Elderly Fall Detection Based on Improved YOLOv5s Network

TINGTING CHEN¹, ZHENGLONG DING²,³, (Graduate Student Member, IEEE), AND BIAO LI²

¹School of Network and Communication, Nanjing Vocational College of Information Technology, Nanjing 210023, China
²Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230601, China
³Key Laboratory of Electric Drive and Control of Anhui Higher Education Institutes, Anhui Polytechnic University, Wuhu, Anhui 241000, China

Corresponding author: Zhenglong Ding (zlding@mail.hfut.edu.cn)

This work was supported in part by the Natural Science Foundation of Jiangsu Higher Education Institutions of China under Grant 20KJB510032, in part by the Opening Project of Key Laboratory of Electric Drive and Control of Anhui Higher Education Institutes under Grant DQKJ2020004, and in part by the Nanjing Vocational College of Information Technology under Grant YK20190501.

ABSTRACT

The problem of aging population in our country is becoming more and more serious, falling on the road accidentally has been the first murder for people over 65 years of age. In this article, a real-time detection method for elderly fall behavior based on improved YOLOv5s is proposed to detect whether the elderly fall in real time, so that they can receive timely and effective treatment. First, the asymmetric convolution blocks (ACB) convolution module is used in the Backbone network to replace the existing basic convolution to improve the feature extraction capability. Then, the spatial attention mechanism module is added to the residual structure of the Backbone network to extract more feature location information. Finally, the feature layer structure is improved to remove the feature layer for small targets so that the network can pay more attention to the semantic level information, and at the same time, the classifier is set. The proposed algorithm is trained on the URFD public dataset, and the test set is used for verification. The experimental results show that the average accuracy of all categories of the algorithm reaches 97.2%, which is increased by 3.5% compared to YOLOv5s. Thus the proposed algorithm can accurately detect the fall behavior of the elderly.

INDEX TERMS

Elderly fall behavior detection, convolution blocks, YOLOv5s, attention mechanism, real-time detection.

I. INTRODUCTION

With the continuous development of economy and society, the problem of aging population in our country is becoming more and more serious. It is estimated that the number of people over 60 will exceed 300 million, accounting for 20.7% of the total population by 2025 [1]. With the continuous increasing in the number of elderly people, the number of elderly people living alone is also increasing day by day, which makes the daily safety of elderly people living alone become a hot topic for their children and society. Domestic research shows that falls have become the second leading cause of death in accidents and unintentional injuries, and it is also the leading cause of death due to injuries for people over 65 years of age [2], [3]. Medical surveys show that if effective treatment can be got in time after a fall, the risk of death can be reduced and the survival rate of the elderly can also be increased [4]. Therefore, an efficient and practical fall detection system for the elderly is needed to be built by advanced science and technology, which can detect and identify fall behaviors in time and send warning to reduce injuries caused by falls and also improve the quality of life of the elderly living alone. It is very necessary to research the fall detection of the elderly, which has important social significance and practical value [5], [6].

The current fall detection methods are mainly divided into three categories [7]: fall detection based on sensors deployed in environmental scenes, fall detection based on wearable sensor devices, and fall detection based on computer vision. For the method based on sensors deployed in environmental
scenes, various monitoring devices need to be installed in the elderly activity area and information such as pressure, vibration, and sound need to be collected to determine whether a fall has occurred. The detection area of this method has certain limitations, and the sensors are easily interfered by environmental factors, and the detection accuracy is poor [8], [9]. Fall detection based on wearable sensor devices requires that the devices, which contain sensors such as accelerometers, gyroscopes, and magnetic needles, are worn on the waist, limbs, or chest and back of the elderly. Then the sensor data is collected and processed to detect and analyze the movement of the elderly in a certain period, which can determine whether there is a fall. This method is simple to install and has a high detection rate, but the device needs to be worn all the time, which will have a certain impact for daily life. If the elderly forget to wear it, the state of the elderly cannot be detected in time, and the device needs to be charged in time, which is less convenient [10], [11]. Fall detection based on computer vision is that the video collected is processed to detect whether there is a fall behavior. This method has received widespread attention and has become a hot spot in fall detection research because of the character that it has a fixed camera to obtain continuous power supply for ensuring real-time monitoring, and that no devices are needed to be worn, so that it is not easy to be interfered by external factors, and that it has a high detection accuracy [12].

