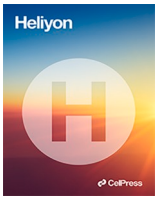
Heliyon 10 (2024) e31614

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Integrating attention mechanism and multi-scale feature extraction for fall detection

Hao Chen [a,*], Wenye Gu [b], Qiong Zhang [a], Xiujing Li [a], Xiaojing Jiang [a]

[a] School of Computer and Information Engineering, Nantong Institute of Technology, China
[b] Affiliated Hospital of Nantong University, China

ARTICLE INFO

Keywords:
Fall events
Spatial attention
Efficient channel attention
Spatial pyramid pooling

ABSTRACT

Addressing the critical need for accurate fall event detection due to their potentially severe impacts, this paper introduces the Spatial Channel and Pooling Enhanced You Only Look Once version 5 small (SCPE-YOLOv5s) model. Fall events pose a challenge for detection due to their varying scales and subtle pose features. To address this problem, SCPE-YOLOv5s introduces spatial attention to the Efficient Channel Attention (ECA) network, which significantly enhances the model's ability to extract features from spatial pose distribution. Moreover, the model integrates average pooling layers into the Spatial Pyramid Pooling (SPP) network to support the multi-scale extraction of fall poses. Meanwhile, by incorporating the ECA network into SPP, the model effectively combines global and local features to further enhance the feature extraction. This paper validates the SCPE-YOLOv5s on a public dataset, demonstrating that it achieves a mean Average Precision of 88.29 %, outperforming the You Only Look Once version 5 small by 4.87 %. Additionally, the model achieves 57.4 frames per second. Therefore, SCPE-YOLOv5s provides a novel solution for fall event detection.

1. Introduction

The prevention of fall events is a crucial task in daily life [1]. The risk of falling can be significantly increased by mobility limitations or environmental factors. Fall events can result in physical injuries, such as fractures and abrasions, as well as a negative impact on a person's quality of life. Therefore, real-time detection of fall events is very significant for injury prevention and personal safety [2].

Traditional detection methods mainly rely on wearable sensors to detect fall events by analyzing the user's motion data [3,4]. This method is effective but has limitations, such as inconvenience of wearing and personal privacy [5]. Video surveillance techniques have gained a lot of attention as technology has advanced. This non-invasive method recognizes fall events through images or video data acquired by cameras. This method is highly accurate and real-time. Therefore, vision-based fall detection methods have become a new focus of fall detection research [6].

With the development of deep learning, Convolutional Neural Network (CNN)-based methods are widely used for target detection [7]. Among them, the You Only Look Once (YOLO) model is widely used in various scenarios [8–10], due to its efficient target detection abilities. However, the YOLO model faces two major challenges in fall detection. Firstly, the fall pose has significant scale variation, which complicates the model's ability to extract multi-scale features. To address this problem, researchers have attempted to

* Corresponding author.
E-mail address: chenhao@ntit.edu.cn (H. Chen).

https://doi.org/10.1016/j.heliyon.2024.e31614
Received 10 February 2024; Received in revised form 11 May 2024; Accepted 20 May 2024
Available online 21 May 2024

enhance the feature extraction module of the YOLO model, aiming to enable it to capture these multi-scale features more efficiently and thus identify fall events more accurately. Secondly, fall postures often involve subtle pose features, such as minor movements of body parts and slight changes in balance. These subtle features are crucial for accurately distinguishing normal activities from fall events, but traditional feature extraction methods often struggle to capture these subtle changes. Therefore, further optimization of the YOLO model is necessary to better focus on the spatial and positional features of falls, to identify fall events more accurately.

The Efficient Channel Attention (ECA) [11] network primarily focuses on channel attention but pays slightly less attention to spatial attention. As a result, when it comes to the spatial distribution of various pose and shape features, models may find it challenging to extract this critical information, posing difficulties in distinguishing between normal activities and fall events. The Spatial Pyramid Pooling (SPP) [12] network can capture features on different scales, which is crucial for recognizing fall events on various scales. However, traditional SPP networks employ a max-pooling strategy, which causes the model to focus more on local features in the image while ignoring the broader background information. In fall detection, the background information is crucial for distinguishing fall events from daily activities. Therefore, traditional SPP networks may lead to misclassification by ignoring this background information. Moreover, fall events usually involve multiple local features and contextual information. However, traditional SPP networks do not fully consider spatial relationships and inter-channel dependencies in images, which limits the performance of the model in complex scenes.

In summary, this paper builds on the You Only Look Once version 5 small (YOLOv5s) and proposes the Spatial Channel and Pooling Enhanced YOLOv5s (SCPE-YOLOv5s) model. The choice of YOLOv5s is justified by existing research that has demonstrated its effectiveness in object detection tasks [13]. The most important contributions of this paper are as follows:

(1) An improved ECA network is proposed and added to the enhanced feature extraction layer of YOLOv5s. By enhancing the spatial attention mechanism, this network enables the model to understand the spatial distribution of the feature's pose more accurately in the fall detection scene, thereby effectively differentiate between normal activities and fall events.
(2) An improved SPP network is proposed. The average pooling layers are introduced into the SPP network, which helps the model to extract multi-scale features of fall events and enhances the model's ability to capture background information in the fall environment. Additionally, an improved ECA network is incorporated into the SPP network to further enhance the global and local feature extraction abilities.
(3) This paper conducted experiments on a public dataset to validate the effectiveness of the proposed model. The experimental results show that, when compared to other state-of-the-art methods, the proposed model is optimal for detecting fall events.

The remainder of the paper is organized as follows: Section 2 describes related work. Section 3 describes the proposed model in detail. Section 4 conducts the experiments and analyzes the results. A discussion is provided in Section 5. A conclusion is provided in Section 6.

2. Related work

Table 1 illustrates a summary of related work. Traditional methods of fall detection rely heavily on wearable sensors. For example, Kerdjidj et al. [14] used a wearable Shimmer device that transmits inertial signals to a computer through a wireless connection, enabling fall detection in elderly patients. Chander et al. [15] proposed a method using a soft robotic stretch sensor to monitor human movement and perform fall detection. This method provides higher flexibility and comfort. Additionally, Alarifi et al. [16] collected information through wearable devices, performed feature analysis and dimensionality reduction, and effectively extracted key features related to falls. The advantage of this method is its ability to improve detection accuracy through machine learning algorithms. Nooruddin et al. [17] proposed an Internet of Things (IoT) system that can be deployed on any type of device. The generality of this method allows it to be adapted to different scenarios and user requirements. Additionally, Al Nahian et al. [18] proposed a novel fall detection scheme based on wearable accelerometer data, where the main features are extracted through feature reduction techniques. Although obtaining fall information through sensors is a very direct and practical method, its limitations cannot be ignored. Firstly, wearable devices can be challenging for older people to wear. Secondly, since the sensors need to continuously collect the user's movement data, these data may include sensitive personal information, posing significant privacy concerns. Therefore, the application

Table 1
Related work summary.

