AMDNet: Adaptive Fall Detection Based on Multi-scale Deformable Convolution Network

Minghua Jiang, Keyi Zhang, Yongkang Ma, Li Liu, Tao Peng, Xinrong Hu, and Feng Yu(✉)

School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China
yufeng@wtu.edu.cn

Abstract. Recent studies by the World Health Organization have shown that human falls have become the leading cause of injury and death worldwide. Therefore, human fall detection is becoming an increasingly important research topic. Deep learning models have potential for fall detection, but they face challenges such as limited utilization of global contextual information, insufficient feature extraction, and high computational requirements. These issues constrain the performance of deep learning on human fall detection in terms of low accuracy, poor generalization, and slow inference. To overcome these challenges, this study proposes an Adaptive Multi-scale Detection Network (AMDNet) based on multi-scale deformable convolutions. The main idea of this method is as follows: 1) Introducing an improved multi-scale fusion module, enhances the network's ability to learn object details and semantic features, thereby reducing the likelihood of false negatives and false positives during the detection process, especially for small objects. 2) Using the Wise-IoU v3 with two layers of attention mechanisms and a dynamic non-monotonic FM mechanism as the boundary box loss function of the AMDNet, improves the model's robustness to low-quality samples and enhances the performance of the object detection. This work also proposes a diversified fall dataset that covers as many real-world fall scenarios as possible. Experimental results show that the proposed method outperforms the current state-of-the-art methods on a self-made dataset.

Keywords: Fall detection · Fall dataset · Multi-scale deformable convolution · Loss function · Multi-scale feature fusion

1 Introduction

Falls are a leading cause of injury and death worldwide [19], especially among older adults. Falling accidents, such as escalator failures, are becoming increasingly frequent. Fall detection is a popular research topic in the field of public safety, and vision-based methods are considered the most promising. However, in practice, accurate fall recognition faces several challenges, as described below:

B. Sheng et al. (Eds.): CGI 2023, LNCS 14497, pp. 3–14, 2024.
https://doi.org/10.1007/978-3-031-50075-6_1

1) Current public fall datasets have several limitations, such as being limited to a single scene, a high degree of sample repetition, and a lack of samples from occluded objects.
2) There are various postures associated with human falls, and existing object detection algorithms often fail to extract meaningful features from these postures due to insufficient utilization of global context information, suboptimal feature extraction, and high computational requirements. These limitations result in low detection accuracy, poor generalization ability, and slow inference speeds.
3) False detection and missed detection can be common issues when attempting to accurately detect falls due to factors such as environmental conditions and occlusion.

To address these limitations, we propose AMDNet, an adaptive fall detection Network based on multi-scale deformable convolutions to improve the accuracy of fall detection. The main contributions of this work are as follows:

1) We construct a diversified falls dataset, which differs from existing publicly available datasets. The images in this dataset are collected from various real-world scenarios, including indoor and outdoor environments, different lighting conditions, varied body types, and different fall poses.
2) We propose a multi-scale deformable convolution network to achieve more accurate and robust fall detection by optimizing the network architecture and convolution operation.
3) We adopt the Wise-IoU regression loss function in the multi-scale deformable convolution network to improve the model's robustness to low-quality samples and enhance the performance of object detection.

2 Related Works

The current mainstream fall detection methods can be roughly divided into two categories: methods based on wearable devices and methods based on computer vision [1,10,18]. Wearable approaches typically use sensors attached to the body, such as accelerometers, gyroscopes, and pressure sensors, to detect falls [6,7,9,16,21]. For example, Montanini et al. [13] used pressure sensors and accelerometers embedded in shoes to detect and identify falling events. Antwi-Afari et al. [2] utilized a wearable insole pressure system to assess the fall risk of construction workers by analyzing their plantar pressure patterns and biomechanical gait stability parameters. Pandya et al. [14] achieved real-time fall detection by integrating accelerometer and gyroscope sensors. Wearable fall detection methods usually require complex devices that need to be charged frequently and are inconvenient to wear. In contrast, computer vision-based fall detection methods only require fixed equipment, such as cameras, and offer the advantages of simplicity, efficiency, and low cost. As a result, computer vision-based methods have received increasing attention in recent years. With the widespread use of IoT and cameras, computer vision-based methods have become ideal for fall detection applications.