Traditional machine vision for feature selection is based on manual selection, and the classifiers are needed to be designed and trained based on specific detection objects. This method has a high subjectivity and a complex design process and it is easily affected by environmental factors. In recent years, convolutional neural network (CNN) has gradually been sought after by scholars in the field of deep learning because that the feature doesn’t need manual selection. Target detection methods based on CNN are mainly divided into two categories [13], one is a two-stage detection algorithm, which divides target detection into two steps, locating and recognition. Region-convolutional neural network (R-CNN) is the classic algorithm, which has low performance and cannot meet real-time requirements. Subsequent improvements are made on the basis of R-CNN, and fast regions with CNN (Fast R-CNN) [14] and faster regions with CNN (Faster R-CNN) [15] are introduced, but they are still far from meeting people’s requirements for real-time performance. The other is the one-stage detection algorithm, which optimizes the positioning and recognition of the target into one step. The classic models of this type of algorithm are the single shot multi-box detector (SSD) series and the you only look once (YOLO) series. In 2019, Lu et al. [16] proposed a fall detection method based on a three-dimensional convolutional neural network (3D CNN), and introduced a spatial visual attention mechanism based on long short-term memory (LSTM). In 2020, Chen et al. [17] proposed a method that used Mask-CNN and an attention guided Bi-directional LSTM model in a complex background to achieve fall detection, which had a certain degree of robustness. Zhang et al. [18] proposed a human fall detection algorithm based on temporal and spatial changes of body posture, and judged whether to fall or not by establishing a temporal and spatial evolution diagram of human behavior. In 2021, Zhu et al. [19] proposed an algorithm based on a deep vision sensor and a convolutional neural network. The convolutional neural network is used to train the extracted three-dimensional posture data of the human body to obtain a fall detection model, but the real-time performance is relatively low. Cao et al. [20] proposed a fall detection algorithm that combined motion features and deep learning. This method used you only look once version3 (YOLOv3) to detect human targets, and fused the human motion features with deep features extracted by CNN to distinguish whether a fall occurred.

With the change of the YOLO algorithm version, the latest YOLOv5 algorithm has been proposed. Compared with YOLOv3, the detection speed of YOLOv5 has been greatly improved on the basis of better accuracy, and the model is also smaller. At present, the YOLOv5 algorithm has not been widely used in the field of fall detection, so this article will improve the model based on the research of YOLOv5 and apply it to the fall behavior detection of the elderly.

The main contributions of this paper are summarized as follows:

1) The asymmetric convolution blocks (ACB) convolution module is used in the Backbone network to replace the existing basic convolution, which not only can extract the basic features, but also can extract the horizontal and vertical features, as well as the position and rotation features of the human body. Therefore, the improved Backbone network has stronger human feature extraction ability.

2) The spatial attention module is introduced into the residual structure of the Backbone network, which can extract more detailed information and improve the overall performance of the network.

3) The feature layer structure is improved and the feature layer of small targets is removed, so that the network can pay more attention to the semantic level information, and at the same time, the classifier is set.

This article first introduces the YOLOv5s network model and then describes some existing problems in the detection of elderly fall behavior. After that, Section III describe the proposed method in detail. Then, experiments have been carried out and the experimental results are analyzed in Section IV. Finally, the summary is given in Section V.

II. RELATED THEORIES
A. YOLOv5s ALGORITHM INTRODUCTION
The target detection network based on YOLOv5 is mainly divided into four network models: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x [21]. Among them, the YOLOv5s network model is the network with the smallest depth and the smallest feature map width in the series of YOLOv5, and the three models of YOLOv5m, YOLOv5l and YOLOv5x are...
the products of continuous deepening and widening on the basis of YOLOv5s [22]. The network structure of YOLOv5 consists of four parts: Input, Backbone, Neck and Prediction, the diagram of which is shown in Fig. 1.

The input of YOLOv5s uses the method of Mosaic data enhancement. The main idea is to perform random cropping, zooming and other operations on four randomly used images, and then stitch them together as training data, thus enriching the image background and making the network more robust, and reducing GPU calculations and increasing the universal applicability of the network. The input adopts adaptive anchor box calculation and adaptive image scaling. During each training process, the network will adaptively calculate the best anchor box in different training sets. After the scaling ratio and scaling size are calculated, a minimum filling value is obtained to adaptively scale and fill the original image. Therefore, the amount of calculation will be reduced and the target detection speed will be improved.

