Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061823

ARTICLE

YOLO-AB: A Fusion Algorithm for the Elders' Falling and Smoking Behavior Detection Based on Improved YOLOv8

Xianghong Cao, Chenxu Li* and Haoting Zhai

College of Building Environment Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450066, China

*Corresponding Author: Chenxu Li. Email: lichenxu@zzuli.edu.cn

Received: 04 December 2024; Accepted: 07 March 2025; Published: 19 May 2025

ABSTRACT: The behavior safety testing of more and more elderly people living alone has become a hot research topic along with the arrival of an aging society. A YOLO-Abnormal Behaviour (YOLO-AB) algorithm for fusion detection of falling and smoking behaviors of elderly people living alone has been proposed in this paper, which can fully utilize the potential of the YOLOv8 algorithm on object detection and deeply explore the characteristics of different types of behaviors among the elderly, to solve the problems of single detection type, low fusion detection accuracy, and high missed detection rate. Firstly, datasets of different types of elderly behavior images such as falling, smoking, and standing are constructed for performance validation of subsequent algorithms. Secondly, the Content-Aware Reassembly of Features Module (CARAFE) is introduced into the YOLOv8 algorithm to enhance content perception, strengthen feature fusion, generate adaptive kernels dynamically, and reduce parameters effectively. Then, the Large Selective Kernel Network (LSKNet) module is added to the backbone network part to strengthen the framing of human targets and improve detection accuracy. Next, the Focal-SCYLLA-IOU (F-SIOU) loss function is used to improve the positioning accuracy of the edge part of the target detection frame. Finally, YOLO-AB and other different algorithms are tested and compared using the falling dataset, the smoking dataset, and the falling and smoking mixed dataset, respectively. The results show that the detection accuracy of the YOLO-AB algorithm is 0.93 on the falling dataset alone, 0.864 on the smoking dataset alone, and 0.923 on the falling and smoking mixed dataset, all of which are better than those of the other algorithms. The performance of YOLO-AB is better than that of YOLOv8 on multiple metrics, such as 4.1% improvement in the mAP50 index, 4.9% increase in the P index, and 3.5% boost in the R index, which verifies the effectiveness of the algorithm.

KEYWORDS: Abnormal behavior of the elderly; feature fusion; deep learning; YOLOv8

1 Introduction

At present, with the increasing severity of the global population aging problem, the number of elderly people living alone is on the rise. The elderly have a high incidence of accidents due to decreased physical functions and self-care abilities. Especially when an accident occurs, it is difficult for the elderly living alone to get help in time due to the lack of specialized personnel, which may lead to tragedy [1]. Therefore, the safety monitoring and health management of the elderly are becoming increasingly crucial.

According to the World Health Organization's Global Report on Fall Prevention in Older Adults, about 30% of people aged 60 years and older have at least one accidental fall each year. If the elderly can be detected at the first time of falling and sent to the hospital for treatment in time, the risk of death can be reduced greatly [2]. The elderly who live alone will try to stand on their own after a fall, which is likely to cause a

second fall, resulting in aggravated injuries and even life-threatening injuries. Therefore, the detection of accidental falls of the elderly has become an important means of safe monitoring of the elderly.

At the same time, the number of people who die from smoking in the world has reached a staggering 60 million every year. Smoking of some older people who have underlying medical conditions can lead to severe coughing and bring sudden risks such as heart or lung disease or cancer, which seriously threaten the health of the elderly. Life-threatening conditions can occur in severe cases. Secondhand smoke can also be potentially harmful to those around you. Therefore, the detection of smoking behavior of the elderly can be used as an important means to monitor the health status of the elderly.

Studies related to the detection of abnormal falling behavior in older adults are the most common. Abnormal smoking behavior in the elderly is relatively scarce. And the detection of the fusion of these two abnormal behaviors is almost non-existent.

Typically, independent designs are adopted. For example, one algorithm framework is used for falling detection, and another is for smoking detection. The synergy between the two is quite weak. Moreover, since only a single behavior is processed, there may be false alarms.

Therefore, a YOLO-Abnormal Behaviour (YOLO-AB) algorithm, which is based on the fusion detection of two types of abnormal behaviors of the elderly, namely falling and smoking, has been proposed in this paper to fully protect the life and health safety of the elderly, especially those living alone. The YOLO-AB algorithm integrates the two behaviors of falling and smoking into one. These two actions are incorporated within a unified framework, using the same network structure. Sharing the feature extraction network, enables the simultaneous detection and classification of behaviors such as falling and smoking. This approach leads to excellent generalization capabilities for these two behaviors during subsequent end-to-end training.

In this study, the object detection ability and behavior analysis ability of YOLO are innovatively combined to detect locations and recognize specific behaviors simultaneously. By means of feature fusion extraction and context information analysis, similar behaviors can be better differentiated, effectively reducing the false alarm rate.

By improving the backbone network, neck network, and loss function of the YOLOv8 algorithm, the detection accuracy of the algorithm and the detection accuracy of the two abnormal behaviors of falling and smoking are improved. The main contributions of this study are as follows:

(1) A total of 9018 experimental images have been labeled to construct a behavioral dataset for subsequent algorithm testing, including 4576 falling pictures, 2482 smoking pictures, and 1960 standing pictures, forming a mixed dataset to support the algorithm performance evaluation of multi-task detection.
(2) The Content-Aware Reassembly of Features Module (CARAFE) has been used to replace the neck network structure of YOLOv8 algorithm, which can reduce the model parameters and complexity, and strengthen feature fusion.
(3) The Large Selective Kernel Network (LSKNet) module has been added to improve the feature learning ability of the model, which can increase the accuracy of target framing, expand the acceptance field, and improve the detection accuracy.
(4) The Focal-SCYLLA-IOU (F-SIOU) loss function has been introduced to adjust the parameters dynamically, reduce the weight of background and increase the weight of edges and blurred areas, to achieve more accurate positioning.
(5) The effectiveness of YOLO-AB has been verified by testing and comparing with other algorithms on the falling dataset, the smoking dataset and the falling and smoking mixed dataset, respectively.

The rest of the jobs of this paper are as follows. Related work on abnormal behavior detection in the elderly is introduced in Section 2. The YOLO-AB algorithm has been described in detail in Section 3. The

comparative experiments of the YOLO-AB algorithm and other algorithms have been analyzed in Section 4. And the conclusions are given in Section 5.

2 Related Work

2.1 Abnormal Behavior Detection Based on Traditional Methods

Early anomalous behavior detection relied on manual observation and feature design, which included analysis based on spatial-temporal points of interest, contour silhouettes, and motion trajectories. Some researchers have proposed Harris 3D points of interest by using the feature recognition method of Harris corners in the spatial dimension, which has expanded the detection ability of spatio-temporal points of interest [3–5]. This method is limited in dynamic environments because Harris 3D points alone cannot effectively capture complex motion features.

Dollar et al. used Gaussian filter and Gabor filter to solve the problem of insufficient detection of the number of points of interest by designing the spatial and temporal dimensions [6].

Wang et al. used the grid and block method to establish the correspondence with the grid, and the field pixels of local observations have been covered in the block [7].

An improved dense grid based on spatiotemporal points of interest is proposed by many authors which is applied to the moving image for behavioral feature extraction by comparing various local descriptors, such as the histogram of oriented gradients (HOG) [8], the boundary motion history map (MHI) [9], and the directional optical flow histogram (HOF), which reduces the influence of time and space on local image changes.

These methods are time-consuming and costly, and the extracted feature maps are not clear, which limits the scope of their use.

With the continuous development of wearable sensor technology and machine learning, researchers use wearable sensors combined with machine learning to detect abnormal behaviors, with the tester wearing the sensor, and judge whether there is an abnormality based on the data values collected by the sensor [10].

Attal et al. used three inertial sensor units, which were worn at the subject's upper and lower limb key points, and the resulting data were processed with a supervised classification algorithm to achieve abnormal behavior detection [11].

Koutli et al. elaborated motion sensor signals to generate contextual indicators, and detected index abnormalities in the elderly by classification and regression methods [12].

These methods exhibit varying degrees of robustness across different environments. Specifically, in complex backgrounds, the accuracy drops significantly. Moreover, the elderly tend to be reluctant to wear wearable devices, often forgetting to do so, which further restricts their application.

2.2 Abnormal Behavior Detection Based on Deep Learning

With the development of deep learning, researchers have begun to use convolutional neural networks for abnormal behavior detection, and the common methods used for abnormal behavior detection in the elderly mainly include R-CNN, Faster R-CNN, LSTM, YOLO, SSD, etc. [13,14]. These methods usually adopt an end-to-end training method, which directly learns the mapping of the position and category of the image, and combines with edge regression technology to achieve object detection and localization, which is widely used in abnormal behavior detection.

Arifoglu et al. explored the use of CNNs to simulate patterns in activity sequences and detect abnormal behaviors such as falls in older adults, considering activity recognition as a sequence labeling problem that is

labeled based on deviations from normal behavior [15]. This method has certain effects in detecting abnormal behavior in the elderly, but it has limitations in the depth and breadth of feature extraction.

Li et al. proposed an MSCS-DenseNet-LSTM anomaly detection model for the elderly based on a multi-scale attention mechanism, which introduced multi-scale features and attention mechanisms, which effectively enhanced the generalization ability of the model and made its performance more stable in complex scenes [16].

He et al. proposed a spatio-temporal feature fusion (SCGAT)-based anomalous behavior recognition method for the elderly, which uses a graph attention neural network to extract non-Euclidean spatial features in sensor signals, and combined with the time-domain features extracted by time series coding network, an accurate classification model is constructed, which is suitable for the fusion and classification of multi-dimensional behavioral features [17].

In order to protect the privacy of the elderly, the authors' team has proposed an accidental fall detection algorithm for the elderly at home, which first uses infrared video images to collect the behavior of the elderly, fits them in ellipses, extracts five geometric feature variables, and then uses the above variables as input for feature extraction and classification to establish an LSTM model, and finally obtains a good correct classification rate and effectively protects the privacy of the elderly [18].

In order to solve the problem of feature information loss and performance degradation in complex environments, the authors' team also proposed a human pose recognition model based on pose estimation-convolutional neural network-bidirectional gated recurrent unit (BiGRU), which combines the deep convolutional network with the Mediapipe framework to extract the key points of the human skeleton, and uses the double-layer BiGRU algorithm and CNN network to perform deep convolution to extract human posture features and detect postures such as climbing, falling, and abdominal pain. Good generalization and robustness were obtained [19].

Although the above studies have made some progress in the accuracy and robustness of abnormal behavior detection, there are still problems of insufficient feature fusion in image-based detection, and the classification of abnormal behaviors has not yet reached a high degree of accuracy. In order to solve this problem, this study mainly divides abnormal behaviors into two categories of falling and smoking, and this section analyzes and summarizes the research progress of these two categories.

2.2.1 Fall-Based Detection Model

In recent years, with the continuous progress of deep learning technology, researchers have gradually applied deep learning algorithms to the field of fall detection in the elderly, and have achieved remarkable results.

Yan et al. proposed a fall recognition method based on skeleton and spatiotemporal convolutional network (ST-GCN), which uses the inertial measurement unit (IMU) to obtain body joint data, construct a human skeleton model, and extract the motion characteristics of human falls from the spatial and temporal dimensions, so as to distinguish daily activities from falling behaviors [20]. This method can accurately identify fall events by capturing the spatiotemporal characteristics of human dynamic changes.