Category	Reference	Description	Key Feature
Wearable Devices	[14]	A wearable Shimmer device.	It offers real-time, precise monitoring but is uncomfortable to wear and may raise privacy concerns.
	[15]	Soft robotic stretch sensors.	
	[16]	A tri-axial device with a magnetometer, gyroscope, and accelerometer.	
	[17]	An IoT system.	
	[18]	Wearable accelerometer sensors.	
YOLO	[19,20]	YOLOv5 and lightweight networks.	It offers non-contact monitoring for various posture scales, but extracting subtle spatial and positional features is challenging.
	[21–23]	Enhanced YOLO with multi-scale features.	
	[24–26]	Enhanced YOLO with attention modules.	
	[27,28]		

of the above methods in the field of fall detection still faces many challenges.

With the rapid development of computer vision technology, video-based fall detection methods have gradually received attention from researchers. Among them, the YOLO model has attracted attention for its efficient target detection ability. Several researches have employed YOLO for the detection of fall events. For example, Bo et al. [19] compared the performance of different YOLO models in fall detection. Kan et al. [20] combined the group shuffle convolution to implement the lightweight YOLOv5, which enhanced the detection abilities. However, these methods still face challenges in dealing with fall poses on different scales. To improve the detection accuracy, researchers improved the design of the feature extraction module. Fan et al. [21] considered multi-scale features in their design to better capture local features and contextual information when a person falls. Lyu et al. [22] enhanced the SPP network by adding an additional 1×1 max-pooling layer, achieving significant detection results. Additionally, to effectively extract multi-scale features, Abas et al. [23] proposed a CNN model combining multiple max-pooling layers and combined it with YOLO to achieve accurate target detection. This suggests that the introduction of multiple pooling layers is crucial for the effective extraction of multi-scale features of fall events. Furthermore, to distinguish between normal activities and fall events, the model needs to capture more detailed features more accurately. For this reason, researchers have attempted to enhance the performance of the YOLO model by introducing an attention mechanism. It is because the introduction of the attention mechanism enables the model to better focus on the

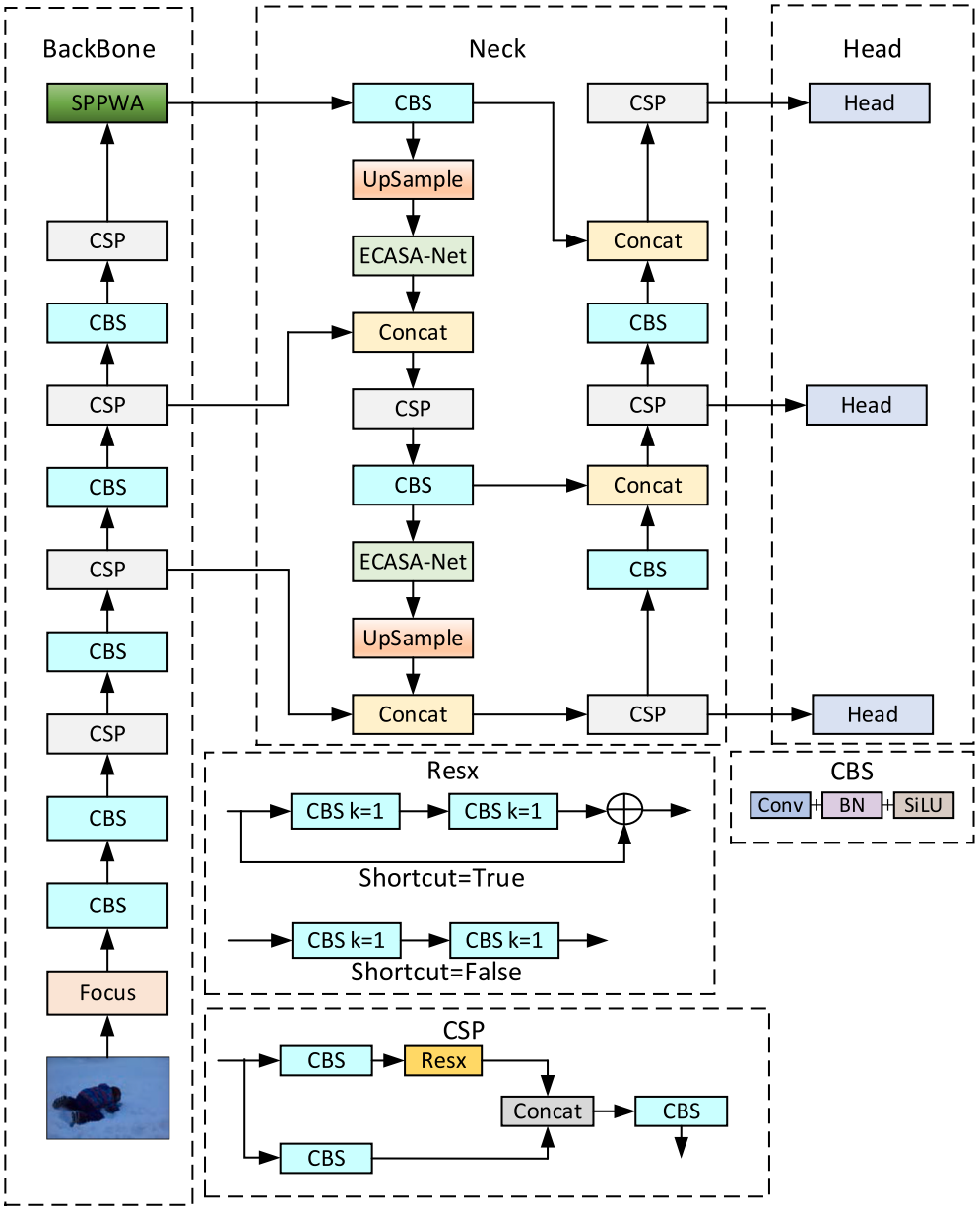


Fig. 1. SCPE-YOLOv5s structure.

key information in the image [24,25]. For example, Chen et al. [26] introduced a spatial attention module in the backbone network to extract more location features. Zhao et al. [27] improved the feature extraction ability for fall target detection by enhancing both the coordinate attention and shuffle attention mechanisms. Wang et al. [28] proposed adding squeeze-and-excitation networks to the last layer of the backbone network to further enhance feature extraction abilities. Given the features of fall movements, it is crucial to pay attention to both spatial and positional information. Therefore, introducing a multi-scale detection module and attention mechanism is of great significance for addressing the problems of varying scales and subtle gesture features in fall detection.

