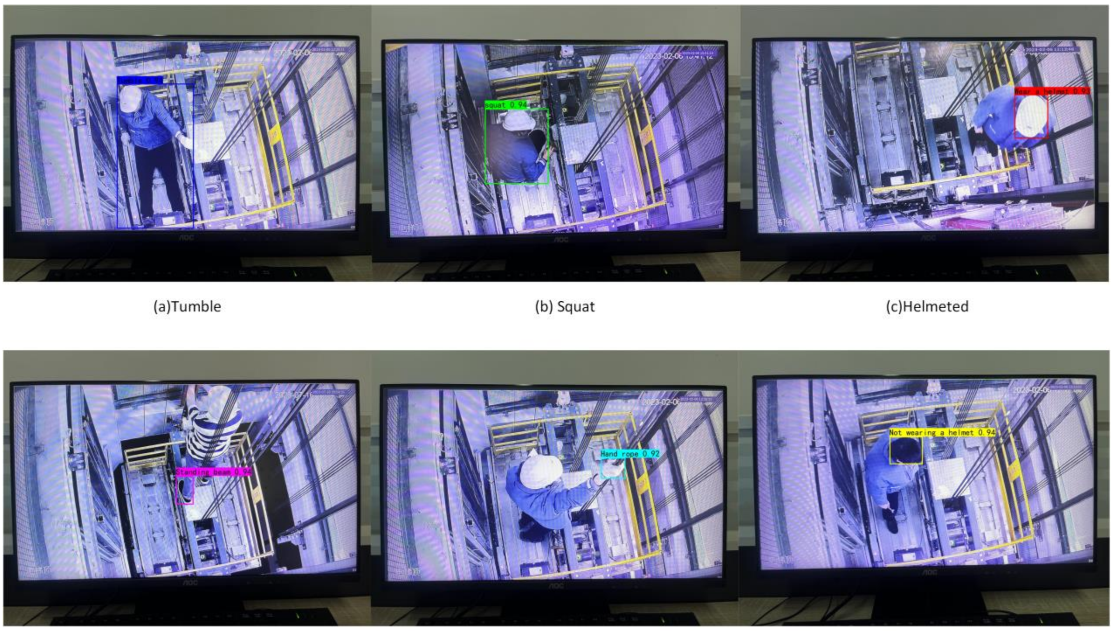
Table 3. Comparison of frame-level AUC of selected algorithms on typical abnormal behavior datasets (%).

Model	USCD Pes1	USCD Ped2
Social Force[27]	67.5	55.6
MPPCA[28]	66.8	69.3
Conv AE[29]	81.0	81.1
ConvLSTM AE[30]	75.5	88.1
Spatiotemporal AE[31]	89.9	87.4
GrowingGas[32]	93.8	94.1
GAN[33]	97.4	93.5
Multi-task Fast R-CNN[34]	-	92.2
Unmasking[35]	68.4	82.2
Stacked RNN[36]	-	92.2
U-Net[37]	83.1	95.4
AMDN[38]	92.1	90.8
Plug-and-play CNN[39]	95.7	88.4
Object-centric AE[40]	-	97.8
Adversarial 3D Conv AE[41]	95.7	96.0
YOLOv8-s	93.8	94.6
FOA-BDNet（Ours）	94.3	95.3

3.2.7. Field testing

To demonstrate the practical performance of the proposed method to detect maintenance personnel behaviors in elevator shafts, the algorithm was deployed on-site, thereby utilizing existing surveillance cameras to provide video streams for the algorithm. The real-time detection results are shown in Figure 12.

Figure 12 displays the captured real-time detection footage, where the proposed FOA-BDNet model was deployed in a real elevator scenario. The experimental personnel entered the top of the elevator cabin to simulate various behaviors of maintenance personnel. It can be observed that the FOA-BDNet algorithm successfully detects all behaviors with a high confidence and without any lag, thus meeting the practical requirements for real-time and accurate detection of unsafe behaviors by maintenance personnel in elevators.



(a)Tumble (b) Squat (c)Helmeted

(d) Standing beam (e) Hand rope (f) Not wearing a helmet

Figure 12. FOA-BDNet real-time detection effect.

4. Affirmation of conclusions and limitations

With the popularization of elevators, elevator maintenance has become frequent and important; however, more and more safety accidents frequently occur in the process of multidimensional maintenance, thereby endangering the life safety of maintenance personnel and the safe operation of elevators. This paper designed a first-order deep object detection model (FOA-BDNet) to detect unsafe behaviors from the perspective of identifying and detecting the maintenance personnel behaviors. Through the lightweight design of the model, the average running speed of 69.51 FPS was obtained, which allowed the model to meet the real-time detection requirements of the video environment. The feature fusion algorithm was designed to enhance the recognition ability of the model with small differences, and the average accuracy of the behavior detection was 98.68%, this included behaviors such as not wearing a helmet, standing on a beam, and grasping a wire cable by the hand. In the case that the depth target detector failed to recognize the squat and fall behaviors, a behavior-assisted computing algorithm was designed to improve the model's ability to recognize the fall behavior. The FOA-BDNet algorithm proposed in this paper meets the need for real-time and high-precision detection of the unsafe behaviors of elevator maintenance personnel and provides some useful technical references for current research and engineering applications in this field.