Li et al. proposed an enhanced ResNet-based three-dimensional fall detection method, which solves the gradient explosion problem by directly extracting spatiotemporal features, reducing parameters and increasing network depth [21]. This method improves the computational efficiency on the premise of ensuring the detection accuracy, and makes the performance of the model more stable in complex scenes.

McCall et al. proposed a new transformer model that extracts 2D poses via a pose extractor and incorporates transfer learning techniques that freeze most of its layers, fine-tune only the last few layers, and use a relatively small dataset for fall detection and prediction tasks [22].

Soni et al. proposed an adaptive two-stage deep learning network (CABMNet) to optimize spatial features by CNN and CBAM, multi-head attention (MHA) enhanced bidirectional long short-term memory (Bi-LSTM) for time analysis, thereby improving the overall performance and efficiency of fall detection, and also integrated a Kalman filter to enhance the noise reduction effect in the preprocessing stage [23].

Li et al. proposed a fall detection method based on future frame prediction, which regards the attention U-Net with flexible global aggregation blocks as a frame prediction network, realizes multiple video frames to predict the next future frame, and combines the common appearance constraints of intensity and gradient with motion constraints to generate higher quality frames and improve the performance of the prediction network [24].

Kang et al. proposed a fall detection algorithm based on YOLO-KCF to identify and respond to the critical situation of elderly people living alone in real time, and detect falls by the change of the shape of the bounding box between the standing type and the falling behavior type [25].

Qin et al. proposed a new high-precision fall detection model ESD-YOLO based on dynamic convolution, which uses Convolution 2D version 3 (C2Dv3) instead of the Cross Stage Partial 2 with Feedforward (C2f) module to enhance the network's ability to capture complex details and deformations, and uses DyHead to unify multiple attention operations, which improves the detection accuracy of targets at different scales and improves the performance in occlusion cases [26].

Zhao et al. proposed an enhanced YOLOv7-fall model, which uses a novel attention module Spatial Depth Integration (SDI) to enhance feature extraction capabilities. The use of Group Shuffle Convolution (GSConv) and Visual Geometry Group-Group Shuffle Cross Stage Partial (VoV-GSCSP) modules is to reduce model parameters and computational complexity, and this method enhances the diversity and robustness of the network by capturing features from different layers of the image, effectively improving the accuracy and efficiency of fall detection [27].

2.2.2 Detection Module Based on Smoking

With the development of deep learning technology, researchers have proposed a variety of innovative methods for the detection of smoking behavior.

Ma et al. designed a smoking image detection model, YOLO-Cigarette, based on deep learning, which fused the new fine-grained spatial pyramid pooling module (FSPP) and multi-spatial attention mechanism (MSAM) to improve the original YOLOv5 network structure, which effectively solved the problem of low detection accuracy caused by small cigarette targets and complex background environment [28].

Senyurek et al. proposed a method to detect smoking events based on gesture regularity [29]. In this method, the regularity of gestures is monitored by a single-axis accelerometer worn on the dominant wrist, and a novel unbiased autocorrelation method is applied to process the time series of gestures and quantify the regularity score, so as to realize the recognition of smoking behavior. This method takes advantage of the portability of the wearable device and can effectively identify the smoking action to a certain extent.

Cho et al. detected typical gases collected from Internet of Things (IoT) sensors and then used machine learning to detect cigarettes with an analysis and machine learning approach, and the results showed that the SVM model showed the best performance [30].

Dong et al. designed and improved the YOLOv4 real-time smoking detection model, through k-means++ clustering, recalculated and designed the prior box, enhanced the adaptability of the scale and obtained better anchors [31]. An attention mechanism has been added to improve the accuracy of cigarette detection.

Wang et al. proposed a new smoke detection algorithm, YOLOv8-MNC, which uses a dedicated layer for a small target and uses the Normalized Wasserstein Distance (NWD) loss function to reduce the impact of small deviations in the position of objects on IoU, so as to improve the training accuracy [32]. However, the algorithm is affected by the dataset. In the case of multi-scale targets, the feature fusion for smoking scenarios is insufficient, which impacts the detection accuracy and stability.

It can be seen that most abnormal behavior detection methods of the elderly are to detect the falling or smoking behavior separately. And the fusion detection of these two abnormal behaviors has not yet been realized. When conducting concurrent detection of falling and smoking behaviors, it is imperative to detect both the falling movement (a large-scale target) and the smoking gesture (a small-scale target). Regarding the YOLOv8 algorithm [32], during the independent detection of large-scale targets (falls) and small-scale targets (smoking), the variance in target dimensions engenders difficulties. Throughout the feature extraction phase, it falls short of comprehensively retrieving all requisite features, frequently disregarding small-scale targets. Such negligence gives rise to inadequate amalgamation of the extracted features, ultimately culminating in a comparatively low precision of the detected results.

YOLO-AB algorithm based on improved YOLOv8 combine falling and smoking behavior has been proposed in this paper, which can effectively solve the problems of insufficient feature extraction, imperfect feature fusion and detection accuracy of the existing models. Specific improvements of YOLOv8 include:

(1) Improve the backbone network structure to improve feature extraction capabilities and identify behavioral details more effectively.
(2) Replace the CARAFE module in the neck network to enhance the feature fusion and make the fusion process more sufficient.
(3) Replace the loss function to reduce the impact of local loss and improve the overall detection accuracy.

Therefor the YOLO-AB algorithm can not only detect falling and smoking behaviors at the same time, but also surpass the existing methods in detection accuracy and efficiency, which provides a new solution for multi-behavior anomaly detection.

3 YOLO-AB Algorithm Based on Improved YOLOv8

3.1 YOLO-AB Overview

YOLOv8 is an algorithm model of the YOLO family, which provides a new SOTA model and consists of four parts, namely the input network part, the backbone network part, the neck network part and the head network part [26]. Improvements have been made to the backbone network part, the neck network part, and the loss function part of the original YOLOv8 network model, which has been named YOLO-AB algorithm and applied to the detection of fusion behaviors of falling and smoking. The network structure of YOLO-AB is shown in Fig. 1.

In order to accurately detect abnormal behaviors, a large selective kernel network (LSKNet) is used in YOLO-AB, which dynamically adjusts its large spatial sensing field and accurately frames the target. The function of nn.Upsample is to perform feature fusion on the input image. CARAFE is used to replace nn.Upsample in the neck network, which strengthens the feature fusion of the images and effectively solves the problem that the feature information between the upper and lower layers cannot be fully fused. The

F-SIOU loss function is used to effectively optimize the boundary imbalance in the detection process, and improve the overall detection accuracy and robustness of the model.

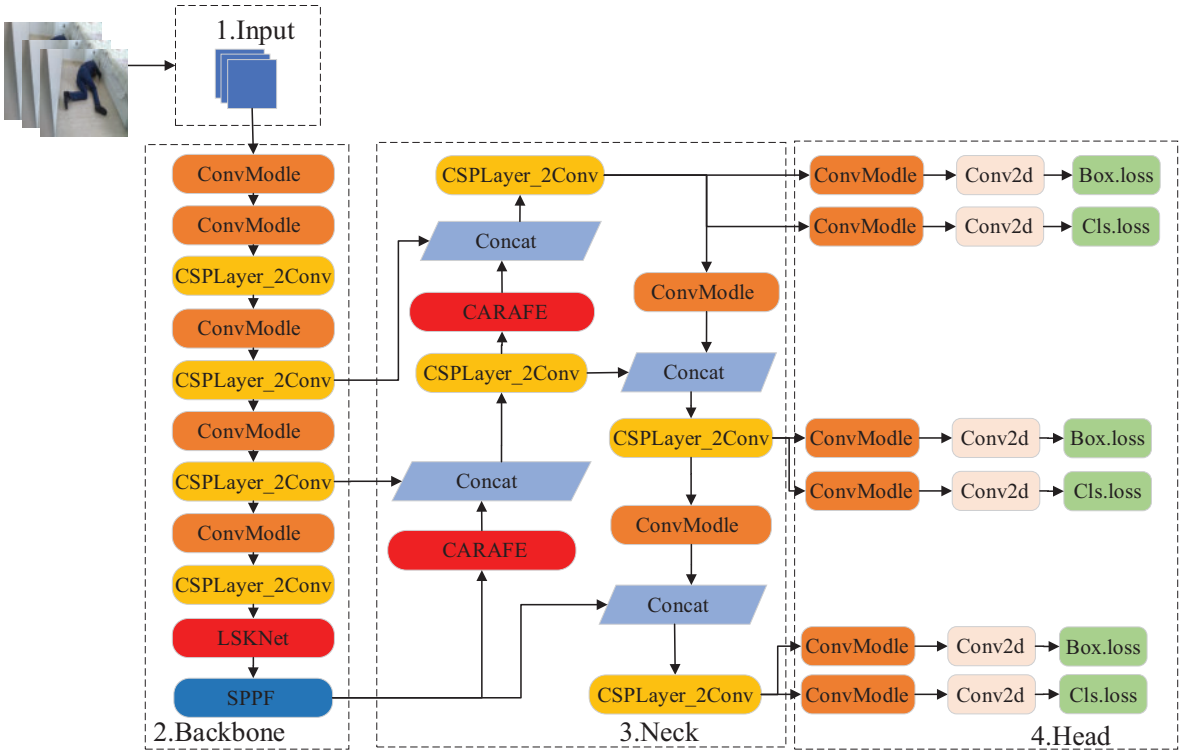

Figure 1: Structure of YOLO-AB network

3.2 Specific Improvement Measures

3.2.1 Network Structure of LSKNet Module

The backbone network part of the YOLOv8 algorithm is responsible for extracting multi-level features from the input images, which serves as the basis for the processing of subsequent modules.

The C2f structure in the backbone network part combines residual connections with cross-stage partial connections, making the gradient flow richer, effectively reducing redundant computations and enhancing the fusion between different features. The C2f structure uses larger receptive fields (such as 5 × 5 or 7 × 7 convolutional kernels) to capture more contextual information, rather than being limited to 3 × 3 convolutional kernels. In the backbone network part, adaptive pooling is adopted instead of the traditional max pooling. Convolution operations with a stride of 2 are used more frequently to replace pooling, which can preserve more feature information.

When detecting abnormal behaviors of the elderly, both large targets and small targets will appear simultaneously. When faced with multi-scale targets, the backbone network fails to extract the key features of the elderly, such as holding a cigarette in the hand, which will result in insufficient feature extraction and affect the model's ability to extract effective features.

The Large Selective Kernel Network (LSKNet [33]) is chosen in order to fully extract the key features of various parts of the elderly people's body. The LSKNet module can focus on processing long-distance context

information and capture multi-scale features. When detecting the elderly, it can better capture the detailed features of the elderly such as smoking with hands, etc.

The LSKNet module consists of a series of large-kernel convolutions and a spatial kernel selection mechanism. The specific structure is shown in Fig. 2.