In summary, although traditional fall detection methods have been automated to a certain extent, their application still faces many challenges due to the inconvenience of wearing devices and privacy concerns. Computer vision-based fall detection methods, especially the target detection method using the YOLO model, offer new insights for fall detection.

3. Methodology

Fig. 1 illustrates the structure of SCPE-YOLOv5s. The model consists of three parts: the BackBone, the Neck, and the Head. Firstly, the BackBone is the main backbone feature extraction network of the model. It includes Focus, Conv-BatchNorm2d-Sigmoid-weighted Linear Unit (SiLU) (CBS), Cross Stage Partial (CSP) and Spatial Pyramid Pooling with Attention (SPPA). Focus is responsible for up-sampling operations and expanding the number of channels. CBS is a combined module integrating convolution, batch normalization, and SiLU activation function. CSP, a special network structure, is designed to extract and integrate the backbone features. SPPA is an improved version of SPP, efficiently enlarges the sensory field of the network by combining average pooling and max-pooling. Next, the Neck is the enhanced feature extraction network. It mainly consists of the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN). In the FPN, the effective feature layers that have been obtained are used to continue extracting features. PAN realizes the fusion between different layers of features through up-sampling and down-sampling operations. After each up-sampling step, an Efficient Channel and Spatial Attention (ECSA) network is introduced. This attention mechanism module allows the model to understand and analyze the image content more comprehensively. Finally, the Head is responsible for predicting the location and category of fall events.

3.1. ECSA network

In fall scenarios, the effectiveness of feature detection can vary significantly due to environmental and other factors. Fall detection requires not only recognizing a person's pose and shape but also understanding the spatial distribution of these features. Spatial attention mechanisms allow the model to focus more accurately on areas where a fall is likely to occur. For example, if a person falls to the floor, the model can use the spatial attention mechanism to focus its attention on the person's pose on the floor while ignoring background regions that are not related to the fall event.

Fig. 2 shows the structure of the ECSA. Suppose the input feature map is denoted as $F \in R^{C\times W\times H}$, where C represents the number of channels, H represents the height, and W represents the width. The output obtained after average pooling is denoted as $F' \in R^{1\times W\times H}$. The size of the convolution kernel for one-dimensional convolution is adapted automatically by a function [11]. The function k can be represented as shown in Equation (1):

$$k = \left| \frac{log_2^C}{\gamma} + \frac{a}{\gamma} \right|, \tag{1}$$

where γ equals 2, and a equals 1. A one-dimensional convolution is applied on the output after average pooling. Subsequently, attention weights are generated using a Sigmoid function activation. The resulting output F'' can be expressed as shown in Equation (2):

$$F'' = Sigmoid(Conv1d(F', \omega, k)), \tag{2}$$

where $Conv1d$ denotes a one-dimensional convolution operation, and ω is the weight of the convolution kernel. Following this, the feature map $F_{channel}$ after channel attention-adjustment is represented as shown in Equation (3):

$$F_{channel} = F \odot F'', \tag{3}$$

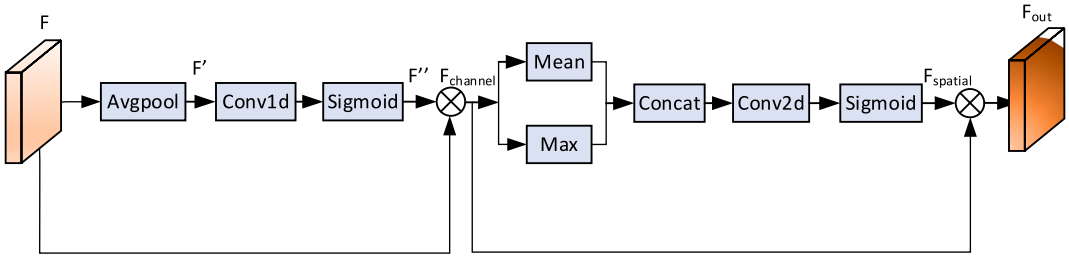


Fig. 2. ECSA structure.

where $\odot$ denotes element-by-element multiplication. Next, spatial attention is considered, and the average and maximum values of the channel attention-adjustment feature maps are calculated. They are represented as shown in Equations (4) and (5):

$$\max\_out = max(X_{channel}, \dim = 1), \tag{4}$$

$$avg\_out = mean(X_{channel}, \dim = 1), \tag{5}$$

where *max* and *mean* represent maximization and averaging, respectively. The obtained results are then stacked in the channel dimension, resulting in the spatial attention expression as shown in Equation (6):

$$F_{spatial} = Sigmoid(Conv2d(Concat(max\_out, avg\_out, dim = 1))), \tag{6}$$

where *Conv2d* denotes the 2D convolution operation, and *Concat* denotes the stacking operation. Finally, the spatial attention weights are applied to the channel attention-adjustment feature maps, and the final output feature map F_{out} obtained is represented as shown in Equation (7):

$$F_{out} = F_{channel} \odot F_{spatial}. \tag{7}$$

3.2. SPPA network

The SPP network is a technique to achieve feature fusion at different scales. It performs feature extraction by max-pooling operation with different convolutional kernel sizes to improve the sensory field of the network. However, there are some issues with the original SPP network. Firstly, the SPP network does not consider the importance of spatial and channel information. Certain feature channels, such as shape and position, are important in fall scenes. Secondly, fall detection requires not only recognizing a person's poses but also understanding the environmental context. The model's understanding of the context is enhanced by average pooling. Additionally, it is critical to capture features on different scales for fall detection because the fall pose and background context can be very different. Multi-scale features help the model to recognize fall poses on various scales more accurately.