The effectiveness of deep learning in various computer vision applications has been proven, with convolutional neural networks (CNNs) receiving much attention in recent years. Chen et al. [5] proposed a robust fall detection method using Mask-CNN and attention-guided Bi-directional LSTM in complex backgrounds, achieving accurate results. Zhang et al. [20] devised a fall detection algorithm based on temporal and spatial body posture changes. Determining the occurrence of falls by creating an evolutionary map of human behavior. Zhu et al. [23] developed an algorithm utilizing a deep vision sensor and a convolutional neural network to train on extracted three-dimensional posture data of the human body and create a fall detection model. However, the algorithm's timeliness is relatively limite is relatively limited. Cao et al. [4] proposed a fall detection algorithm that integrates motion features and deep learning, leveraging YOLOv3 for human target detection and fusing motion features with deep CNN-extracted features to precisely detect falls. Li et al. [11] proposed an optimized YOLOv5s-based fall detection method that uses MobileNetV3 as the backbone network for feature extraction and applies a lightweight attention mechanism to enhance the detection accuracy of the model. Although these fall detection methods have achieved high accuracy in controlled environments, they still face challenges in practical applications. One of the main challenges is the limited generalization ability of the model, which reduce performance when tested on data from different environments or populations. This is mainly due to limited diversity and quantity of training data, which cannot fully represent real-world scenarios. Another challenge is the slow speed of model inference, which makes it difficult to meet the real-time requirements of some applications. The main reason for this is that deep learning models usually require a lot of computation, which is time-consuming and resource-intensive. To address these challenges, we propose an adaptive fall detection method based on multi-scale deformable convolution network, which aims to enhance the model's robustness and reduce its computational complexity without sacrificing performance.

3 AMDNet Framework

This paper proposes a fall detection algorithm based on an Adaptive Multi-scale Detection Network (AMDNet) that can rapidly and accurately recognize human fall events. The algorithm architecture is inspired by the YOLO object detection network [15] and incorporates multi-scale deformable convolutions to achieve adaptive fall detection.

AMDNet is a deep learning-based algorithm designed for fall detection. The framework of AMDNet is as shown in Fig. 1. Its backbone network is based on the CSPDarknet53 architecture [3] and consists of CBL and C2f modules. The CBL module encapsulates three main functions: convolutional layer, batch normalization [8], and a LeakyReLU activation function. This lightweight design allows the CBL module to be easily embedded into deep learning models for feature extraction. The C2f module is used to enhance the network's receptive field and representation capability. It is a lightweight convolutional block that performs

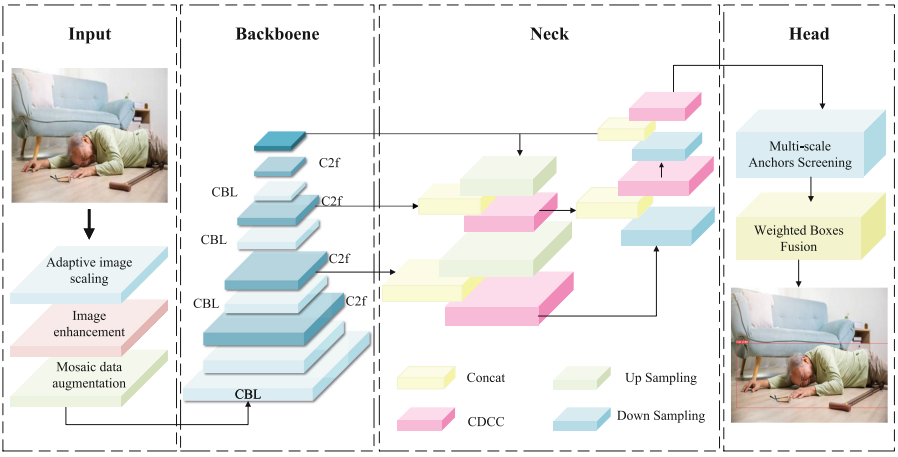

Fig. 1. AMDNet Framework.