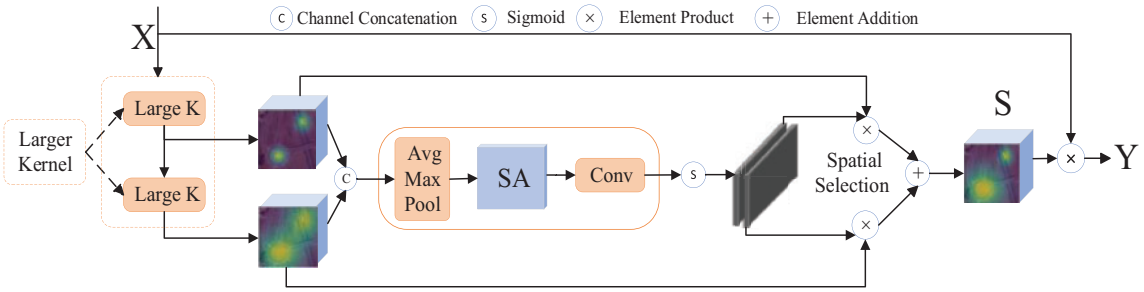

Figure 2: Schematic diagram of LSKNet

The operation steps of LSKNet are as follows:

(1) In the large-kernel convolution, the feature map is simultaneously subjected to average pooling and maximum pooling operations through channel concatenation, and then the two obtained pooling results are spliced and fused.
(2) Perform weighted feature fusion on the fused kernels to obtain feature S.
(3) Multiply the obtained feature S by the input feature X to get the final output.

The core idea of the LSKNet is to decompose large-kernel convolutions layer by layer, and combine depthwise separable convolutions and dilated convolutions to adjust the receptive field. In depthwise separable convolutions, each input channel is convolved with a single convolution kernel, respectively, and then 1×1 convolution kernels are used for channel mixing to integrate the information of each channel. Dilated convolutions enable the convolution kernel to sense a larger area through the dilation rate, expanding the receptive field of the convolution so as to achieve adaptive feature selection.

To enable the module to achieve adaptive selection, we explicitly decompose the large-kernel convolution into a series of depthwise separable convolutions. Larger convolution kernels are constructed by using convolution kernels that grow incrementally in size along with an increasing dilation rate, as presented in Eqs. (1) and (2).

$$h_{i-1} \leq h_i; l_{i-1} < l_i \leq SE_{i-1}, l_1 = 1 \tag{1}$$

$$SE_1 = h_1, SE_i = l_i\,(h_1 - 1) + SE_{i-1} \tag{2}$$

where i is the kernel of the i-th deep convolution, h is the size of the expansion, l is the expansion velocity, and SE is the receiving field.

The expansion rate of the receptive field is guaranteed by the increase in kernel size and expansion rate. Set an upper limit for the dilation rate to ensure that the dilated convolution does not introduce gaps between the feature maps. Generally, the upper limit of the dilation rate is usually 4 or 6. The dilation rate is set to 4 because of small targets such as the situation of smoking with hands. It can display multiple features with various large acceptance domains so as to select kernels more easily. And sequential decomposition is more efficient than simply applying a large kernel.

In order to enhance the spatial context region and improve the ability to detect targets, the spatial selection mechanism is used to select the feature map from large convolution kernels of different scales. The traits obtained from different nuclei are connected with different receptive fields firstly, as shown in Eq. (3).

$$\tilde{U} = \left[\widetilde{U_1}; \widetilde{U_2}; \ldots; \widetilde{U_i}\right] \tag{3}$$

where $\widetilde{U_i}$ is a variable that connects the characteristics of the nucleus with the receptive field.

The channel-based average and maximum pooling is then applied to $\tilde{U}$, as shown in Eq. (4).

$$SA_{avg} = P_{avg}\left(\tilde{U}\right), SA_{max} = P_{max}\left(\tilde{U}\right) \tag{4}$$

where SA stands for pooling and SA_{avg} and SA_{max} are the spatial descriptions of the average and maximum pooling.

The connection of spatial pooled features can ensure the information exchange between different spatial descriptions, as shown in Eq. (5).

$$\widehat{SA} = F^{2\to N}\left(\left[SA_{avg}; SA_{max}\right]\right) \tag{5}$$

where $F^{2\to N}$ represents the conversion function of the pooling features of the two channels into N spatial attention plots.

The sigmoid function is used to decode the SA, and then the features in the decomposed large kernel sequence are weighted to obtain the feature S, as shown in Eqs. (6) and (7).

$$\widetilde{SA_i} = \sigma\left(\widehat{SA_i}\right) \tag{6}$$

$$S = F\left(\sum_{i=1}^{N}\left(\widetilde{SA_i} * \tilde{U}_i\right)\right) \tag{7}$$

where σ is the sigmoid function.

The final output Y of the LSKNet module is the product of the input features X and S. It can be seen clearly from Fig. 2 that the large selection kernel adaptively collects the corresponding large reception fields of different objects to work. The LSKNet module enhances the flexibility and adaptability of the model in different tasks. Especially when dealing with multi-scale targets, it can extract the key features of the targets more effectively. The specific results are presented in Section 4.

3.2.2 CARAFE Module

The neck network part mainly fuses the features from different layers of the backbone network part. The neck network of YOLOv8 uses the Feature Pyramid Network (FPN) and the Path Aggregation Network (PAN) for feature fusion. FPN adopts a top-down feature fusion mechanism to propagate high-level semantic information to the low-level feature maps. PAN adopts a bottom-up feature fusion mechanism, which transmits the spatial information of low-level features to high-level feature maps, enhancing the spatial information of the targets.

The neck network uses the Task-Aligned Assigner method for positive and negative sample matching, which can dynamically allocate thresholds. During the abnormal behavior detection process, small targets (smoking actions) and large targets (falling actions) need to be detected simultaneously. The Task-Aligned Assigner dynamically adjusts the assignment of positive and negative samples according to the quality of the prediction boxes. It lowers the threshold for small targets and raises the threshold for large targets, so that

more prediction boxes are included in the range of positive samples and the problem of bounding box shift is reduced.

However, during feature fusion, the multi-scale features of some elderly people (such as hand movements when smoking and falling) are not fully fused, resulting in lower detection accuracy or even missing targets. By increasing the number of convolutional layers to expand the parameter scale of the fusion network, the fusion effect can be enhanced. However, this will increase the computational complexity of the model and reduce the processing speed.

The Content-Aware Reassembly of Features Module (CARAFE [34]) for up-sampling the feature maps has been introduced in order to enhance the feature fusion ability and ensure that the number of parameters does not increase simultaneously. CARAFE predicts the reassembly kernels using the underlying information and reassembles the features within the nearby pre-processed area. It can use adaptive and optimized reassembly kernels at different positions, achieving better feature extraction performance than other mainstream up-sampling operators.

The structure of CARAFE is shown in Fig. 3, where the feature map with the size of $C \times H \times W$ is up-sampled and δ is the up-sampling factor. CARAFE consists of a kernel prediction module and a content-aware reassembly module. The kernel prediction module predicts the convolution kernel for each local region according to the input feature map and adaptively assigns different convolution kernels to different spatial positions. In this way, each position can be processed differently, enhancing the flexibility of the model. The content-aware reassembly module generates an appropriate reassembly method according to the content of the input feature map. For example, in the edge areas of the pictures of the elderly, the reassembly operation will pay more attention to the detailed parts, while in the flat areas, it will focus on large-scale smooth processing. It can adaptively adjust according to the data content, making the feature fusion more flexible. This can ensure that the reassembled feature maps contain more information about the local environment.

With this design, CARAFE can dynamically adjust the size of the reassembly kernels in different regions, enabling YOLO-AB to effectively fuse low-level and high-level information during the feature extraction and up-sampling processes. This not only improves the feature expression ability of YOLO-AB, allowing it to better capture the subtle changes in the behaviors of the elderly, but also enhances the accuracy and robustness of the model in complex scenarios when detecting abnormal behaviors such as falling and smoking among the elderly.

The kernel prediction module is mainly responsible for generating the recombinant kernel in a content-aware manner, which is mainly composed of a channel compressor, a content encoder and a kernel normalizer. The channel compressor compresses the channel of the input feature map, and then the content encoder takes the compressed feature map as the input, encodes its content to generate a recombination kernel, and finally uses the softmax function to apply it to the recombination kernel. The specific calculation process is as follows.

(1) Assuming that the shape of a feature map $C \times H \times W$ is given, the upsampling multiple is σ, and the output feature map size is $C \times \sigma H \times \sigma W$. The convolutional layer used by the channel compressor compresses the feature channel from C to C1, which can reduce the parameters and computational cost.

(2) Suppose the up-sampled nucleus size is $k \times k$. Recombinant kernels are generated by the content encoder. A convolutional layer with a convolutional kernel $k_l \times k_l$ is used to predict the size $C1 \times H \times W$ of the up-sampled kernel. The size of the output feature map is $H \times W \times \sigma^2 \times k^2$. The channel dimension of the feature map is expanded in the spatial dimension to obtain an upsampling kernel of size of $\sigma H \times \sigma W \times k^2$.

(3) Softmax normalization is performed on the predicted upsampling kernels, so that the sum of the weights of the convolution kernels is 1.
(4) The predicted upsampling kernel and the input features are convoluted to obtain the final output.

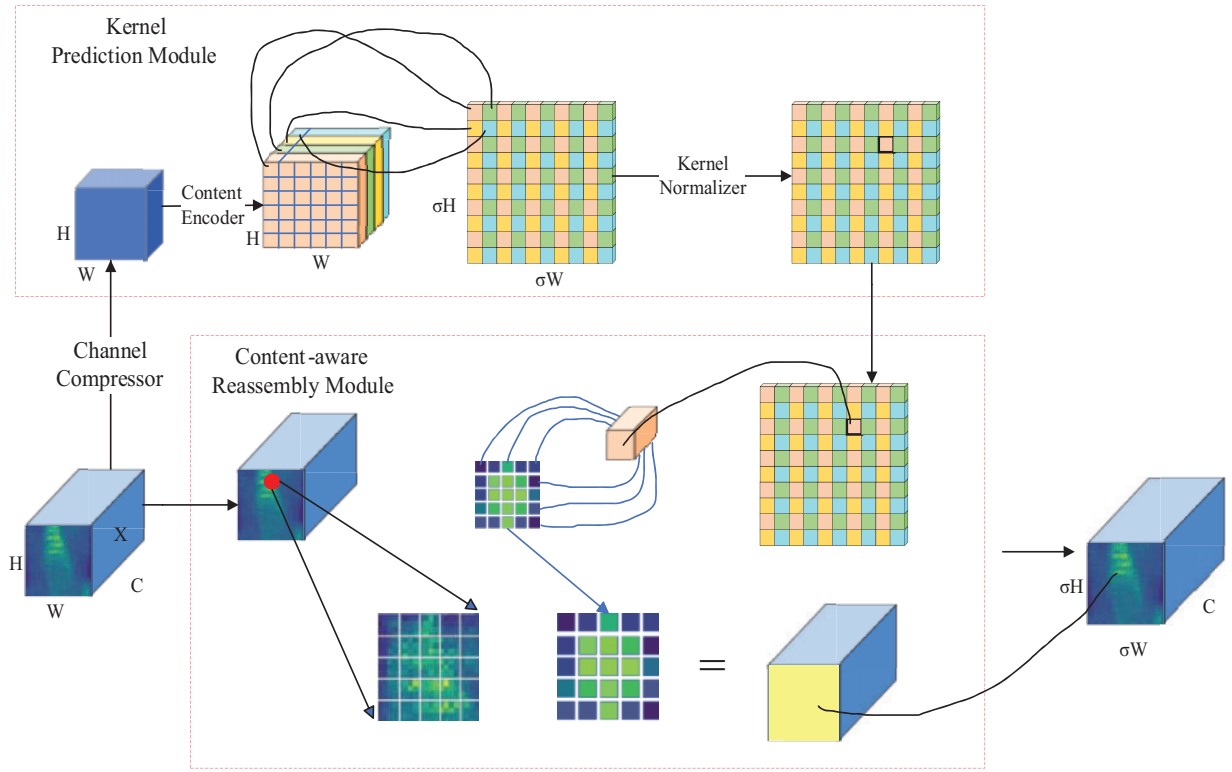

Figure 3: Overall framework diagram of CARAFE

The up-sampled kernel of CARAFE is predicted by the input features, which contains stronger semantic information and has a larger receptive field. The information of the relevant points in the local region can get more attention. CARAFE introduces only a small number of parameters and calculations, but achieves an increase in accuracy. The amount of parameters for the CARAFE sampling process is shown in Eq. (8).

$$M = CC_1 + \left(C_1 k_1^2 \sigma^2 + 1\right) \sigma^2 k^2 + \sigma^2 k^2 C \tag{8}$$

The CARAFE module strengthens the feature fusion of pictures of the elderly at different scales, enhancing the feature fusion among the pictures and effectively improving the accuracy in detecting abnormal behaviors of the elderly. The specific results are shown in Section 4.