Fig. 3 (a) illustrates the original SPP [12] structure. Fig. 3 (b) illustrates the SPPA structure. Suppose the input feature map is denoted as $X \in R^{C\times W\times H}$. The output representation of the input feature map obtained after the CBS module is as shown in Equation (8):

$$X' = Relu\left(\alpha\left(\frac{Conv(F,\omega) + b - \delta}{\sqrt{\varepsilon^2}}\right) + \beta\right), \tag{8}$$

where $Conv(F, w)$ denotes the convolution operation, ω is the weight of the convolution kernel, b is the bias of the convolution, α and β are the learnable parameters, and δ and ε^2 are the mean and variance, respectively. The activation function can be represented as shown in Equation (9):

$$Relu = max(0, x). \tag{9}$$

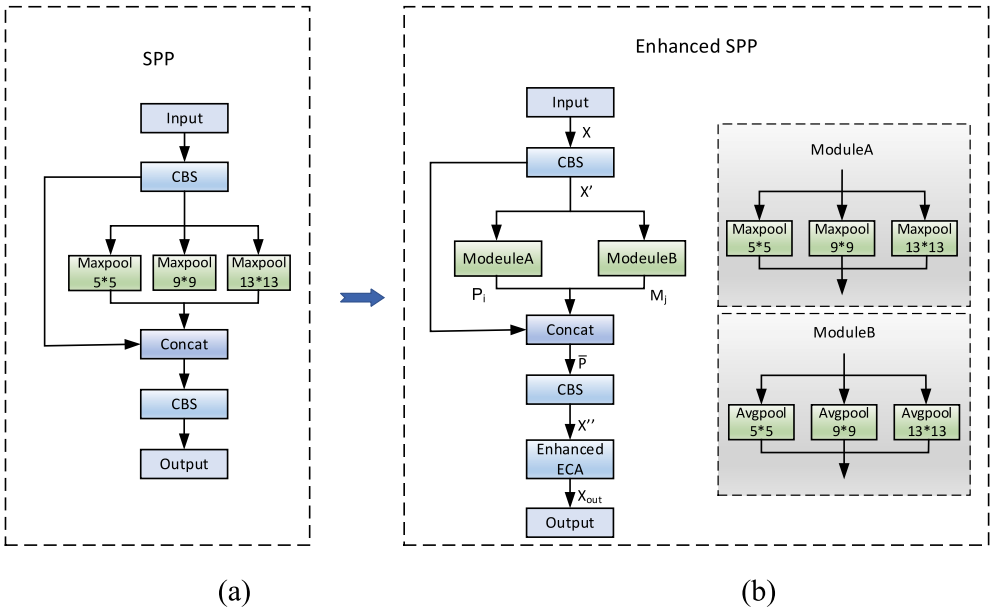


Fig. 3. Network structure. (a) SPP. (b) SPPA.

The max-pooling operation is calculated as shown in Equation (10):

$$P_i = \sum_{i=1}^{3} max_{w,h}\left(X'_{C:W:H}\right), \tag{10}$$

where $X'_{C:W:H}$ denotes the local area in width, height, and channel, and *max* denotes the max-pooling operation. The average pooling operation is calculated as shown in Equation (11):

$$M_j = \sum_{j=1}^{3} avg_{w,h}\left(X'_{C:W:H}\right), \tag{11}$$

where *avg* denotes the average pooling operation. The outputs of all pooling layers are then spliced together to form a new feature vector. The feature of the spliced vector is denoted as: $\overline{P} = [P_1, P_2, P_3, M_1, M_2, M_3]$. Then, the output feature map can be obtained as X'' according to Equation (8). Finally, the output obtained after the ECSA network can be expressed as shown in Equation (12):

$$X_{out} = F_{out}\left(X''\right). \tag{12}$$

4. Experiments and results

4.1. Dataset

The dataset has a total of 1440 images, containing both normal and fall states. The total number of fall event labels is 1360. There are 1170 images in the training set, 130 images in the validation set, and 140 images in the test set. The training and validation sets are divided into a 9:1 ratio. To improve the reliability of the experimental results, this paper employs 10-fold cross-validation [29]. Specifically, the training data and the validation data are divided into ten sub-samples in total. For each validation, nine of these sub-samples are used for training and one for validation.

4.2. Experimental setting and training results

Software environment: the operating system is Windows 10, the deep learning framework is PyTorch, and the programming language is Python. Hardware environment: CPU is Intel Core i7, GPU is NVIDIA RTX 3060, memory is 12G, and CUDA version is 11.7.

To ensure the model's training effect and stability, the parameters are set as follows: the image input size is 640*640. The optimizer used is Stochastic Gradient Descent (SGD). The initial learning rate is set at 0.01. The epoch is 400. The batch size is 16. Fig. 4 shows the loss function during the training and validation process. As can be seen from the figure, the loss value decreases as the training time increases. When epoch is 80, the loss function begins to stabilize, indicating that the model is beginning to converge.

4.3. Evaluation indicators

Assume *TP* denotes that positive samples are considered positive, *FP* denotes that negative samples are considered positive, and *FN* denotes that positive samples are considered negative. The formulas for precision and recall [22] are shown below as Equations (13)

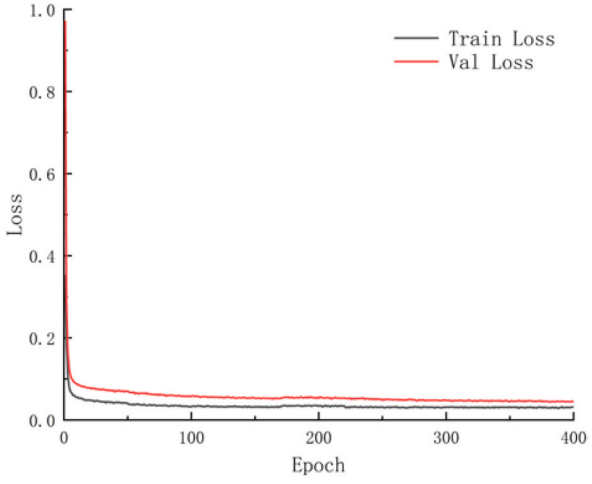


Fig. 4. Loss function curve.

and (14):

$$P = \frac{TP}{TP + FP}, \tag{13}$$

$$R = \frac{TP}{TP + FN}. \tag{14}$$

Average Precision (AP) [20] evaluates the algorithm's ability to balance precision and recall across different categories. AP is calculated as shown in Equation (15):

$$AP = \sum_{n} (R_n - R_{n-1}) P_n, \tag{15}$$

where R_n and P_n are recall and precision at threshold $n = 0.5$. Fig. 5 displays the model's P–R curves during the training process. The mean Average Precision (mAP) [20] is the average value of each type of AP. mAP is calculated as shown in Equation (16):

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k}, \tag{16}$$

where k denotes the number of categories. Because there is only one label for fall events in the dataset, $k = 1$ is used. Higher mAP values indicate higher accuracy of the model in predicting fall events.

Frames Per Second (FPS) is used to evaluate the detection speed of a model. The higher the FPS, the faster the model's processing speed, indicating that it can complete object recognition in a shorter amount of time.

To verify the significance of the performance improvements of SCPE-YOLOv5s more rigorously over YOLOv5s, this paper employs the Wilcoxon test [30]. This test is a nonparametric statistical method for comparing the central tendencies of two related samples. The Wilcoxon test does not require the data to follow a normal distribution, making it particularly suitable for analyzing small samples or data that do not satisfy the conditions for normality.