The Backbone of the network is mainly composed of Focus structure and cross stage partial (CSP) structure. Among them, the Focus structure is mainly used for slicing operations. In the network model of YOLOv5s, a normal image with a size of 608 × 608 × 3 is input into the network, and the input image is copied into four copies. The slicing operation will cut these four images into four slices, each of which has a size of 304 × 304 × 3, and then connect the four slices together, thus a feature map with the size of 304 × 304 × 12 is output. Then the feature map is input into convolution layer with a convolution kernel of 32 to become a feature map with the size of 304 × 304 × 32. The Focus module increases the speed by reducing the amount of calculation and the number of layers.

Neck uses feature pyramid networks (FPN) and pyramid attention network (PAN) structure. FPN transfers and integrates high-level feature information from top to bottom through up-sampling to convey strong semantic features. PAN is a bottom-up feature pyramid that conveys strong positioning features. Both are used at the same time to strengthen network feature fusion capabilities. In the figure, “Concat” means connection, which connects the four slices cut by the slicing operation in the Backbone of the network.

Prediction includes bounding box loss function and non-maximum suppression (NMS). The loss function of the bounding anchor box is improved from complete intersection over union (CIOU) loss to generalized intersection over union (GIoU) loss, which effectively solves the problem of non-coincidence of bounding boxes and improves the speed and accuracy of prediction box regression. In the post-processing process of target detection, YOLOv5 uses weighted NMS operation to filter multiple target anchor box, which enhances the recognition ability for multiple targets and occluded targets, and obtains the optimal target detection box.

Compared with YOLOv4, the Focus structure has been added to the Backbone network of YOLOv5. Different from the YOLOv4 network model that only uses the CSP structure in the Backbone network, the YOLOv5 network model designs two new CSP structures. Taking the YOLOv5s network model as an example, the Backbone network uses the CSP1_1 structure and the CSP1_3 structure, and the Neck uses the CSP2_1 structure to strengthen feature fusion between the networks.

B. PROBLEMS IN THE DETECTION OF FALLING BEHAVIOR USING YOLOV5S ALGORITHM

Due to the large differences in human clothing, posture, etc., the features are relatively complex, coupled with environmental factors such as the illumination of the human activity
scene, YOLOv5s has some problems in falling behavior detection: (1) YOLOv5s only uses $3 \times 3$ convolution to extract human body features, which can only extract basic features in the image, and it has insufficient ability to extract features such as rotation features. (2) The YOLOv5s algorithm is easy to lose some detailed information during feature extraction, resulting in false detection and missed detection.

III. THE PROPOSED METHOD

Aiming at the above problems of YOLOv5s in fall behavior detection, this paper mainly improves it from the following two aspects: (1) ACB convolution module is used in the Backbone network to replace the existing basic convolution, to improve the feature extraction ability of the Backbone network; (2) Introduce the spatial attention module into the residual structure of the Backbone network to extract more detailed information such as feature locations and improve the overall performance of the network.

A. ASYMMETRIC CONVOLUTION BLOCKS

Inspired by Ac.net [23], ACB is used in the YOLOv5s network to replace the original basic convolution. Specifically, it is to replace the existing $3 \times 3$ convolution kernel with ACB. As shown in Figure 2, the ACB contains three parallel layers with convolution kernel sizes $3 \times 3$, $1 \times 3$ and $3 \times 1$, where the $3 \times 3$ convolution kernel is a regular convolution that can extract the basic features in the abnormal human behavior image, and the other two convolution kernels are used to extract the horizontal and vertical features in the abnormal human behavior images, as well as the position and rotation features of the human body. Therefore, the improved Backbone network has stronger human feature extraction ability.

![FIGURE 2. Schematic diagram of ACB structure.](image)

According to the superposition principle in the convolution operation, the designed ACB module can directly replace the convolution kernel in the current YOLOv5s network. After the feature extraction of the image, it can be superimposed according to the operation method in formula (1), where $I$ is the input, and $K^1$ and $K^2$ are two convolution kernels of compatible sizes.

$$I \ast K^1 + I \ast K^2 = I \ast (K^1 \oplus K^2) \quad (1)$$

Similar to the conventional convolutional neural network, each layer is used as a branch after batch normalization operation, and then the outputs of the three branches are fused as the output of ACB. At this point, the network can be trained using the same configuration as the original model without tuning any additional hyper parameters. The specific implementation steps are as follows:

1. **BN normalization**
   $I \ast F \xrightarrow{BN} O_1 = (I \ast F - \mu) \frac{\gamma}{\sigma} + \beta \quad (2)$
   $O_2 = (I \ast F - \overline{\mu}) \frac{\overline{\gamma}}{\overline{\sigma}} + \overline{\beta} \quad (3)$
   $O_3 = (I \ast F - \hat{\mu}) \frac{\hat{\gamma}}{\hat{\sigma}} + \hat{\beta} \quad (4)$

   where, $I$ represents the input, let $F$, $\overline{F}$ and $\hat{F}$ be the convolution kernel of the $3 \times 3$ layer, $3 \times 1$ layer, and $1 \times 3$ layer. $O_1$, $O_2$ and $O_3$ respectively represent the normalized output of the corresponding convolutional layer branch. $\mu$, $\overline{\mu}$ and $\hat{\mu}$ are the batch normalized mean corresponding to the three convolution kernels, respectively. $\sigma$, $\overline{\sigma}$ and $\hat{\sigma}$ are the variances corresponding to the three convolution kernels. $\gamma$, $\overline{\gamma}$ and $\hat{\gamma}$ are the weights learned by the corresponding convolution kernel. $\beta$, $\overline{\beta}$ and $\hat{\beta}$ are the learned biases corresponding to the convolution kernels.

2. **Branch fusion**
   $O = O_1 + O_2 + O_3 = F' + b \quad (5)$
   $F' = \frac{\gamma}{\sigma} F + \frac{\overline{\gamma}}{\overline{\sigma}} \overline{F} + \frac{\hat{\gamma}}{\hat{\sigma}} \hat{F} \quad (6)$
   $b = \frac{\mu \gamma}{\sigma} - \frac{\mu \overline{\gamma}}{\overline{\sigma}} - \frac{\mu \hat{\gamma}}{\hat{\sigma}} + \beta + \overline{\beta} + \hat{\beta} \quad (7)$

   where $O$ represents the output of the ACB convolution block, $F'$ represents the fused convolution kernel, $b$ represents the fused bias.

   In the training phase of the network, the convolution kernels in the proposed ACB are trained separately. In the later inference phase, the weights of the three convolution kernels are fused into a regular convolution form through an algorithm, and then the inference operation is performed. Therefore, the actual inference time does not increase.

In this paper, the ACB convolution block is used to replace the convolution kernels in different positions of the YOLOv5s model, and the detection results are tested. According to the structural characteristics of the network model of YOLOv5s, the ACB is used to replace the basic convolution of Backbone, Neck and Prediction respectively. The specific positions are shown in Figure 3(a), 3(b) and 3(c), and the corresponding networks are represented by ACB-YOLOv5s-Backbone, ACB-YOLOv5s-Neck and ACB-YOLOv5s-Prediction, respectively.

The network after replacing the basic convolution in three different positions with the ACB convolution module is compared with the original network. The results are shown in Table 1. AP50% refers to the average accuracy (AP) when the IoU threshold is 0.5. mAP@ 0.5% refers to the mean
average precision (mAP) of each category when IoU threshold is 0.5.

As can be seen from Table 1, using the ACB convolution block to replace the base convolution of the CSP1 structure in the Backbone network improves the mean average precision by 2.1%. However, in the Neck and Prediction modules, mAP is reduced by 0.9% and 0.3%. Therefore, the ACB convolution module is used in the Backbone network to replace the basic convolution, which can improve the detection ability of the model.

### Table 1. Comparison of ACB replacement results.

| Model                      | AP50/% | mAP@ 0.5% |
|----------------------------|--------|-----------|
| YOLOv5s                    | 94.2   | 93.7      |
| ACB-YOLOv5s-Backbone       | 96.5   | 95.8      |
| ACB-YOLOv5s-Neck           | 93.9   | 92.8      |
| ACB-YOLOv5s-Prediction     | 94.1   | 93.4      |

### B. ATTENTION MECHANISM

The attention mechanism is a resource allocation strategy, which is very similar to human visual attention and is widely used in many directions of computer vision [24], [25]. By adding a visual attention mechanism to the convolutional neural network, the network itself can pay more attention to the target area that needs to be Focused, and selectively ignore some irrelevant information to improve the overall performance of the network. The convolutional block attention module (CBAM) [26] is a hybrid domain attention mechanism composed of channel attention and spatial attention in series. Channel attention enhances the network’s attention to meaningful input features, and helps to improve the granularity of resource allocation between convolutional channels.