3.2.3 F-SIOU (Focal-SCYLLA-IOU) Loss Function

Usually, the target of anomalous behavior detection is only a part of the overall image, and the background also occupies the other parts of the image. This phenomenon can lead to imbalances when classifying. The original loss function may assign too many or the same weights to the background pixels, which will result in lower detection accuracy. The F-SIOU loss function is used to solve such a problem in our research.

The F-SIOU loss function introduces Focal Loss, which optimizes the problem of boundary imbalance in the detection of the elderly. This loss function not only considers the weights of positive and negative samples but also considers the angle, distance, shape, IOU, and other factors between the real bounding box and the predicted bounding box. The function calculates the loss according to the relationship between the spatial intersection and union of the measured and predicted bounding boxes and the real bounding boxes, which helps to improve the accuracy of the loss calculation, accelerates the convergence speed of the model, and improves the accuracy of the model in the regression task [35,36].

The F-SIOU loss function consists of four indicators, namely Angle cost, Distance cost, Shape cost, and IOU cost.

The angle cost represents the rotation angle of the target box (the angle between the box and the horizontal line) when detecting the elderly. If the angular deviation of the predicted box is too large, even if the center and size of the box are correct, the IOU value will be relatively low, affecting the accuracy of the model. Only by precisely locating the behavior can the position of the target be accurately found.

The distance cost measures the difference in the distance between the center of the predicted box and the center of the real box. It is necessary to ensure that the center position of the target box coincides with that of the real box. By minimizing the distance between the center points, namely, diminishing the distance between the two box centers to the greatest extent, the pinpoint localization of the target can be optimized.

The shape cost measures the shape difference between the predicted box and the ground-truth box, mainly taking into account the aspect ratio of the target box. Even if the overlapping area of the target boxes is relatively large, the IOU will decrease if the aspect ratios of the two differ significantly.

The IOU cost measures the intersection loss between the predicted box and the ground-truth box. IOU refers to the proportion of the overlap between the predicted box and the ground-truth box. The IOU cost minimizes the IOU difference between the predicted box and the ground-truth box, that is, it makes the two boxes overlap as much as possible. As the degree of box overlap increases, the detection accuracy will also improve.

The Angle cost minimizes distance-related variables and bring the prediction to the X or Y axis. Then it continued to approach along the relevant axis, and tried to minimize the angle α when it is less than $\frac{\pi}{4}$, otherwise β will be minimized. The calculation scheme of the Angle cost is shown in Fig. 4, where α or β is the relative angle of the center point of the two bounding boxes, x is the sin value of this angle, c_h is the vertical distance between the centers of the two bounding boxes, and γ is the straight-line distance between the centers of the two bounding boxes.

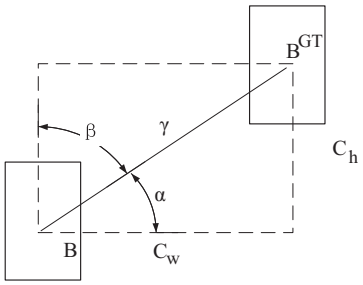

Figure 4: Calculation of the angle indicator

The calculation is shown in Eqs. (9)–(12).

$$\Lambda = 1 - 2sin^2\left(\sin^{-1}(x) - \frac{\pi}{4}\right) \tag{9}$$

$$\mathrm{x} = \frac{c_h}{\gamma} = \sin\alpha \tag{10}$$

$$\gamma = \sqrt{\left(b_{cx}^{gt} - b_{cx}\right)^2 + \left(b_{cy}^{gt} - b_{cy}\right)^2} \tag{11}$$

$$c_h = \max\left(b_{c_y}^{gt}, b_{c_y}\right) - \min(b_{c_y}^{gt}, b_{c_y}) \tag{12}$$

where $\mathbf{\Lambda}$ is the Angle cost, $b_{c_x}^{gt}$ and b_{c_x} are the x coordinates of the real bounding box and the predicted bounding box, respectively, and $b_{c_y}^{gt}$ and b_{c_y} are the y coordinates of the bounding box and the predicted bounding box, respectively.

Then the Distance cost is redefined, as is shown in Eqs. (13)–(15).

$$\Delta = \sum t = x, y\left(1 - e^{-rpt}\right) \tag{13}$$

$$p_x = \left(\frac{b_{cx}^{gt} - b_{cx}}{c_w}\right)^2, p_y = \left(\frac{b_{cy}^{gt} - b_{cy}}{c_y}\right)^2 \tag{14}$$

$$r = 2 - \Lambda \tag{15}$$

where Δ is the Distance cost, which represents the distance difference between the prediction box and the real box, p_x and p_y represent the loss of the bounding box on the x-axis and y-axis, respectively.

Δ uses r to weight the loss of p_x and p_y to control the Distance cost. r changes with Λ. When α is small, r is larger, and the contribution of Distance cost decreases. When αapproaching $\frac{\pi}{4}$, r becomes smaller, and the contribution of Distance cost increases. This situation indicates that distance optimization is more important.

Shape cost is represented by a symbol Ω, and it is calculated as shown in Eqs. (16)–(18).

$$\Omega = \sum_{t=w,h}\left(1 - e^{-wt}\right)^{\theta} \tag{16}$$

$$w_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})} \tag{17}$$

$$w_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{18}$$

where w_w and w_h represent the relative difference in width and height, respectively, w and w^{gt} represent the width of the predicted bounding box and the true bounding box, respectively, h and h^{gt} represent the height of the predicted and true bounding boxes, respectively.

θ controls the level of attention paid to Shape cost which defines the cost of the shape and is unique for each dataset. The higher the value of θ, the greater the weight of the Shape cost. The value with range from 3 to 5 can be calculated by genetic algorithm.

IOU cost is used to measure the crossover loss between the predicted and the actual, detecting the degree of repetition of the framework. The smaller the *IOU*, the greater the *IOU* cost. The final regression loss function L_{box} consists of Angle cost, Distance cost, Shape cost, and *IOU* cost, and the calculation process of the loss function L_{box} is shown in Eq. (19).

$$L_{box} = 1 - IOU + \frac{\Delta + \Omega}{2} \tag{19}$$

In order to alleviate the problem of imbalance between positive and negative samples in the regression process, the Focal *EIOU* loss function L_{F-EIOU} integrates the Focal Loss and *EIOU* loss functions, which are calculated according to Eq. (20).

$$L_{F-EIOU} = IOU^{\tau} L_{box} \tag{20}$$

According to the principle of Focal *EIOU* loss function, the *SIOU* loss function and Focal Loss are combined to obtain a new Focal *SIOU* loss function, which is used for the regression loss function of YOLOv8, and is calculated according to Eq. (21).

$$L_{F-SIOU} = IOU^{\tau} L_{box} \tag{21}$$

where τ is a parameter for controlling the sample weights, and its value is generally set to 0.5.

The use of the F−SIOU loss function enables the model to capture various factors such as the rotation, morphological changes, and displacement of the target more precisely, providing higher positioning accuracy and robustness for the detection of the elderly's behaviors. The specific results are presented in Section 4.

4 Experiments

4.1 Experimental Environment and Parameter Settings

The experimental development environment is Anaconda1.7.2 and PyTorch1.2. The experimental platform is Intel i5-6500 CPU, equipped with an NVIDIA RTX 3060 Ti graphics card, and Windows 10 operating system. The initial learning rate is set to 0.0001, the number of training rounds (epochs) is set to 200, and the batch size is set to 4, which ensure the convergence and performance of model training.

4.2 Experimental Data and Evaluation Indicators

There aren't many publicly accessible datasets of human anomalous behavior currently. The authors independently created a human abnormal behavior detection dataset called Abnormal Behaviour-Dataset (AB-Dataset), and the images in the dataset were mainly obtained through their own image collection and YouTube videos. In this study, the collected images were selectively processed. Since the research primarily focuses on elderly behaviors such as falling and smoking, images depicting the elderly falling, smoking, and standing were selected as the image dataset for this investigation. To enhance the generalization capability of the data, data augmentation techniques were applied, including flipping, color space transformation, and resizing. Specific examples are illustrated in Fig. 5. To ensure the multifaceted nature of the data in this study, the dataset includes samples from both indoor and outdoor environments, as specifically illustrated in Fig. 6.

The images were annotated using Labelimg software and three label types, namely falling, smoking, and standing, were added. The dataset was divided into a training set, a test set, and a validation set at a ratio of 8:1:1. A total of 9018 images were labeled, which were divided into 7214 training sets, 902 test sets, and 902 validation sets. Fig. 7 illustrates the distribution and proportion of the dataset annotations. It can be observed that the data exhibit good uniformity and diversity across different categories. The annotation boxes are relatively evenly distributed and predominantly centered within the images, providing reliable data support for the generalization capability of the model.