4.4. Experiments results

4.4.1. Ablation experiment

In this paper, ablation experiments are conducted to assess the effect of various modules in SCPE-YOLOv5s on model performance. Fig. 6 demonstrates the changes in mAP for the four models during the training process. By comparison, it is found that SCPE-YOLOv5s performs best at epoch 320, indicating that the model achieved the best performance in the validation set after 320 rounds of training. During training, the mAP of the other three models also fluctuated; however, their highest mAP did not exceed that of SCPE-YOLOv5s.

Table 2 presents the experimental results of four models. It reveals the optimal performance of SCPE-YOLOv5s, with a mAP of 88.29 %. Compared with YOLOv5s, the mAP of SCPE-YOLOv5s improved by 4.87 %; compared with YOLOv5s + SPPA, the mAP improved by 3.83 %; compared with YOLOv5s + ECSA, the mAP improved by 2.73 %. The SCPE-YOLOv5s′ performance is demonstrated by these data. Additionally, the introduction of either SPPA or ECSA in YOLOv5s brings about an increase in mAP. This indicates that the two networks, SPPA and ECSA, are performing well. As shown in the table, the FPS of YOLOv5s is 62.6, highlighting its excellent processing speed. After introducing either the SPPA or ECSA network alone, the model's performance remains high, although there is a slight decrease in FPS. It is worth mentioning that when both networks are applied simultaneously, the FPS of SCPE-YOLOv5s

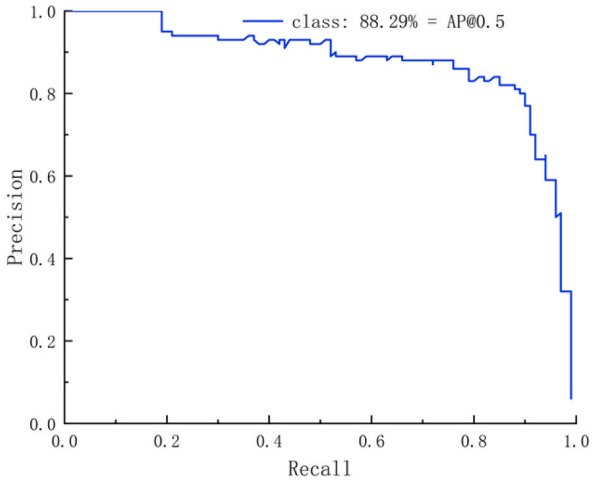


Fig. 5. P–R curve.

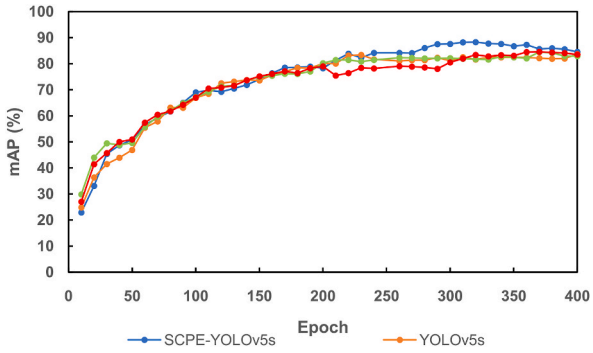


Fig. 6. Trend plots of mAP for different models during training.

Table 2
Ablation results on a dataset.

Model	SPPA	ECSA	mAP (%)	FPS (f/s)
YOLOv5s			83.42	62.6
	✓		84.46	60.8
		✓	85.56	59.6
	✓	✓	88.29	57.4

stabilizes at 57.4. This fully demonstrates the excellent performance of SCPE-YOLOv5s.

To demonstrate the performance of different models more intuitively, Fig. 7 shows the actual detection results of YOLOv5s and SCPE-YOLOv5s. From Fig. 7 (a), it has been discovered that YOLOv5s occasionally fails to completely detect the target. In contrast, as shown in Fig. 7 (b), SCPE-YOLOv5s provides more comprehensive detection results. Furthermore, the confidence level of the targets detected by YOLOv5s is generally low, which is significantly lower than the detection results of SCPE-YOLOv5s. Also, the heat maps are being generated. The heat map visually demonstrates how much attention the model pays to different regions. From the heat map, we can see that SCPE-YOLOv5s exhibits a higher confidence level in the detected regions with a more distinct red color, which is consistent with its higher confidence score. In comparison, YOLOv5s, while also correctly labeling the target, exhibits a lower intensity of color, indicating that the model is indeed less confident in its own predictions than SCPE-YOLOv5s.

4.4.2. Comparison experiment

To evaluate the performance of different models on the fall event dataset, You Only Look Once version 3 (YOLOv3) [31], You Only Look Once version 4 (YOLOv4) [32], Improved YOLOv5s [26], and SCPE-YOLOv5s are selected for comparative experiments. Among them, YOLOv3 and YOLOv4 are classical models in the field of target detection; Improved YOLOv5s further optimizes the feature extraction process by integrating asymmetric convolutional blocks and spatial attention mechanisms. The performance of these models in real applications can be more accurately evaluated by comparing their performance on the same dataset.

The experimental results of each comparison model are shown in Table 3. From the table, SCPE-YOLOv5s has the best performance on mAP, with an improvement of 6.53 % compared to YOLOv3; 6.01 % compared to YOLOv4; and 4.08 % compared to Improved YOLOv5s. These data fully demonstrate the superior performance of SCPE-YOLOv5s on the fall event detection task.

Fig. 8 illustrates the detection results of the comparative models. Fig. 8(a) shows YOLOv3 accurately detecting targets, whereas Fig. 8(b) indicates that YOLOv4 has target detection errors in some cases. Fig. 8(c) presents the improved YOLOv5s, which also accurately detects targets. It is particularly noteworthy that Fig. 8(d), depicting SCPE-YOLOv5s, detects targets with higher confidence than the other models. Furthermore, this higher confidence level is clearly visible when analyzing the heatmaps. SCPE-YOLOv5s heatmaps have a higher level of confidence in the target area.

This paper compares the performance of YOLOv5s and SCPE-YOLOv5s on a dataset using 10-fold cross-validation. Table 4 illustrates the performance data of both models on the same dataset, showing that SCPE-YOLOv5s consistently achieves higher mAP than YOLOv5s. Furthermore, the results have been statistically analyzed using the Wilcoxon test. Table 5 demonstrates the Wilcoxon test results, with a statistic of 0 and a p-value of 0.002. This result supports the conclusion that the performance improvement of SCPE-YOLOv5s is statistically significant.