feature extraction and integration in both the channel and contextual directions. By splitting and integrating input feature maps, the C2f module can enhance the network's representation capability and adaptability, reduce the network's computational complexity and parameter count, and improve the network's computational efficiency. The combination of CBL and C2f modules achieves higher accuracy and detection speed. By using an efficient backbone network model, computational speed can be greatly increased and training parameters can be reduced, while still ensuring detection accuracy. This results in more efficient fall detection.

The feature fusion layer of the AMDNet adopts the core idea of PANet [12], which aims to improve the network's perception and expression abilities for features of different scales and levels. Specifically, the layer first constructs a feature pyramid by obtaining feature maps of different scales from the backbone network. Then, the top-down path upsamples the low-resolution feature maps to match the size of high-resolution feature maps for feature fusion at different levels. The fused feature maps are then extracted by CDCC blocks, which consist of a series of deformable convolution blocks for feature extraction and enhancement. Meanwhile, the bottom-up path downsamples the high-resolution feature maps to reduce their resolution and extracts higher-level semantic information through CDCC blocks for prediction. The CDCC blocks effectively combine deep and shallow information, preserving image information on the original feature map layers, and possess strong receptive fields and expression abilities for detecting objects of different sizes. The feature fusion layer plays a role in feature fusion and integration in the object detection task, enabling the detector to better handle feature information of different scales and semantic levels. This improves the accuracy and robustness of object detection.

3.1 CDCC Block

To better adapt to the diverse range of human body posture changes involved in the fall detection task, we propose the CDCC block. This module mainly uses multi-scale deformable convolution [22] to improve the network's ability to focus on the target area and better adapt to changes in human body posture during falls. This enhances the network's ability to extract fall target features. Additionally, the CDCC block can effectively extract features and achieve multi-scale feature fusion, which improves the accuracy of fall detection. Overall, the CDCC block improves the network's adaptability and robustness in fall detection tasks, enabling the network to better handle the diversity of human body posture changes. The CDCC block structure is shown in Fig. 2.

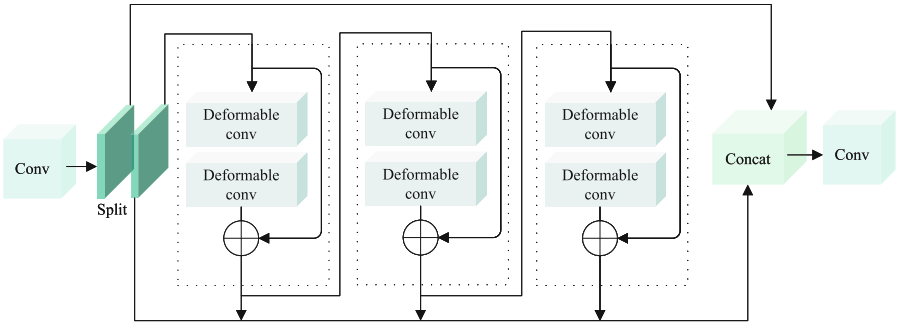

Fig. 2. The strucutre of CDCC blocks. CDCC feature enhancement module consists of a residual structure composed of multi-scale deformable convolutions.