Author contributions

Zengming Feng completed the conceptualization, software, funding acquisition, validation, and methodology for this paper and participated in writing the original draft of the manuscript; Tingwen Cao completed data curation, resource, supervision, investigation, visualization, project administration, and finally reviewed and edited the manuscript. All authors have read and approved the final version of the manuscript for publication.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. D. C. Balmer, Impact of the A18. 1 ASME Standard on platform lifts and stairway chairlifts on accessibility and usability, *Assist. Technol.*, **22** (2010), 46−50. https://doi.org/10.1080/10400430903520264
2. C. Cheng, S. Zhang, Z. Wang, L. Qiu, L. Tu, L. Zhu, et al., Surrogate-model-based dynamic-sensing optimization method for high-speed elevator car horizontal vibration reduction, *Proc. I. Mech. Eng. Part C*, 2024. https://doi.org/10.1177/09544062231217926
3. P. C. Liao, Z. Guo, T. Wang, J. Wen, C. H. Tsai, Interdependency of construction safety hazards from a network perspective: A mechanical installation case, *Int. J. Occup. Saf. Ergo.*, **26** (2020), 245−255. https://doi.org/10.1080/10803548.2018.1426272
4. J. Lei, W. Sun, Y. Fang, N. Ye, S. Yang, J. Wu, A model for detecting abnormal elevator passenger behavior based on video classification, *Electronics*, **13** (2024), 2472. https://doi.org/10.3390/electronics13132472
5. H. Hasegawa, S. Aida, *Elevator monitoring system to guide user's behavior by visualizing the state of crowdedness*, In: Lee, R. (eds) Big Data, Cloud Computing, and Data Science Engineering, Springer, Cham, 2020, 85−98. https://doi.org/10.1007/978-3-030-24405-7_6
6. S. Liang, D. Niu, K. Huang, H. Wu, L. Ding, Y. Yang, An elevator door blocking behavior recognition method based on two-stage object detection networks, *IEEE*, 2022, 1374−1378. https://doi.org/10.1109/YAC57282.2022.10023898
7. Z. Wang, J. Chen, P. Yu, B. Feng, D. Feng, SC-YOLOv8 network with soft-pooling and attention for elevator passenger detection, *Appl. Sci.*, **14** (2024), 3321. https://doi.org/10.3390/app14083321
8. S. Chai, X. I. Li, Y. Jia, Y. He, C. H. Yip, K. K. Cheung, et al., A non-intrusive deep learning based diagnosis system for elevators, *IEEE Access*, **9** (2021), 20993−21003. https://doi.org/10.1109/ACCESS.2021.3053858
9. S. C. Lai, M. L. Yang, R. J. Wang, J. Y. Jhuang, M. C. Ho, Y. C. Shiau, Remote-control system for elevator with sensor technology, *Sensor. Mater.*, **34** (2022). https://doi.org/10.18494/SAM3827
10. Z. Li, J. Ning, T. Li, Design of non-intrusive online monitoring system for traction elevators, *Appl. Sci.*, **14** (2024), 4346. https://doi.org/10.3390/app14114346
11. W. Yao, A. Wang, Y. Nie, Z. Lv, S. Nie, C. Huang, et al., Study on the recognition of coal miners' unsafe behavior and status in the hoist cage based on machine vision, *Sensors*, **23** (2023), 8794. https://doi.org/10.3390/s23218794
12. T. Kong, W. Fang, P. E. D. Love, H. Luo, S. Xu, H. Li, Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction, *Adv. Eng. Inform.*, **50** (2021), 101400. https://doi.org/10.1016/j.aei.2021.101400
13. M. Casini, Extended reality for smart building operation and maintenance: A review, *Energies*, **15** (2022), 3785. https://doi.org/10.3390/en15103785
14. R. D'Souza, *IoT and the future of elevator maintenance business*, Master Thesis, Technische Universität Wien, 2022. https://doi.org/10.34726/hss.2022.103532
15. X. P. Zhang, J. H. Ji, L. Wang, Z. He, S. Liu, A review of video-based human abnormal behavior recognition and detection methods, *Control Decis.*, **37** (2022), 14−27.