5. Discussion

In daily life, early detection of fall events can significantly reduce the risk of injury. Therefore, this paper proposes a fall detection method based on SCPE-YOLOv5s, aiming to enhance the recognition abilities in complex fall scenarios. The main contribution of this paper is the development of enhanced ECA and SPP networks, integrated into YOLOv5s, which significantly improves the accuracy of fall detection.

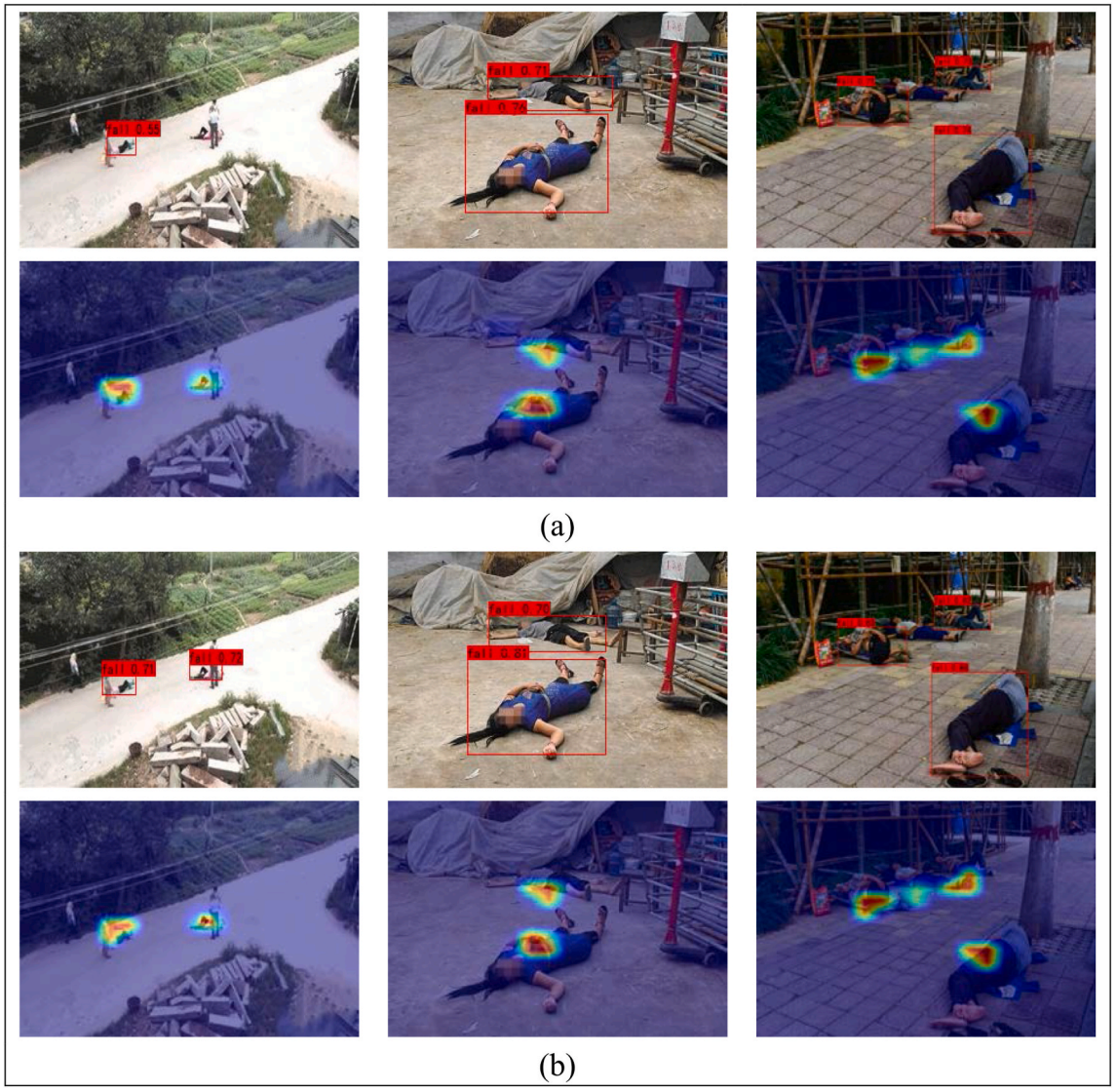


Fig. 7. Detection results. (a) YOLOv5s. (b) SCPE-YOLOv5s.

Table 3
Comparison results on a dataset.

Model	mAP (%)
YOLOv3 [31]	81.76
YOLOv4 [32]	82.28
Improved YOLOv5s [26]	84.21
SCPE-YOLOv5s	88.29

The mechanism of model performance improvement is investigated in this paper through ablation experiments. mAP is improved by 4.87 % for SCPE-YOLOv5s compared to YOLOv5s. This indicates that by improving the model's feature extraction ability, particularly in capturing spatial and background information, the performance of fall detection can be effectively improved. Compared to the traditional ECA, ECSA enhances the understanding of spatial distribution in a person's pose for fall detection by improving spatial attention in feature extraction. Compared with the traditional SPP, SPPA enhances the ability to capture background information in the fall environment by introducing an average pooling layer. Additionally, the improvements of the SPPA network incorporate max-pooling and average pooling operations, which help the model to extract multi-scale features of the fall event more easily. Because of the diversity of person poses and environments in a fall scene, the extraction of multi-scale features is critical for the