After performing a convolution operation on the input feature map, the CDCC block splits the output feature map into two parts based on channel numbers. One part is used for subsequent feature fusion, while the other part is used for feature extraction. This avoids operating on all channels during extraction, reducing computational complexity to make the model more lightweight. Moreover, splitting the feature map into two parts enables different operations to be applied to different parts, further improving the feature representation ability. Specifically, the feature extraction part of the CDCC module uses a residual structure composed of deformable convolutions. Deformable convolution adaptively extracts local features, which can extract more abundant and accurate feature information, thereby enhancing the model's receptive field and detection ability of small targets. The residual connection is used to prevent information loss and gradient vanishing by directly transferring the previous feature information to the following layers. This is achieved by adding the input feature map to the output feature map processed by the deformable convolutional layer. The feature fusion part uses the channel concatenation operation to merge the feature information of the branch part and the extraction part, resulting in a more comprehensive feature representation. Therefore, the CDCC module helps to improve feature representation capabilities and computational efficiency, and can also improve the performance of object detection.

3.2 Loss Function

The performance of object detection depends on the design of the loss function. The bounding box loss function is an important part of the object detection loss function, and an appropriate loss function can significantly improve the performance of the object detection model. In this paper, we employ the Wise-IoU loss function [17] as the bounding box regression loss, which combines a dynamic nonmonotonic focusing mechanism and the use of "outliers" to evaluate the quality of anchor frames. This avoids excessive punishment of geometric factors such as distance and aspect ratio, and improves the performance of the model in cases where the quality of the training data annotations is low. Additionally, when the predicted box overlaps significantly with the target box, the loss function weakens the punishment of geometric factors, enabling the model to achieve better generalization ability with less intervention during training. Therefore, AMDNet uses Wise-IoU v3 with two layers of attention mechanism and dynamic non-monotonic FM mechanism, whose expressions are shown in Eqs. 1 and 2.

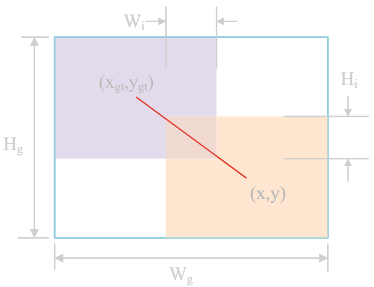

Fig. 3. The smallest enclosing box (purple) and the central points' connection (red), where the area of the union is $S_u = wh + w_{gt}h_{gt} - W_iH_i$. (Color figure online)

$$L = \left(1 - \frac{W_iH_i}{S_u}\right) \exp\left(\frac{(x_p - x_{gt})^2 + (y_p - y_{gt})^2}{\left(W_g^2 + H_g^2\right)^*}\right)\gamma \tag{1}$$

$$\gamma = \beta/\delta\alpha^{\beta-\delta} \tag{2}$$

where β is used to measure the abnormality of the predicted box, where a smaller abnormality value indicates a higher quality of the anchor box. Therefore, β is used to construct a non-monotonic focus number, which assigns smaller gradient gains to predicted boxes with larger abnormality values. In the formula, α and δ are hyperparameters, while the meanings of other parameters are shown in Fig. 3.

4 Dataset

Currently, there are several publicly available datasets for vision-based fall detection, including UR Fall Detection (URFD) Dataset, Multicam (Multiple

Cameras Fall Dataset), Fall Detection Dataset (FDD), and Le2i Fall Detection Dataset. The URFD dataset contains 30 simulated fall videos and 40 non-fall videos. The Multicam dataset contains 24 performances, each captured by 8 synchronized cameras from different angles. The FDD dataset contains 22,636 images, while the Le2i dataset contains 191 videos, covering varying numbers of falls and non-falls videos. These datasets share the following common characteristics:

1) High degree of sample repetition.
2) Unbalanced number of positive and negative samples.
3) Usually only contain indoor scenes.
4) All falling events are simulated situations.
5) Lack of small and occluded object samples.