Spatial attention preserves key information when spatial information of the original image is transformed into another space, which helps the network pay more attention to the feature location information. Considering that this article detects whether the elderly falls, there are only two categories. Therefore, it has lower requirements for the classification ability of the network model, but higher requirements for the positioning ability. Combining with the idea of lightweight, this article only uses the spatial attention module (SAM) in CBAM. SAM is to perform maximum pooling and average pooling operations on the input feature map in the channel dimension to generate two 2-dimensional spatial feature map matrices. The two feature maps are spliced in the channel dimension, and then a 7 × 7 convolutional layer is used to optimize the weights. Then the optimized feature map is input into Sigmoid activation function to obtain the spatial attention map. Finally, the new feature of spatial attention can be obtained by multiplying the two map point by point. SAM is defined as follows:

\[
M_S(F) = \sigma(f^{7 \times 7}(P_{\text{avg}}(F); P_{\text{max}}(F)))
\]

where \( F \) is input feature map, \( P_{\text{max}} \) and \( P_{\text{avg}} \) denote maximum pooling and average pooling operations respectively, \( f^{7 \times 7} \) is a 7 × 7 convolutional layer, \( \sigma() \) is Sigmoid activation function, \( M_S \) is spatial attention map. Figure 4 shows the schematic diagram of the spatial attention mechanism.

The detection model used in this article is YOLOv5s. In order to further enhance the network’s ability to extract...
features of elderly fall behavior and improve the accuracy of fall detection. SAM is added to the residual structure of the Backbone part, which can increase the receptive field of the network and adaptively refine the features. The improved residual structure is shown in the Fig. 5.

FIGURE 5. Diagram of improved residual structure.

C. IMPROVED FEATURE LAYER STRUCTURE

Based on the YOLOv5s model, three different scale feature layers of 19 × 19, 38 × 38, and 76 × 76 are used to predict large, medium and small targets. The smaller the size of the feature layer, the larger the neuron’s receptive field, which means that the semantic level is richer, but local and detailed features will be lost. On the contrary, when the convolutional neural network is shallower and the receptive field becomes smaller, the neurons in the feature map tend to be local and detailed information. 76 × 76 is mainly used to predict a target with a smaller size. In order to adapt to the size characteristics of human body of this dataset, the 76 × 76 feature layer is removed, while the 19 × 19 and 38 × 38 feature layers are retained for prediction and the human behavior feature detection layer is established.

D. CLASSIFIER SETTINGS

The classifier contains 80 categories of different sizes in the original model of YOLOv5. After clustering, the classifier needs to be modified. The model uses multi-scale feature layers to detect targets of different sizes. The YOLOv5s model sets 3 prediction boxes for each network unit, and each prediction box contains 5 basic parameters (x, y, w, h, confidence), and requires probabilities of 80 categories, so the dimension of the model output is 3 × (5 + 80) = 255. In this paper, fall behavior of the elderly is divided into two categories of fall and up according to the needs, so the output dimension tensor is 3 × (5 + 2) = 21. Therefore, in this experiment, the classifier is modified on the basis of the original model, the output of which is 21-dimension tensor.

Based on the improvements in the above aspects, the schematic diagram of the improved YOLOv5s network structure is shown in Figure 6.

IV. EXPERIMENTAL RESULTS AND ANALYSIS FOR ELDERLY FALL BEHAVIOR

A. LAB ENVIRONMENT

The experimental dataset is the public dataset of UR Fall Detection Dataset (URFD) which is collected from the Interdisciplinary Center of Computational Modeling, University of Rzeszow, Poland [27]. The dataset includes 70 videos, which consists of 40 videos of daily life behaviors and 30 videos of falling behaviors. The daily life behavior videos include actions such as bending over, squatting, and sitting down. The falling behaviors include the process from walking upright to falling and the process from sitting on a chair to falling. There are several types of falling in different directions, falling forward and backward. These videos were taken from two perspectives, parallel to the ground and looking down on the ground. The size of each frame is 720 × 480, and the frame rate is 30fps. A part of the data set is shown in Fig. 7.