Figure 5: Image augmentation processing

Figure 6: Some experimental images in the dataset

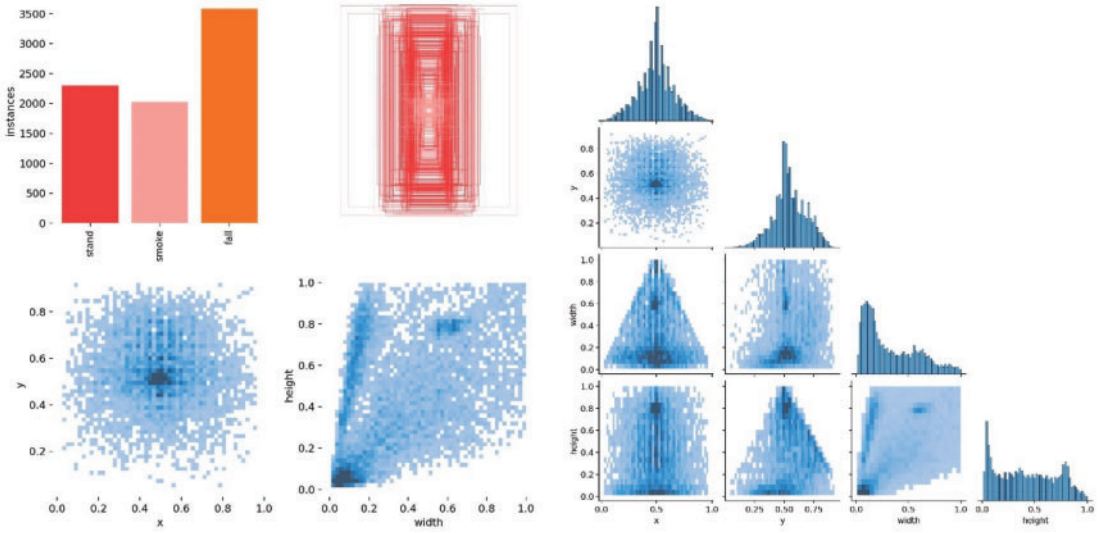

Figure 7: The distribution and proportion of annotated data in the dataset

Several indicators are mainly used to evaluate the detection performance of the model, including precision of detecting targets (P), recall rate (R), mean average precision with 50% overlap ($mAP50$), frames per second, referring to the number of processed images per second (FPS), the number of model parameters (Parameters), and standard deviation (SD). Definitions of them are shown in Eqs. (22)–(25).

$$P = \frac{TP}{TP + FP} \tag{22}$$

$$R = \frac{TP}{TP + FN} \tag{23}$$

$$mAP50 = \frac{1}{n_{class}} \sum_{i=1}^{n_{class}} \frac{1}{n_{recall}} \sum_{r=0}^{n_{recall}} AP \tag{24}$$

$$SD = \sqrt{\frac{\sum_{y=1}^{N} \left(x_y - \overline{x}\right)^2}{N}} \tag{25}$$

where TP is the correct positive sample, FP is the positive sample that was detected incorrectly, FN is the number of undetected samples, n_{class} denotes the number of object categories, n_{recall} denotes the number of recalled values set, AP represents the average precision rate of a single category, N represents the number of experiments, y represents the number of experimental times, x_y represents the result of each experiment, and $\overline{x}$ represents the average value of the experiments.

4.3 Comparative Experiments

4.3.1 Experiments on the Improvement of the Backbone Network

It is a common strategy to use various network modules to improve the performance of the algorithm model. The commonly used improved modules are AFPN, LSKNet, BiFPN, C2f_DCN, Ghost, FasterNet, Slim, and OD-Conv modules. Fig. 8 shows the experimental comparison results of the backbone network improvement modules. It can be seen from Fig. 8 that the P and R values of the improved LSKNet module reach 0.887 and 0.856, respectively, and the mAP50 reaches 0.892. Among the many modules, the LSKNet module shows significant advantages in improving the backbone network. This shows that the LSKNet module has superior performance in object detection, which not only improves the accuracy of abnormal behavior detection, but also significantly reduces the incidence of false detection and missed detection.

4.3.2 Experiments on the Improvement of the Neck Network

To strengthen the feature fusion ability of the algorithm, the neck network of the algorithm is improved in this experiment. The improved replacement modules are DySample, DEA-Net, EVC, BGF-YOLO, CSP-Stage, GELAN, and CARAFE modules. A series of experiments were conducted to analyze the results by comparing multiple performance indicators such as P, R, mAP50, FPS, and Parameters to comprehensively evaluate the advantages of CARAFE over other modules. The experimental comparison results of the neck network improvement module are shown in Figs. 9 and 10.

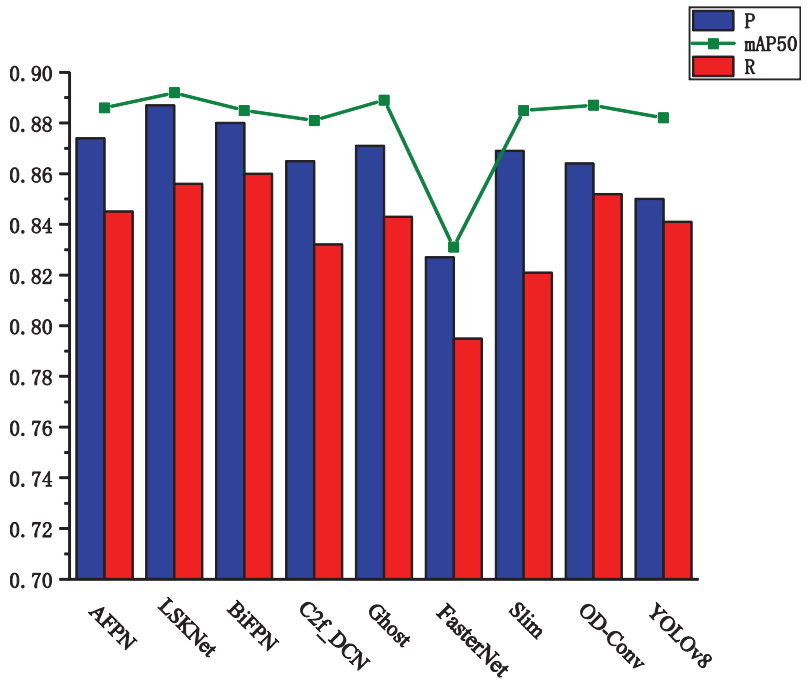

Figure 8: Experimental comparison of backbone network improvement modules

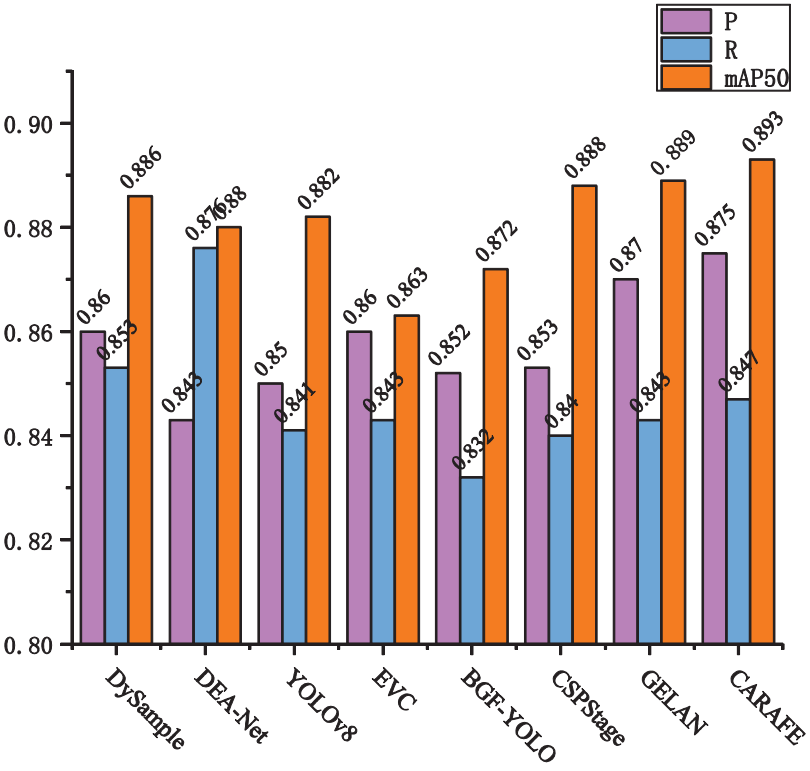

Figure 9: Experimental comparison results of the neck network improvement module

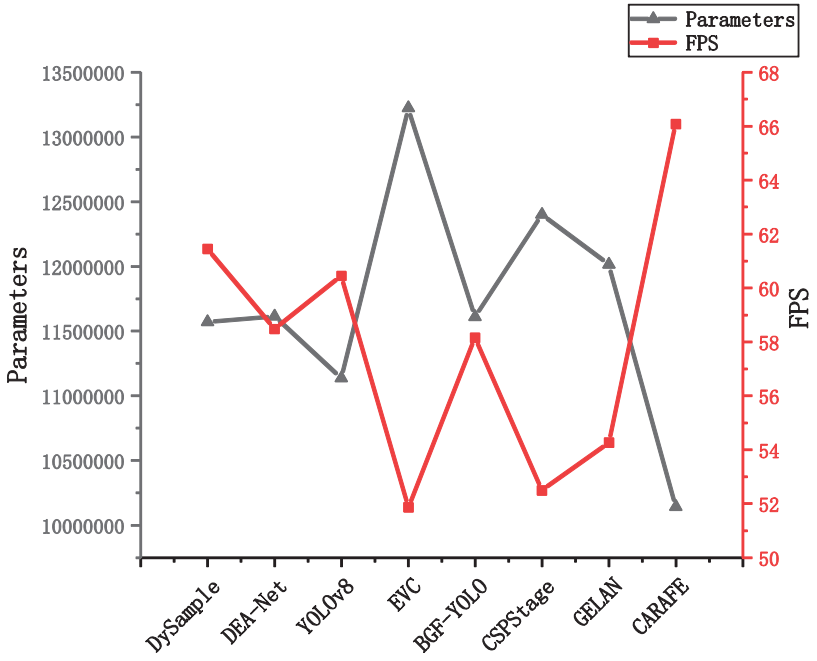

Figure 10: Experimental comparison of FPS and Parameters of improved rear neck network

The results presented in Figs. 9 and 10 illustrate that the CARAFE module outperforms other modules across a range of performance metrics. In particular, the model that employed the CARAFE module achieved a mAP50 of 0.893, with P and R of 0.876 and 0.847, respectively. Moreover, it demonstrated an FPS of 66.08 while concurrently reducing the number of parameters involved. The experimental findings demonstrate that the CARAFE module not only achieves superior accuracy in detecting anomalous behaviors but also significantly reduces both false positive rates and missed detections, all while ensuring a high level of detection stability. This capability markedly enhances the model's capacity to detect anomalous behaviors of small targets against complex backgrounds. The CARAFE module exemplifies its ability to maintain a high FPS while concurrently reducing the number of parameters, thereby further substantiating its computational efficiency and enhanced detection accuracy.

4.3.3 Experiments on the Improvement of the Neck Network

Select the appropriate loss function to train the model and evaluate it experimentally. The loss functions of CIOU, DIOU, EIOU, GIOU, WIOU, F-DIOU, F-EIOU and F-GIOU were combined as comparisons, and the results were analyzed by key performance indicators such as P, R, and mAP50, as shown in Fig. 11.

It can be seen from Fig. 11 that the P and R of the F-SIOU loss function can reach 0.88 and 0.859, indicating that high-precision object detection can be achieved while maintaining a high recall rate. The mAP50 value of the F-SIOU loss function can reach 0.898, indicating that the F-SIOU loss function has excellent performance. In summary, compared to other loss functions, the F-SIOU loss function performs well on all key performance indicators and can provide excellent performance in high-precision object detection tasks.