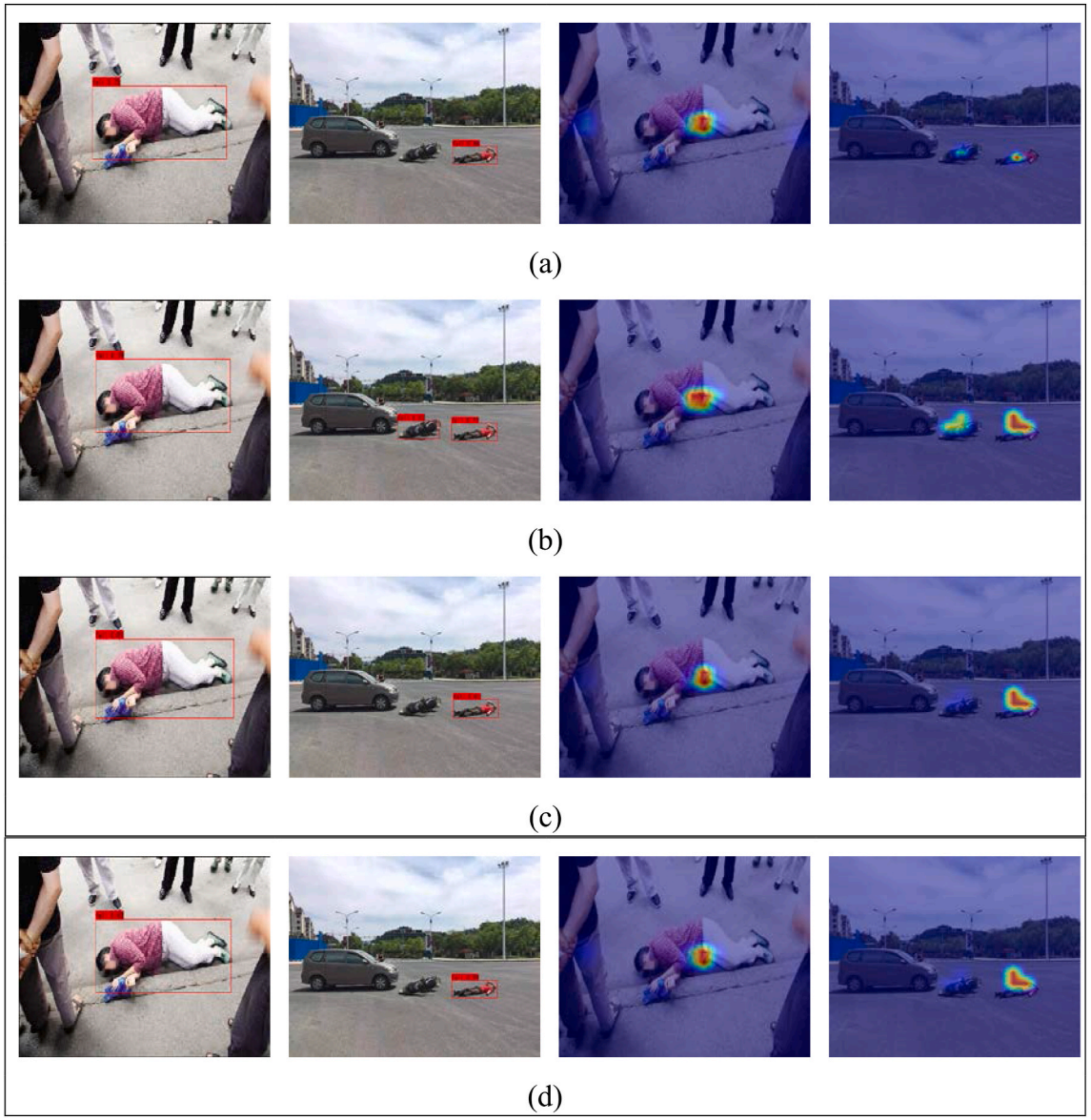


Fig. 8. Detection results. (a) YOLOv3. (b) YOLOv4. (c) Improved YOLOv5s. (d) SCPE-YOLOv5s.

Table 4
10-fold cross-validation results.

Model	mAP (%)									
	1	2	3	4	5	6	7	8	9	10
YOLOv5s	83.56	83.41	84.42	82.12	82.89	84.45	82.50	83.70	83.82	83.35
SCPE-YOLOv5s	88.56	86.74	88.30	89.27	88.57	87.89	86.26	88.85	88.90	89.52

Table 5
Wilcoxon test results on YOLOv5s and SCPE-YOLOv5s.

Indicator	Value
Wilcoxon Statistic	0
P-value	0.002

model's recognition ability. The max-pooling operation allows the model to capture the larger features of the image, whereas the average pooling operation allows the model to extract the detailed information in the image. By combining these two pooling operations, the SPPA network can extract the features of the fall scene more comprehensively, thus improving the recognition performance of the model. Furthermore, this paper combines ECSA and SPPA. Experimental results show that by strengthening the global and local feature extraction abilities, this combination can improve the model's recognition ability in fall scenes. To demonstrate the model's performance more intuitively, this paper includes actual target detection comparison charts and heat maps in the experiments. As a result of these visualization results, the superiority of SCPE-YOLOv5s in the fall detection task can be seen more clearly.

In the comparison experiments, this paper compares the proposed method with YOLOv3, YOLOv4, and improved YOLOv5s. The experimental results show that the proposed method has a significant advantage in mAP when compared to other methods. This indicates that the proposed method is more effective in the fall detection task. It is worth noting that although YOLOv3 and YOLOv4 show strong performance in the target detection task, they do not perform as well as YOLOv5s in the specific scenario of fall detection. This is primarily due to their limited ability to extract features and attention mechanisms, which are insufficient to accurately capture the nuances of fall behavior. Additionally, Improved YOLOv5s significantly improves the feature extraction performance by using asymmetric convolutional blocks and spatial attention mechanisms. These enhancements improve the model's ability to better understand the semantic information in the scene. However, in practice, the method still has limitations when it comes to dealing with the spatial distribution and background complexity that is specific to fall detection. The ability to capture the dynamic interaction between the background and the character during the fall process needs to be improved. Through comparison and analysis, we have discovered that SCPE-YOLOv5s is much better at understanding complex fall scenarios, which is attributed to the combined improvements of the ECSA and SPPA. These improvements not only enhance the spatial attention of the model, but also improve the ability to extract multi-scale features.

This paper clearly reveals the significant performance enhancement of SCPE-YOLOv5s compared to YOLOv5s through rigorous 10-fold cross-validation. This enhancement is fully reflected in the mAP statistics and is solidly supported by statistical analysis using the Wilcoxon test. Specifically, the Wilcoxon test yields a statistic of 0 with a p-value of 0.002. This result demonstrates that the advantages of the improved model are not only significant but also highly statistically reliable.

Even though SCPE-YOLOv5s has achieved good results in experiments, it still has some limitations in distinguishing between lying and falling events. The existing dataset may not fully cover all potential lying and falling postures, making it difficult for the model to distinguish between them when faced with new postures.

6. Conclusion

This paper proposes a SCPE-YOLOv5s model that aims to allow for the detection of fall events in daily life. By incorporating spatial attention paths into the ECA network, the model can provide a better understanding of the features of the distribution of a person's pose in a fall scenario. Additionally, the improved ECA network is embedded into the up-sampling process of the enhanced feature extraction network, which significantly improves the local and global feature extraction abilities of the model. Furthermore, adding average pooling layers to the SPP network not only enhances the multi-scale feature extraction ability of the model, but also optimizes the background information grasping ability. In this paper, the fall event dataset is validated. The ablation experiment results show that SCPE-YOLOv5s improves the mAP by 4.87 % over YOLOv5s. In addition, compared with other state-of-the-art algorithms, SCPE-YOLOv5s is still optimal.

Future work will further extend the dataset, investigate the model's ability to recognize different fall postures, and improve the accuracy and reliability of detection. The human detection frame and pose estimation results are used as inputs to the model to distinguish between lying events and falling events.