16. A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, *IEEE T. Pattern Anal.*, **23** (2001), 257−267. https://doi.org/10.1109/34.910878
17. H. Wang, A. Kläser, C. Schmid, C. L. Liu, Dense trajectories and motion boundary descriptors for action recognition, *Int. J. Comput. Vis.*, **103** (2013), 60−79. https://doi.org/10.1007/s11263-012-0594-8
18. L. Xu, C. Gong, J. Yang, Q. Wu, L. Yao, Violent video detection based on MoSIFT feature and sparse coding, *IEEE*, 2014, 3538−3542. https://doi.org/10.1109/ICASSP.2014.6854259
19. H. Fujiyoshi, A. J. Lipton, T. Kanade, Real-time human motion analysis by image skeletonization, *IEICE T. Inf. Syst.*, **87** (2004), 113−120.
20. M. S. Alzahrani, S. K. Jarraya, H. Ben-Abdallah, M. S. Ali, Comprehensive evaluation of skeleton features-based fall detection from Microsoft Kinect v2, *Signal, Image Video P.*, **13** (2019), 1431−1439. https://doi.org/10.1007/s11760-019-01490-9
21. Z. Liao, H. Hu, J. Zhang, C. Yin, Residual attention unit for action recognition, *Comput. Vis. Image Und.*, **189** (2019), 102821. https://doi.org/10.1016/j.cviu.2019.102821
22. C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, 1933−1941. https://doi.org/10.1109/CVPR.2016.213
23. S. Sudhakaran, O. Lanz, Learning to detect violent videos using convolutional long short-term memory, *IEEE*, 2017, 1−6. https://doi.org/10.1109/AVSS.2017.8078468
24. C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, I. H. Yeh, *CSPNet: A new backbone that can enhance learning capability of CNN*, Montreal, BC, Canada, 2020, 390−391.
25. X. Hu, D. Kong, X. Liu, J. Zhang, D. Zhang, FM-STDNet: High-speed detector for fast-moving small targets based on deep first-order network architecture, *Electronics*, **12** (2023), 1−15. https://doi.org/10.3390/electronics12081829
26. S. Wang, Z. Miao, Anomaly detection in crowd scene, *IEEE*, 2010, 1220−1223. https://doi.org/10.1109/ICOSP.2010.5655356
27. R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, *IEEE*, 2009, 935−942. https://doi.org/10.1109/CVPR.2009.5206641
28. J. Kim, K. Grauman, Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates, *IEEE*, 2009, 2921−2928. https://doi.org/10.1109/CVPR.2009.5206569
29. M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, L. S. Davis, Learning temporal regularity in video sequences, *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, 733−742. https://doi.org/10.1109/CVPR.2016.86
30. W. Luo, W. Liu, S. Gao, Remembering history with convolutional lstm for anomaly detection, *IEEE*, 2017, 439−444. https://doi.org/10.1109/ICME.2017.8019325
31. Y. S. Chong, Y. H. Tay, *Abnormal event detection in videos using spatiotemporal autoencoder*, Springer, Cham, 2017, 189−196. https://doi.org/10.1007/978-3-319-59081-3_23
32. Q. Sun, H. Liu, T. Harada, Online growing neural gas for anomaly detection in changing surveillance scenes, *Pattern Recogn.*, **64** (2017), 187−201. https://doi.org/10.1016/j.patcog.2016.09.016
33. M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe, Abnormal event detection in videos using generative adversarial nets, *IEEE*, 2017, 1577−1581. https://doi.org/10.1109/ICIP.2017.8296547
34. R. Hinami, T. Mei, S. Satoh, Joint detection and recounting of abnormal events by learning deep generic knowledge, *Proc. IEEE Conf. Comput. Vis.*, 2017, 3619−3627. https://doi.org/10.1109/ICCV.2017.391

35. R. I. Tudor, S. Smeureanu, B. Alexe, M. Popescu, Unmasking the abnormal events in video, *Proc. IEEE Conf. Comput. Vis.*, 2017, 2895−2903.
36. W. Luo, W. Liu, S. Gao, A revisit of sparse coding based anomaly detection in stacked RNN framework, *Proc. IEEE Conf. Comput. Vis.*, 2017, 341−349. https://doi.org/10.1109/ICCV.2017.45
37. W. Liu, W. Luo, D. Lian, S. Gao, Future frame prediction for anomaly detection—a new baseline, *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, 6536−6545. https://doi.org/10.1109/CVPR.2018.00684
38. D. Xu, Y. Yan, E. Ricci, N. Sebe, Detecting anomalous events in videos by learning deep representations of appearance and motion, *Comput. Vis. Image Und.*, **156** (2017), 117−127. https://doi.org/10.1016/j.cviu.2016.10.010
39. M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, N. Sebe, Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection, *IEEE*, 2018, 1689−1698. https://doi.org/10.1109/WACV.2018.00188
40. R. T. Ionescu, F. S. Khan, M. I. Georgescu, L. Shao, Object-centric auto-encoders and dummy anomalies for abnormal event detection in video, 2019, 7842−7851. https://doi.org/10.1109/CVPR.2019.00803
41. C. Sun, Y. Jia, H. Song, Y. Wu, Adversarial 3d convolutional auto-encoder for abnormal event detection in videos, *IEEE T. Multimedia*, **23** (2020), 3292−3305. https://doi.org/10.1109/TMM.2020.3023303