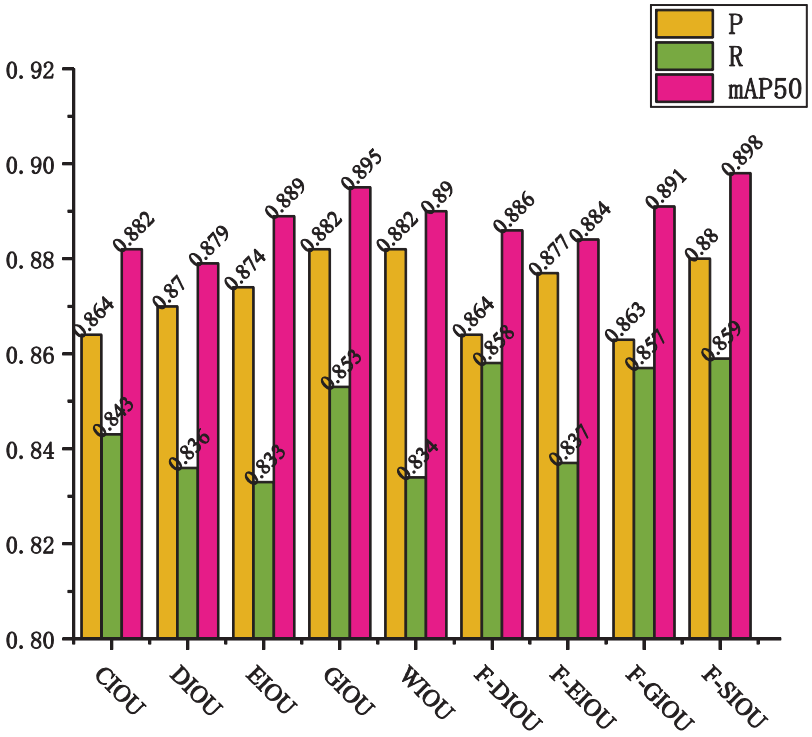

Figure 11: Comparative experiment of loss functions

4.3.4 Comparison of Loss Curve Results during Training

To enhance the effectiveness and robustness of the algorithm, the loss curves of YOLOv8 and YOLO-AB have been compared. The loss curve during training is shown in Fig. 12. It can be seen that with the increase of the number of training epochs, the loss curve shows a downward trend. When the training reaches 200 epochs, the model gradually converges and tends to be stable. Moreover, there is no sign of overfitting in the whole process. Compared with YOLOv8, the training loss of YOLO-AB is smaller and the downward trend is more obvious, indicating that YOLO-AB has better data fitting ability, which verifies the effectiveness and stability of the algorithm.

4.3.5 Comparison of Fall and Smoking Detection Separately and Fall and Smoking Fusion Detection

In order to verify the effectiveness of the YOLO-AB algorithm for fusion detection of abnormal behaviors in the elderly, comparative experiments were carried out on falling detection, smoking detection, and fusion detection, respectively. Comparative experiments and validation were conducted with other mainstream algorithms such as SSD, Faster R-CNN, RetiaNet, Resnet [21], U-Net [24], MobileNetV3-YOLO, and YOLO series. The performance indicators of P, R, mAP50, FPS, and Parameters were mainly used to evaluate. The abnormal behavior detection in this study was mainly to evaluate falling detection, smoking detection, and fusion detection of the two. The experimental results are shown in Table 1.

In the YOLO-AB algorithm, the precision (P) values in fall detection and smoking detection are 0.893 and 0.889, respectively, while the recall (R) values are 0.855 and 0.809, respectively. In the fused detection of falling and smoking, the P value and R value are 0.899 and 0.876, respectively. Compared with other algorithms, the YOLO-AB algorithm has higher precision (P value) and recall (R value). The mAP50 values of the YOLO-AB algorithm for fall detection and smoking detection can reach 0.93 and 0.864, respectively.

The *mAP50* value for the fused detection of falling and smoking is 0.923. Compared with other algorithms, it has better balance performance between accuracy and recall.

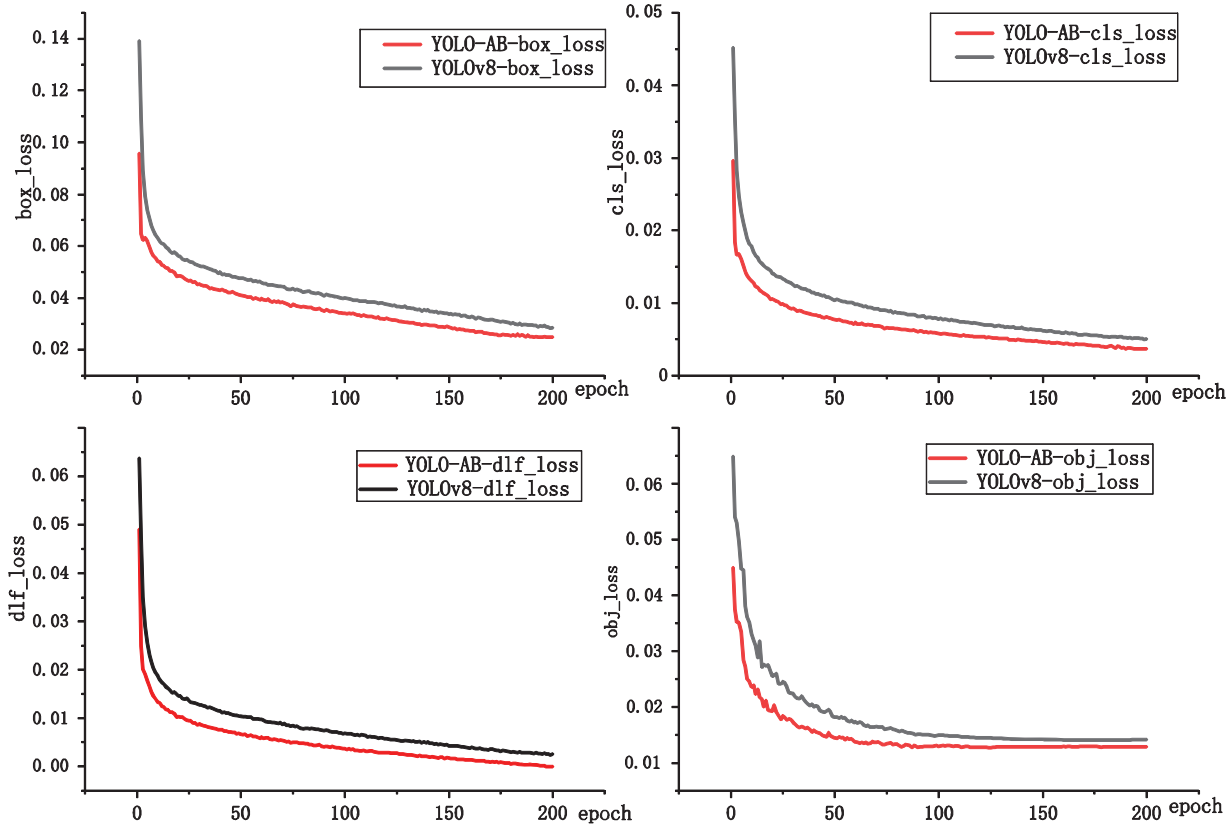

Figure 12: The loss curve during training

Table 1: Comparative experiment of fall detection, smoke detection and fusion detection

Model	Class	P	R	mAP50	FPS	Parameters
YOLOv10	Fall	0.836	0.881	0.898		
	Smoke	0.878	0.588	0.713	70.35	7,201,433
	All	0.867	0.815	0.863		
YOLOv5	Fall	0.835	0.852	0.891		
	Smoke	0.859	0.787	0.813	72.61	2,654,816
	All	0.863	0.868	0.889		
YOLOv8	Fall	0.837	0.826	0.883		
	Smoke	0.818	0.732	0.8	60.45	11,136,761
	All	0.85	0.841	0.882		
YOLOv6	Fall	0.786	0.812	0.84		
	Smoke	0.811	0.765	0.798	64.25	4,500,080
	All	0.823	0.811	0.863		

(Continued)

Table 1 (continued)

Model	Class	P	R	mAP50	FPS	Parameters
YOLOv7	Fall	0.768	0.77	0.892		
	Smoke	0.812	0.81	0.82	54.13	6,510,075
	All	0.882	0.834	0.87		
SSD	Fall	0.754	0.875	0.894		
	Smoke	0.794	0.749	0.849	64.8	9,058,161
	All	0.76	0.858	0.857		
Faster R-CNN	Fall	0.648	0.758	0.825		
	Smoke	0.612	0.71	0.84	46.45	13,671,518
	All	0.625	0.762	0.834		
RetinaNet	Fall	0.854	0.843	0.885		
	Smoke	0.805	0.829	0.857	50.64	3,635,854
	All	0.824	0.84	0.876		
MobileNetV3-YOLO	Fall	0.875	0.794	0.835		
	Smoke	0.832	0.751	0.849	70.12	3,055,947
	All	0.847	0.785	0.84		
Resnet [21]	Fall	0.811	0.756	0.785		
	Smoke	0.762	0.745	0.823	51.45	15,262,153
	All	0.823	0.805	0.862		
U-Net [24]	Fall	0.824	0.832	0.852		
	Smoke	0.782	0.812	0.834	45.19	18,548,459
	All	0.833	0.821	0.892		
YOLO-AB	Fall	0.893	0.855	0.93		
	Smoke	0.889	0.809	0.864	58.53	11,570,111
	All	0.899	0.876	0.923		

Meanwhile, the YOLO-AB algorithm also has a smaller number of parameters compared with most other algorithms, which only has more parameters than RetinaNet and MobileNetV3-YOLO. But the mAP50 values of these two algorithms are 0.876 and 0.84, respectively, which are much lower than the 0.923 of the YOLO-AB algorithm. It also shows some deficiencies in terms of accuracy and stability. Other algorithms, such as Faster R-CNN and U-Net, have a larger number of parameters, and they require more computational resources and thus have lower applicability.

Overall, the YOLO-AB algorithm demonstrates excellent performance across multiple metrics, especially in terms of precision (P), recall (R), and mAP50 values. Although some other algorithms might have a slight edge in FPS, YOLO-AB offers stable performance, achieving good results in both FPS and accuracy. With a relatively small number of parameters, it is suitable for deployment under limited-resource conditions. It is also highly applicable to real-time video surveillance, as it can promptly capture abnormal behaviors among the elderly, make rapid judgments, and issue early warnings, thus providing reliable protection for the daily safety of the elderly.

4.3.6 Standard Deviation Analysis

To comprehensively evaluate the stability of the YOLO-AB algorithm, a standard deviation analysis of the P-values, R-values and mAP50 values of the mainstream algorithms and the YOLO-AB algorithm has been conducted, which is shown in Table 2.