B. LAB ENVIRONMENT AND TRAINING

The experimental environment of this article is: operating system Windows10, processor Intel(R) Core(TM) i7-8550U, image processor GeForce GTX3080, deep learning framework PyTorch, compute unified device architecture (CUDA) parallel computing architecture, and the CUDA deep neural network (CU-DNN) acceleration library is integrated into the PyTorch framework to accelerate computer computing capabilities. The development environment is PyCharm and the programming language Python 3.6.

When training the improved YOLOv5s model, the initial value of the learning rate is 0.001, and a total of 300 epochs are set, and the learning rate momentum is set to 0.925. Figure 8 below shows the loss function curve of the model structure. From the graph, we can see that during the training process of the YOLOv5s model, the value of the loss function drops sharply from 0 to 40 epochs, and starts to converge near the 50th epoch, with a faster convergence rate.

C. EVALUATE

In the field of target detection, precision (P), recall (R), average precision (AP) and mean average precision (mAP) are commonly used as indicators to evaluate the performance of training models. They are defined as follows:

\[ P = \frac{TP}{TP + FP} \]  
\[ R = \frac{TP}{TP + FN} \]  
\[ AP = \sum_{i=1}^{n} P(i) \Delta r(i) = \int_0^1 p(r)dr \]  
\[ mAP = \frac{\sum_{n=1}^{N} AP(n)}{N} \]
where TP is the number that positive samples are predicted to be positive, FP is the number that negative samples are predicted to be positive, TN is the number that negative samples are predicted to be negative, FN is the number that positive samples are predicted to be negative, n is category, N is class number.

The relationship between P and R can be expressed by the PR curve. PR curve during model training is shown in Fig. 9, where the horizontal axis is the recall rate, and the vertical axis is the precision rate.

D. EXPERIMENTAL RESULTS ANALYSIS

In order to evaluate the impact of different improvement methods on the performance of the model on the detection of elderly falling behavior, an ablation experiment has been performed on the URFD public dataset, and the effects of different improvements are analyzed, where F represents the improved feature layer. The experimental results are shown in Table 2.

From the first two rows in Table 2, we can see that the ACB convolution block designed in the Backbone network improves the evaluation indicators AP and mAP by 2.3% and 2.1%, respectively. It can be seen that the ACB convolution block can enhance the feature extraction ability of the Backbone network for the detection target and improve the detection effect. From the first and third rows in the
TABLE 2. Evaluation indicators comparison of different improvements.

| ACB | SAM | F   | AP50% | mAP@0.5% |
|-----|-----|-----|-------|----------|
|     |     |     | 94.2  | 93.7     |
| √   |     |     | 96.5  | 95.8     |
|     |     |     | 96.1  | 95.4     |
| √   |     |     | 94.9  | 94.3     |
|     |    √|     | 93.9  | 92.8     |
| √   |    √|     | 94.1  | 93.4     |

Table, we can see that the introduction of the spatial attention mechanism in the Backbone network increases the evaluation indicators AP and mAP by 1.9% and 1.7%, respectively. It can be seen that the spatial attention mechanism is conducive to focusing on feature positions information to make fall detection more accurate. From the first row and the fourth row, we can see that using two-scale feature layers to predict the fall behavior of the elderly can accurately classify the behavior of the elderly while reducing the amount of calculation. From the fifth and sixth rows, we can see that when different improvement methods are superimposed, the performance of the model is not directly superimposed, but is further improved slightly on the basis of a certain improvement. To sum up, the superposition of different improvement methods improves the detection ability of the model, indicating that the improvement is feasible and necessary.

The improved YOLOv5s model is used to test on the URFD dataset, and some of the detection results are shown in Figure 10 and Figure 11. Fall behavior is represented by down, and non-fall behavior is represented by up.

As can be seen from Figure 10, the improved YOLOv5s model has a good detection effect for different objects and different forms of falls. It can be seen from Figure 11 that in daily activities, the improved model also has good detection results for non-falling behaviors whose postures are similar to falling behaviors in different scenes and under different lighting conditions. In order to more intuitively show the detection effect of different algorithms, YOLOv5s model and the improved YOLOv5s model are used for detection on the dataset. In addition to this, we also have collected 300 images of different life scenes as a validation set, and some test results are shown in the figure 12 and 13.