Data availability statement

The data that support the findings of this study are available in [AI Studio] at https://aistudio.baidu.com/datasetdetail/94809/1.

CRediT authorship contribution statement

Hao Chen: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wenye Gu:** Supervision, Software, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Qiong Zhang:** Writing – review & editing, Validation, Supervision, Resources, Investigation. **Xiujing Li:** Validation, Resources, Investigation, Formal analysis. **Xiaojing Jiang:** Visualization, Resources, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by Natural Science Research of Jiangsu Higher Education Institutions of China (Grant 23KJD520011) and Directive Projects of Nantong municipal science and technology plan (Grant MSZ2023175). The authors are thankful to all the personnel who either provided technical support or helped with data collection. We also acknowledge all the reviewers for their useful comments and suggestions.

References

[1] T. Xu, Y. Zhou, J. Zhu, New advances and challenges of fall detection systems: a survey, Appl. Sci. 8 (3) (2018) 418.
[2] R. Igual, C. Medrano, I. Plaza, Challenges, issues and trends in fall detection systems, Biomed. Eng. Online 12 (1) (2013) 66.
[3] A. Singh, S.U. Rehman, S. Yongchareon, P.H.J. Chong, Sensor technologies for fall detection systems: a review, IEEE Sensor. J. 20 (13) (2020) 6889–6919.
[4] X. Wang, J. Ellul, G. Azzopardi, Elderly fall detection systems: a literature survey, Front. Robot. AI 7 (2020) 71.
[5] M. Mubashir, L. Shao, L. Seed, A survey on fall detection: principles and approaches, Neurocomputing 100 (2013) 144–152.
[6] Y. Zhang, X. Zheng, W. Liang, S. Zhang, X. Yuan, Visual surveillance for human fall detection in healthcare IoT, IEEE MultiMedia 29 (1) (2022) 36–46.
[7] A.N. Tabata, A. Zimmer, L. dos Santos Coelho, V.C. Mariani, Analyzing CARLA's performance for 2D object detection and monocular depth estimation based on deep learning approaches, Expert Syst. Appl. 227 (2023) 120200.
[8] S.A. Tarimo, M.A. Jang, E.E. Ngasa, H.B. Shin, H. Shin, J. Woo, WBC YOLO-ViT: 2 Way-2 stage white blood cell detection and classification with a combination of YOLOv5 and vision transformer, Comput. Biol. Med. 169 (2024) 107875.
[9] G.H. Aly, M. Marey, S.A. El-Sayed, M.F. Tolba, YOLO based breast masses detection and classification in full-field digital mammograms, Comput. Methods Progr. Biomed. 200 (2021) 105823.
[10] O.G. Ajayi, J. Ashi, B. Guda, Performance evaluation of YOLO v5 model for automatic crop and weed classification on UAV images, Smart Agric. Technol. 5 (2023) 100231.
[11] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.
[12] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[13] B.J. Souza, S.F. Stefenon, G. Singh, R.Z. Freire, Hybrid-YOLO for classification of insulators defects in transmission lines based on UAV, Int. J. Electr. Power Energy Syst. 148 (2023) 108982.
[14] O. Kerdjidj, N. Ramzan, K. Ghanem, A. Amira, F. Chouireb, Fall detection and human activity classification using wearable sensors and compressed sensing, J. Ambient Intell. Hum. Comput. 11 (2020) 349–361.
[15] H. Chander, R.F. Burch, P. Talegaonkar, D. Saucier, T. Luczak, J.E. Ball, R.K. Prabhu, Wearable stretch sensors for human movement monitoring and fall detection in ergonomics, Int. J. Environ. Res. Publ. Health 17 (10) (2020) 3554.
[16] A. Alarifi, A. Alwadain, Killer heuristic optimized convolution neural network-based fall detection with wearable IoT sensor devices, Measurement 167 (2021) 108258.
[17] S. Nooruddin, M.M. Islam, F.A. Sharna, An IoT based device-type invariant fall detection system, Internet of Things 9 (2020) 100130.
[18] M.J. Al Nahian, T. Ghosh, M.H. Al Banna, M.A. Aseeri, M.N. Uddin, M.R. Ahmed, M.S. Kaiser, Towards an accelerometer-based elderly fall detection system using cross-disciplinary time series features, IEEE Access 9 (2021) 39413–39431.
[19] L.U.O. Bo, Human fall detection for smart home caring using yolo networks, Int. J. Adv. Comput. Sci. Appl. 14 (4) (2023).
[20] X. Kan, S. Zhu, Y. Zhang, C. Qian, A lightweight human fall detection network, Sensors 23 (22) (2023) 9069.
[21] X. Fan, Q. Gong, R. Fan, J. Qian, J. Zhu, Y. Xin, P. Shi, Substation personnel fall detection based on improved YOLOX, Electronics 12 (20) (2023) 4328.
[22] L. Lyu, Y. Liu, X. Xu, P. Yan, J. Zhang, EFP-YOLO: a quantitative detection algorithm for marine benthic organisms, Ocean Coast Manag. 243 (2023) 106770.
[23] S.M. Abas, A.M. Abdulazeez, D.Q. Zeebaree, A YOLO and convolutional neural network for the detection and classification of leukocytes in leukemia, Indonesian J. Electr. Eng. Computer Sci. 25 (1) (2022) 200–213.
[24] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[25] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: In Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[26] T. Chen, Z. Ding, B. Li, Elderly fall detection based on improved YOLOv5s network, IEEE Access 10 (2022) 91273–91282.
[27] D. Zhao, T. Song, J. Gao, D. Li, Y. Niu, YOLO-fall: a novel convolutional neural network model for fall detection in open spaces, IEEE Access 12 (2024) 26137–26149.
[28] Y. Wang, Z. Chi, M. Liu, G. Li, S. Ding, High-performance lightweight fall detection with an improved YOLOv5s algorithm, Machines 11 (8) (2023) 818.
[29] J.G. Moreno-Torres, J.A. Sáez, F. Herrera, Study on the impact of partition-induced dataset shift on k-Fold cross-validatio, IEEE Transact. Neural Networks Learn. Syst. 23 (8) (2012) 1304–1312.
[30] G. Divine, H.J. Norton, R. Hunt, J. Dienemann, A review of analysis and sample size calculation considerations for Wilcoxon tests, Anesth. Analg. 117 (3) (2013) 699–710.
[31] Redmon J., Farhadi A., Yolov3: an incremental improvement, arXiv preprint arXiv:1804.02767 (2018). http://arxiv.org/abs/1804.02767.
[32] A. Bochkovskiy, C.Y. Wang, H.Y.M. Liao, Yolov4: optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934 (2020). https://arxiv.org/abs/2004.10934.