Table 2: Standard deviation analysis

Model	Class	P	SD	R	SD	mAP50	SD
YOLOv5	Fall	0.835	±0.023	0.852	±0.015	0.891	±0.015
	Smoke	0.859	±0.017	0.787	±0.013	0.813	±0.016
	All	0.863	±0.02	0.868	±0.019	0.889	±0.014
YOLOv10	Fall	0.815	±0.018	0.814	±0.015	0.87	±0.019
	Smoke	0.798	±0.024	0.675	±0.013	0.748	±0.016
	All	0.832	±0.022	0.816	±0.01	0.833	±0.017
YOLOv8	Fall	0.837	±0.016	0.826	±0.013	0.883	±0.02
	Smoke	0.818	±0.014	0.732	±0.015	0.8	±0.015
	All	0.85	±0.013	0.841	±0.012	0.882	±0.014
YOLOv6	Fall	0.786	±0.026	0.812	±0.025	0.84	±0.021
	Smoke	0.811	±0.025	0.765	±0.029	0.798	±0.016
	All	0.823	±0.023	0.811	±0.024	0.863	±0.019
YOLOv7	Fall	0.768	±0.017	0.77	±0.015	0.892	±0.017
	Smoke	0.812	±0.021	0.81	±0.013	0.82	±0.018
	All	0.882	±0.018	0.834	±0.011	0.87	±0.014
SSD	Fall	0.754	±0.029	0.875	±0.016	0.894	±0.028
	Smoke	0.794	±0.026	0.749	±0.013	0.849	±0.025
	All	0.76	±0.021	0.858	±0.012	0.857	±0.02
Faster R-CNN	Fall	0.648	±0.016	0.758	±0.016	0.825	±0.018
	Smoke	0.612	±0.018	0.71	±0.011	0.84	±0.014
	All	0.625	±0.017	0.762	±0.015	0.834	±0.017
RetinaNet	Fall	0.854	±0.015	0.843	±0.018	0.885	±0.012
	Smoke	0.805	±0.013	0.829	±0.014	0.857	±0.015
	All	0.824	±0.012	0.84	±0.015	0.876	±0.014
MobileNetV3-YOLO	Fall	0.875	±0.018	0.794	±0.013	0.835	±0.021
	Smoke	0.832	±0.015	0.751	±0.011	0.849	±0.015
	All	0.847	±0.016	0.785	±0.01	0.84	±0.012
Resnet [21]	Fall	0.811	±0.01	0.756	±0.012	0.785	±0.011
	Smoke	0.762	±0.013	0.745	±0.009	0.823	±0.013
	All	0.823	±0.014	0.805	±0.008	0.862	±0.012
U-Net [24]	Fall	0.824	±0.012	0.832	±0.015	0.852	±0.009
	Smoke	0.782	±0.015	0.812	±0.01	0.834	±0.014
	All	0.833	±0.01	0.821	±0.009	0.892	±0.01
YOLO-AB	Fall	0.893	±0.013	0.855	±0.008	0.93	±0.012
	Smoke	0.889	±0.011	0.809	±0.014	0.864	±0.01
	All	0.899	±0.009	0.876	±0.01	0.923	±0.011

It can be seen that:

(1) The standard deviations of the P-values for fall detection and smoking detection of YOLO-AB are 0.013 and 0.011 respectively, and the standard deviation for the fusion detection is 0.009.
(2) The standard deviations of the R-values for fall detection and smoking detection of YOLO-AB are 0.008 and 0.014 respectively, and the standard deviation for the fusion detection is 0.01.
(3) The standard deviations of mAP50-values for fall detection and smoking detection of YOLO-AB are 0.012 and 0.01 respectively, and the standard deviation for the fusion detection is 0.011.

Compared with other mainstream algorithms, it can be seen that the YOLO-AB algorithm maintains higher precision and good stability with a higher recall rate and smaller data fluctuations in terms of standard deviation.

4.4 Ablation Experiments

A series of ablation experiments about YOLO-AB were performed to evaluate the improvements of CARAFE, LSKNet, and F-SIOU on the performance of YOLOv8. During the experiment, it is necessary to ensure the consistency of the experimental environment and parameter settings. Each improvement point of the algorithm model was evaluated to understand its effect on the overall algorithm performance, and the results are shown in Table 3, where √ indicates that this module was used.

Table 3: Ablation experiments

Experiment serial number	CARAFE	LSKNet	F-SIOU	P	R	mAP50	FPS	Parameters
1				0.85	0.841	0.882	60.45	11,136,761
2	√			0.875	0.847	0.893	66.08	10,141,865
3	√	√		0.877	0.871	0.909	63.91	11,570,111
4	√	√	√	0.899	0.876	0.923	64.59	11,570,111

From the data in Table 3, it can be seen that each improvement point has a significant contribution to the performance improvement of the YOLOv8 anomalous behavior detection model. Experiment 1 exhibits the result of the original YOLOv8 object detection algorithm, which is used as a reference for several subsequent experiments. In Experiment 2, the CARAFE module was used, and P and R increased by 0.025 and 0.006, respectively, and mAP50 increased by 0.011. In Experiment 3, the LSKNet module was used to adjust the backbone network on the basis of Experiment 2, and the performance was further improved, with P and R increased by 0.002 and 0.024, respectively, and mAP50 increased by 0.016. In Experiment 4, the F-SIOU loss function was sampled on the basis of Experiment 3 to further improve the performance of the algorithm, with P and R increased by 0.022 and 0.005, and mAP50 increased by 0.014. As can be seen from the data, these results highlight the assisting nature of the improvement work, and the YOLO-AB anomalous behavior detection model has achieved significant performance improvements in various fields by adding various modules.

4.5 Nested K-Fold Validation

To thoroughly evaluate the generalization capability and stability of the dataset and model in this study, Nested K-Fold validation was employed to analyze metrics such as mAP50, P, and R. A cross-validation process with 10 iterations was conducted, and the specific validation results are presented in Fig. 13. As can

be seen from the figure, as the number of folds increases, the fluctuations in the dataset's mAP50, P, and R are minimal, and the overall trend is upward. This phenomenon may indicate that increasing the number of folds enhances the stability and performance of the model. The validation results of the YOLO-AB model were compared with those of YOLOv8, demonstrating that the proposed algorithm model exhibits superior performance, achieving higher values in mAP50, P, and R, along with improved stability. These results indicate that the dataset used in this study demonstrates excellent detection performance on YOLO-AB, exhibiting strong generalization capabilities without signs of overfitting.

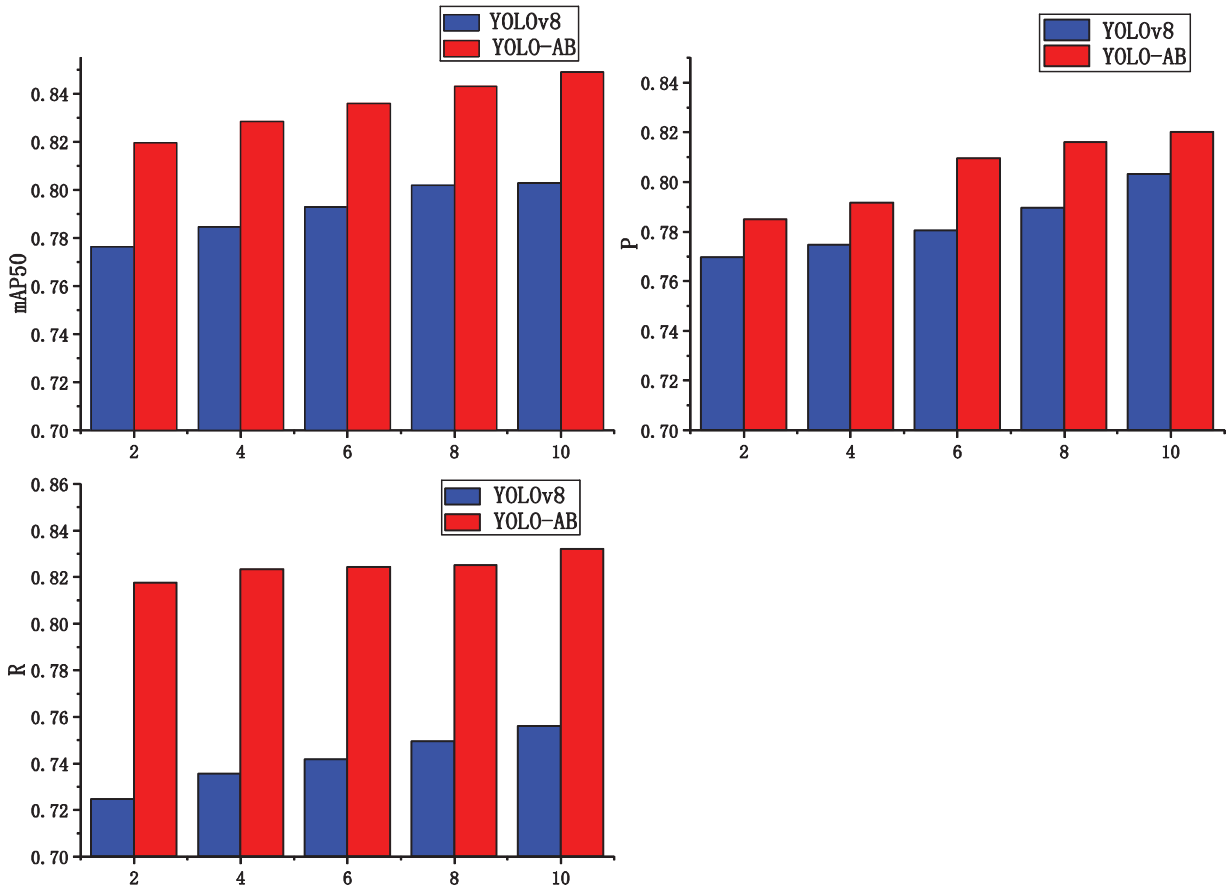

Figure 13: The mAP50, P and R values of Nested K-Fold validation

4.6 Testing Experiments

A test experiment was planned and carried out in order to verify the feasibility, and effectiveness of the YOLO-AB algorithm in practical application scenarios more in-depth, comprehensive and systematic manner. The core algorithms used in this experiment are YOLOv8 and the YOLO-AB algorithm. These algorithms focus on two behaviors falling and smoking, which are of great social concern in real scenes. The dataset contains data related to falling and smoking scenarios. The results of the experiment are presented in an intuitive visual form, as shown in Fig. 14. The specific experimental data are shown in Table 4.

Figure 14: Comparison of the experimental results of YOLOv8 and YOLO-AB algorithms; (a) The detection results of the YOLOv8 algorithm model; (b) The detection results of the YOLO-AB algorithm model

Fig. 14a shows the detection results of the YOLOv8 algorithm, and Fig. 14b shows the detection results of the YOLO-AB algorithm. The "Left" column in Table 3 shows the detection results of the left picture in Fig. 14. It can be seen that the detection accuracy of the YOLOv8 algorithm for falling and smoking can reach 0.9 and 0.8, while the YOLO-AB algorithm can reach 0.93 and 0.87, respectively. The "Middle" column in Table 3 shows that the detection accuracy of the YOLOv8 algorithm for falling and smoking can reach 0.78 and 0.75, while the YOLO-AB algorithm can reach 0.85 and 0.84, respectively. The "Rright" column in Table 3 shows the detection results of the right picture in Fig. 14, while the accuracy of the YOLOv8 algorithm for falling and smoking can reach 0.9 and 0.83, and the YOLO-AB algorithm can reach 0.96 and 0.88, respectively. The experimental results show that the YOLO-AB algorithm has better detection accuracy

than the YOLOv8 algorithm on the falling and smoking dataset, and has stable and accurate detection ability in different scenarios.