It can be seen from figure 12 that the improved YOLOv5s model performs better for daily activities in terms of recognition probability and accuracy. 12(c) and 12(d) show that YOLOv5s model has false detection for non-falling behaviors whose postures are similar to falling behaviors. From figure 13, we can see that improved YOLOv5s model also has a better performance on self-built dataset. 13(a) shows that YOLOv5s model has a missed detection, 13(b) shows that YOLOv5s model has a false detection.
In order to further verify that the improved YOLOv5s algorithm has a better effect on the detection of falling behavior of the elderly, the same number of test sets are used to conduct comparative experiments with other mainstream algorithms under the same configuration conditions. AP and mAP are selected as evaluation indicators, and the performance comparison of different algorithms is shown in Table 3.

| Algorithm       | AP50/% | mAP@0.5% |
|-----------------|--------|----------|
| SSD             | 77.2   | 76.6     |
| Faster R-CNN    | 84.3   | 85.3     |
| YOLOv3          | 89.7   | 89.2     |
| YOLOv5s         | 94.2   | 93.7     |
| YOLOv5s+CBAM    | 96.1   | 95.5     |
| Ours            | 97.8   | 97.2     |

It can be seen from the comprehensive index mAP that the algorithm in this paper is 3.5% higher than the original YOLOv5s model, and the detection effect is the best among several mainstream algorithms, with mAP reaching 97.2%, which can accurately detect the elderly fall behavior.

V. CONCLUSION

In order to improve the behavioral safety of the elderly, especially the elderly living alone, an improved YOLOv5s algorithm is proposed in this paper. In the Backbone network, the ACB convolution block is used to replace the existing basic convolution, which improves the feature extraction ability. The spatial attention mechanism module is added to the residual structure, which makes the network pay more attention to the feature location information and has stronger localization ability. At the same time, the feature layer structure is improved, and the classifier is set, so that the improved network can better detect the fall behavior of the elderly. The experimental results show that the average accuracy of all categories of the algorithm reaches 97.2%, which is increased by 3.5% compared to YOLOv5s, which improves the accuracy of fall detection and recognition for the elderly and has certain practical value for real-time detection and early warning of falls.

In future work, we will continue to explore how to reduce the number of network model parameters and improve the detection rate of the network model.

REFERENCES

[1] Y. M. Chen, Z. F. Liu, X. D. Li, and Y. X. Huang, “The aging trend of Chinese population and the prediction of aging population in 2015–2050,” Chin. J. Social Med., vol. 35, no. 5, pp. 480–483, 2018.
[2] M. Zhao, M. Yu, and S. K. Zhu, “The prevalence of falls in the elderly in the community and the progress of prevention,” Injury Med., vol. 7, no. 1, pp. 61–66, 2018.
[3] Y. Chen, R. Du, K. Luo, and Y. Xiao, “Fall detection system based on real-time pose estimation and SVM,” in Proc. IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE), Mar. 2021, pp. 990–993.
[4] C. Mao, “Research Progress of intervention for fear of falling in the aged at home and abroad,” Chin. J. Modern Nursing, vol. 24, no. 7, pp. 865–868, 2018.
[5] F. F. Liu, “Research on detection and recognition of indoor falling behavior based on video surveillance,” Shan Dong Univ., Jinan, China, Tech. Rep., 2016.
[6] J. Q. Ma, H. Lei, and M. Y. Chen, “Fall behavior detection algorithm for the elderly based on AlphaPose optimization model,” J. Comput. Appl., vol. 42, no. 1, pp. 294–301, 2022.

[7] L. Ren and Y. Peng, “Research of fall detection and fall prevention technologies: A systematic review,” IEEE Access, vol. 7, pp. 77702–77722, 2019.

[8] K. Wang, G. Zhan, and W. Chen, “A new approach for IoT-based fall detection system using commodity mmWave sensors,” in Proc. 7th Int. Conf. Inf. Technol., IoT Smart City, Dec. 2019, pp. 197–201.

[9] L. Ma and N. Wang, “Room-level fall detection based on ultra-wideband (UWB) monostatic radar and convolutional long short-term memory (LSTM),” Sensors, vol. 20, no. 4, pp. 1105–1106, 2020.

[10] Z. Sheng-lan, Y. Yi-fan, G. Li-fu, and W. Diao, “Research and design of a fall detection system based on multi-axis sensor,” in Proc. 4th Int. Conf. Intell. Inf. Process., Nov. 2019, pp. 303–309.