To address the limitations of existing fall datasets, we create a Diversified Fall dataset (DFD), which contains 9,696 samples from two classes: 'fall' and 'non-fall'. Figure 4 shows some examples from DFD. The images in DFD include indoor and outdoor environments, various lighting conditions, different body types, and different fall postures to cover as many realistic fall situations as possible. This dataset is of great significance for studying human fall detection in real-world scenarios. To increase the diversity of samples in our limited dataset, we apply several data augmentation methods to some of the images. For example, we add random noise to the original images, apply blurring, and perform color transformations by changing the order of color channels.

5 Experimental Results and Discussion

The experimental environment consists of a Windows 10 operating system, a NVIDIA RTX 3050 graphics processor, and the PyTorch deep learning framework. The development environment is PyCharm and the Python 3.9.16 programming language, along with torchvision 0.14.1, numpy 1.23.0, PyTorch

Fig. 4. Sample examples from the proposed dataset. The proposed dataset includes different fall cases from different scenarios, different perspectives, and different individuals.

1.13.1, and OpenCV 4.7.0. During model training, the initial learning rate is set to 0.01, with a total of 300 epochs, and the learning rate momentum is set to 0.937.

5.1 Evaluation Metrics

The fall detection problem is treated as a binary classification problem, where each input has two possible outcomes: True (T) and False (F). In the fall detection problem, T means that a fall has occurred, while F means that no fall has occurred. To compare and evaluate the performance of experiments carried out by different researchers, various evaluation metrics have been devised. Some widely used key indicators are listed below.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$F1 - Score = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

The important indicators mentioned above can be defined by True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Among them, TP indicates that a fall event is correctly detected, FP indicates that a non-fall event is falsely detected as a fall event, and FN is the opposite, indicating that a real fall event is incorrectly detected as a non-fall event. TN means correctly identifying non-fall events as non-fall events. These metrics are commonly used definitions in binary classification problems and are crucial for evaluating the performance of fall detection systems.

5.2 Comparative Experiment

To evaluate the performance of AMDNet in human fall detection, we conducted comparison experiments with Faster-RCNN, YOLOV4, YOLOV5, YOLOX,

Table 1. The result of comparative experiment.

Model	Precision	Recall	mAP	F1-Score
Faster-RCNN	45.13%	92.59%	85.53%	59.50%
YOLOv4	88.15%	49.94%	72.05%	61%
YOLOv5	89.20%	91.90%	93.60%	90.52%
YOLOX	93.28%	86.59%	90.33%	89.50%
YOLOv7	90.70%	94.20%	94.90%	92.41%
YOLOv8	92.60%	94.50%	96.80%	93.54%
AMDNet	96.50%	95.90%	98.20%	96.19%

YOLOV7, and YOLOV8. To ensure a fair comparison, all models are trained on the same platform. Currently, YOLOV8 is one of the best-performing versions of the YOLO series object detection algorithms [15]. Table 2 shows the model experimental results of six classic networks and AMDNet on our self-collected human fall dataset DFD. The experimental results demonstrate that AMDNet achieves excellent detection performance and high accuracy on the self-made DFD dataset. Compared with traditional human detection algorithms and other object detection algorithms, AMDNet recognizes fall events more accurately with better robustness and generalization (Table 1).

5.3 Ablation Experiment

To comprehensively evaluate the proposed network framework and understand the performance of each proposed module, we conducted ablation experiments. These experiments gradually dissected each part to determine its impact on the overall performance of the model, in order to fully evaluate the contribution of each module to the speed and accuracy of the network.

From Table 2, it can be seen that the introduction of the CDCC block in the network increases the evaluation metrics Precision, Recall, and F1-Score by 2.6%, 1.2%, and 1.9%, respectively. This indicates that the CDCC block can enhance the ability to extract human fall features, thereby making fall detection more accurate. The effectiveness of the Wise-IoU loss function has been demonstrated in the fall detection task, as shown in the first and third rows of the table.