Table 4: Comparative detection accuracy

Model	Category	Precision		
		Left	Middle	Right
YOLOv8	Fall	0.9	0.78	0.9
	Smoke	0.8	0.75	0.83
YOLO-AB	Fall	0.93	0.85	0.96
	Smoke	0.87	0.84	0.88

4.7 Generalization Experiments

To further verify the generalization ability of the YOLO-AB algorithm, this study conducted experimental verification on the publicly available UR Fall Detection Dataset [37] and the author's own real-time video data using the YOLO-AB algorithm. The images collected in the experiment are shown in Fig. 15. Among them, Fig. 15a presents the verification results of the publicly available UR Fall Detection Dataset, and Fig. 15b shows the verification results of the author's own real-time video data. It can be seen that the YOLO-AB algorithm exhibits relatively high detection accuracy in recognizing falls, smoking, standing, and other situations on both datasets. The experiments demonstrate that the YOLO-AB algorithm proposed in this study has good robustness and generalization ability in the field of abnormal behavior detection, which provides strong support for the application of the algorithm in real life.

(a)

Figure 15: (Continued)

(**b**)

Figure 15: Verification results of the YOLO-AB algorithm on the author's real-time video data and the UR dataset; (a) Verification results of the YOLO-AB algorithm model on the UR Fall Detection Dataset; (b) Verification results of the YOLO-AB algorithm model in my own tests

5 Conclusions

In this study, a high-precision YOLO-AB algorithm for abnormal behavior detection was proposed by improving the structure of the YOLOv8 algorithm. This algorithm solved the problems of a single type of elderly abnormal behavior detection, low precision in the fused detection of multiple abnormal behaviors, insufficient feature extraction, and a high rate of missed detections. The improvement schemes include the modification of the backbone network, the replacement of the neck network part, and the substitution of the loss function. These improvements have effectively enhanced the detection performance. For two abnormal behaviors, falling and smoking, the algorithm can maintain stable and accurate detection capabilities, ensuring the accuracy and reliability of the detection results.

However, there is still room for further improvement in recall, which needs to be addressed in follow-up studies. In the future, we will focus on improving the recall rate and further improving the performance of the algorithm, which may require further improvement of the backbone network and detection head to better capture and reflect the characteristic information of abnormal behaviors of the elderly. In addition, we intend to broaden the application scope of the algorithm. We are devoted to exploring how the algorithm performs in detecting more kinds of abnormal behaviors among the elderly under diverse lighting conditions and at different distances from the camera. For instance, detecting staggering gaits in the dark or long bouts of inactivity from afar requires the construction of more datasets. It is expected to further improve the versatility and practicability of the algorithm so that it has a wider application prospect in the field of safe monitoring of the elderly.

Acknowledgement: The authors would like to thank the editors and reviewers for their valuable comments and suggestions. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding Statement: This work was supported by the Henan Provincial Science and Technology Research Project (242102211022), the Starry Sky Creative Space Innovation Space Innovation Incubation Project of Zhengzhou University of Light Industry (2023ZCKJ211) and Research and Practice Project of Higher Education Teaching Reform in Henan Province for Graduate Education (2023SJGLX160Y).

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Xianghong Cao; Data collection: Chenxu Li, Haoting Zhai; Analysis and interpretation of results: Chenxu Li, Haoting Zhai; Draft manuscript preparation: Xianghong Cao, Chenxu Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author C.L upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Fang EF, Xie C, Schenkel JA, Wu C, Long Q, Cui H, et al. A research agenda for ageing in China in the 21st century: focusing on basic and translational research, long-term care, policy and social networks. Ageing Res Rev. 2020;64(1):101174. doi:10.1016/j.arr.2020.101174.
2. Pani-Harreman KE, Bours GJ, Zander I, Kempen GI, van Duren JM. Definitions, key themes and aspects of 'ageing in place': a scoping review. Ageing Soc. 2020;41:1–34.
3. Laptev I. On space-time interest points. Int J Comput Vis. 2005;64(2):107–23. doi:10.1007/s11263-005-1838-7.
4. Zhang W. Robust registration of SAR and optical images based on deep learning and improved Harris algorithm. Sci Rep. 2022;12(1):5901. doi:10.1038/s41598-022-09952-w.
5. Chen H, Sun D, Liu W, Wu H, Liang M, Liu PX. A novel approach to the extraction of key points from 3-D rigid point cloud using 2-D images transformation. IEEE Trans Geosci Remote Sens. 2022;60(4):1–15. doi:10.1109/TGRS.2022.3175758.
6. Dollár P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. In: 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance; 2005 Oct 15–16; Beijing, China. p. 65–72.
7. Wang H, Ullah MM, Klaser A, Lapter I, Schmid C. Evaluation of local spatio-temporal features for action recognition. In: Bmvc 2009-British Machine Vision Conference; 2009 Sep 7–10; London, UK. p. 124.1–11.
8. Wang W, Zhao C, Li X, Zhang ZQ, Yuan X, Li H. Research on multimodal fusion recognition method of upper limb motion patterns. IEEE Trans Instrum Meas. 2023;72(8):1–12. doi:10.1109/TIM.2023.3289556.
9. Thummala J, Pumrin S. Fall detection using motion history image and shape deformation. In: 2020 8th International Electrical Engineering Congress (iEECON); 2020 Mar 4–6; Chiang Mai, Thailand. p. 1–4.
10. Ramanujam E, Perumal T, Padmavathi S. Human activity recognition with smartphone and wearable sensors using deep learning techniques: a review. IEEE Sens J. 2021;21(12):13029–40. doi:10.1109/JSEN.2021.3069927.
11. Attal F, Mohammed S, Dedabrishvili M, Chamroukhi F, Oukhellou L, Amirat Y. Physical human activity recognition using wearable sensors. Sensors. 2015;15(12):31314–38. doi:10.3390/s151229858.
12. Koutli M, Theologou N, Tryferidie A, Tzovaras D. Abnormal behavior detection for elderly people living alone leveraging IoT sensors. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE); 2019 Oct 28–30; Athens, Greece. p. 922–6.
13. Li J, Liang X, Shen SM, Xu T, Feng JS, Yan SC. Scale-aware fast R-CNN for pedestrian detection. IEEE Trans Multimed. 2017;20(4):985–96. doi:10.1109/TMM.2017.2759508.
14. Fu Y, Ran T, Xiao W, Yuan L, Zhao J, He L, et al. GD-YOLO: an improved convolutional neural network architecture for real-time detection of smoking and phone use behaviors. Digit Signal Process. 2024;151(4):104554. doi:10.1016/j.dsp.2024.104554.
15. Arifoglu D, Bouchachia A. Detection of abnormal behaviour for dementia sufferers using convolutional neural networks. Artif Intell Med. 2019;94(2):88–95. doi:10.1016/j.artmed.2019.01.005.
16. Li C, Li Y, Wang B, Zhang Y. Research into the applications of a multi-scale feature fusion model in the recognition of abnormal human behavior. Sensors. 2024;24(15):5064. doi:10.3390/s24155064.

17. He Y, Huang H, Wu Y, Zhu G. Research on abnormal behavior recognition of the elderly based on spatial-temporal feature fusion. In: Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences; 2022 Oct 13–15; Amsterdam, The Netherlands. p. 85–92.
18. Cao X, Zhang H. Falling detection research based on elderly behavior infrared video image contours ellipse fitting. Int J Pattern Recognit Artif Intell. 2021;35(2):2154004. doi:10.1142/S0218001421540045.
19. Cao X, Wang X, Geng X, Wu D, An H. An approach for human posture recognition based on the fusion PSE-CNN-BiGRU model. Comput Model Eng Sci. 2024;140(1):385–408. doi:10.32604/cmes.2024.046752.
20. Yan J, Wang X, Shi J, Hu S. Skeleton-based fall detection with multiple inertial sensors using spatial-temporal graph convolutional networks. Sensors. 2023;23(4):2153. doi:10.3390/s23042153.
21. Li S, Song X, Cao J, Xu S. Enhanced 3D residual network for human fall detection in video surveillance. KSII Trans Internet Inf Syst. 2022;16(12):3991–4007.
22. McCall S, Kolawole SS, Naz A, Gong L, Ahmed SW, Prasad PS, et al. Computer vision based transfer learning-aided transformer model for fall detection and prediction. IEEE Access. 2024;12(7):28798–809. doi:10.1109/ACCESS.2024.3368065.
23. Soni V, Yadav H, Bijrothiya S, Semwal V. CABMNet: an adaptive two-stage deep learning network for optimized spatial and temporal analysis in fall detection. Biomed Signal Process Control. 2024;96(8):106506. doi:10.1016/j.bspc.2024.106506.
24. Li S, Song X. Future frame prediction network for human fall detection in surveillance videos. IEEE Sens J. 2023;23(13):14460–70. doi:10.1109/JSEN.2023.3276891.
25. Kang KW, Park SY. The modified fall detection algorithm based on YOLO-KCF for elderly living alone care. J Inst Converg Signal Process. 2020;21(2):86–91.
26. Qin Y, Miao W, Qian C. A high-precision fall detection model based on dynamic convolution in complex scenes. Electronics. 2024;13(6):1141. doi:10.3390/electronics13061141.
27. Zhao D, Song T, Gao J, Li D, Niu Y. YOLO-fall: a novel convolutional neural network model for fall detection in open spaces. IEEE Access. 2024;12(3):26137–49. doi:10.1109/ACCESS.2024.3362958.
28. Ma Y, Yang J, Li Z, Ma Z. YOLO-cigarette: an effective YOLO network for outdoor smoking real-time object detection. In: Ninth International Conference on Advanced Cloud and Big Data (CBD); 2022 Mar 26–27; Xi'an, China. p. 121–6.
29. Senyurek VY, Imtiaz MH, Belsare P, Tiffany S, Sazonov E. Smoking detection based on regularity analysis of hand to mouth gestures. Biomed Signal Process Control. 2019;51:106–12. doi:10.1016/j.bspc.2019.01.026.
30. Hyuk JC. Detection of smoking in indoor environment using machine learning. Appl Sci. 2020;10(24):8912. doi:10.3390/app10248912.
31. Wang D, Yang J, Hou FH. Design of intelligent detection system for smoking based on improved YOLOv4. Sens Mater. 2022;34(8):3271. doi:10.18494/SAM3878.
32. Wang Z, Lei L, Shi P. Smoking behavior detection algorithm based on YOLOv8-MNC. Front Comput Neurosci. 2023;17:1243779. doi:10.3389/fncom.2023.1243779.
33. Li Y, Hou Q, Zheng Z, Cheng J, Yang MM, Li X. Large selective kernel network for remote sensing object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France. p. 16794–805.
34. Wang J, Chen K, Xu R, Liu ZW, Loy CC, Lin DH. Carafe: content-aware reassembly of features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 3007–16.
35. Du S, Zhang B, Zhang P. Scale-sensitive IOU loss: an improved regression loss function in remote sensing object detection. IEEE Access. 2021;9:141258–72. doi:10.1109/ACCESS.2021.3119562.
36. Zhang YF, Ren WQ, Zhang Z, Jia Z, Wang L, Tan TN. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing. 2022;506(9):146–57. doi:10.1016/j.neucom.2022.07.042.
37. Kwolek B, Kepski M. Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput Methods Programs Biomed. 2014;117(3):489–501. doi:10.1016/j.cmpb.2014.09.005.