[11] Y. P. Er and K. K. Tan, “Wearable solution for robust fall detection,” in Assistive Technology for the Elderly. Cambridge, MA, USA: Academic Press, 2020, pp. 81–105.

[12] L. J. Zhu, Z. Y. Chen, and C. L. Tian, “Review of fall detection method based on wearable devices,” Comput. Eng. Appl., vol. 55, no. 18, pp. 8–14, 2019.

[13] Z. Sheng-lan, Y. Yi-fan, G. Li-fu, and W. Diao, “Research and design of a fall detection system based on multi-axis sensor,” in Proc. 4th Int. Conf. Intell. Inf. Process., Nov. 2019, pp. 303–309.

[14] K.-H. Chen, Y.-W. Hsu, J.-J. Yang, and F.-S. Jaw, “Enhanced characterization of an accelerometer-based fall detection algorithm using a repository,” Instrum. Sci. Technol., vol. 45, no. 4, pp. 382–391, Jan. 2017.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[16] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[17] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[18] N. Lu, Y. Wu, L. Feng, and J. Song, “Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data,” IEEE J. Biomed. Health Inform., vol. 23, no. 1, pp. 314–323, Jan. 2019.

[19] Y. Chen, W. Li, L. Wang, J. Hu, and M. Ye, “Vision-based fall event detection in complex background using attention guided bi-directional LSTM,” IEEE Access, vol. 8, pp. 161337–161348, 2020.

[20] J. Zhang, C. Wu, and Y. Wang, “Human fall detection based on body posture spatio-temporal evolution,” Sensors, vol. 20, no. 3, p. 946, Feb. 2020.

[21] Y. Zhu, Y. P. Zhang, and S. S. Li, “Fall detection algorithm based on deep vision sensor and convolutional neural network,” Opt. Technique, vol. 47, no. 1, pp. 56–61, 2021.

[22] L. Z. Wu, X. L. Wang, and Q. Zhang, “An object detection method of falling person based on optimized YOLOv5s,” J. Graph., pp. 1–13, 2022. [Online]. Available: https://kns.cnki.net/kcms/detail/10.1034.T20220629.1803.002.html

[23] S. L. Zhang, L. P. Zhang, and W. Q. Zheng, “Identification and localization of walnut varieties based on YOLOv5s,” J. Chin. Agricult. Mechanization, vol. 43, no. 7, pp. 167–172,2022.

[24] J. R. Cao, J. J. Lu, and X. Y. Wu, “Fall detection algorithm combining motion features and deep learning,” Comput. Appl., vol. 41, no. 2, pp. 583–589, 2021.

[25] C. J. Xu, X. F. Wang, and Y. D. Yang, “Attention-Yolo: Yolo detection algorithm with attention mechanism,” Comput. Eng. Appl., vol. 55, no. 6, pp. 13–23, 2019.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. ECCV (Lecture Notes in Computer Science), vol. 11211. Cham, Switzerland: Springer, 2018, pp. 3–19.

[27] B. Kwok and M. Kepski, “Human fall detection on embedded platform using depth maps and wireless accelerometer,” Comput. Methods Programs Biomed., vol. 117, no. 3, pp. 489–501, Dec. 2014.

**TINGTING CHEN** was born in Hebei, China, in 1987. She received the M.S. degree from the Beijing Jiaotong University, in 2015. She is currently a Lecturer with the Nanjing Vocational College of Information Technology. Her main research interests include human action recognition and image processing.

**ZHENGLONG DING** (Graduate Student Member, IEEE) was born in Wuhu, Anhui, China, in 1988. He received the M.S. degree in mechanical engineering from Zhejiang University (ZJU), Hangzhou, China, in 2015. He is currently pursuing the Ph.D. degree in information and communication engineering with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei. Since 2019, he has been an Associate Professor with the Anhui Institute of Information Technology, Wuhu, China. His research interests include image processing, object detection, and automatic measurement.

**BIAO LI** was born in Huaibei, Anhui, China, in 1993. He received the B.S. degree in mechanical engineering from the Anhui Institute of Information Technology, Wuhu, China, in 2018. He is currently pursuing the M.S. degree with the Hefei University of Technology, Hefei. His research interests include image processing and automatic measurement.

**TINGTING CHEN** was born in Hebei, China, in 1987. She received the M.S. degree from the Beijing Jiaotong University, in 2015. She is currently a Lecturer with the Nanjing Vocational College of Information Technology. Her main research interests include human action recognition and image processing.