From the fourth and fifth rows, it can be seen that when different improvement methods are combined, the performance of the model is further slightly improved on the basis of certain improvements, rather than simply adding up. In summary, the combination of different methods can enhance the detection performance of the model, indicating that the proposed method is feasible and necessary.

Table 2. The result of ablation experiment.

AMDNet	CDCC	WIou	Precision	Recall	F1-Score
✓	–	–	92.60%	94.50%	93.54%
✓	✓	–	95.20%	95.70%	95.44%
✓	–	✓	95.30%	96.10%	95.69%
✓	✓	✓	96.50%	95.90%	96.19%

5.4 Detection Results

After testing in real falling scenarios, the algorithm's detection results are shown in Fig. 5. The figure presents several common types of falls, where group (a) indicates that the algorithm can accurately detect human falls in good outdoor

ambient lighting conditions; group (b) shows scenes with relatively dark lighting that are prone to falling, but the algorithm can still accurately detect human falls; and group (c) shows some relatively complex and occluded human falls, but the algorithm can still accurately detect them. Therefore, in most cases, the algorithm is less affected by environmental disturbances and can accurately detect human falls, providing protection for personal safety.

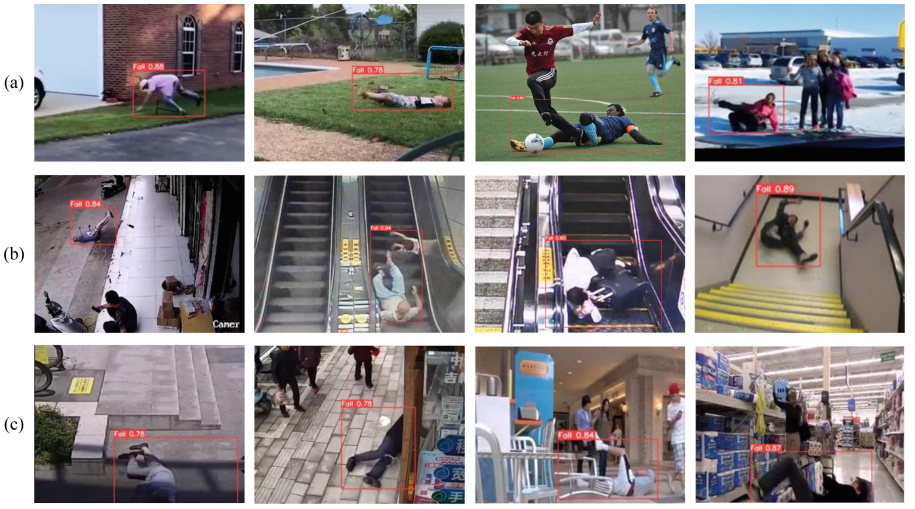

Fig. 5. Group (a) demonstrates the effectiveness of fall detection in a conventional scene, while group (b) shows the algorithm's ability to detect falls in challenging environments such as elevators and stairs. Group (c) presents the algorithm's performance in complex and occluded scenes, where it is still able to detect falls accurately. These results demonstrate the robustness of the AMDNet in various real-world scenarios and its potential for improving personal safety.

6 Conclusion and Outlook

The proposed deformable convolution-based fall detection network and diverse fall dataset have effectively improved the accuracy of human fall detection and have practical value for real-time detection and early warning of human falls. Our proposed method achieved an F1-Score of 96.19%, outperforming six classical object detection networks. However, there is still room for improvement, especially in complex scenes and objects with more occlusions. To address this issue, in future work, we plan to expand the dataset by adding more postures of falling in complex scenes and continue exploring ways to reduce the number of network model parameters while improving the detection rate. Additionally, we will investigate the use of other advanced techniques to further enhance the detection performance. We believe that these improvements will help make the proposed fall detection system more robust and effective in real-world scenarios.

Acknowledgment. This work was supported by national natural science foundation of China (No. 62202346), Hubei key research and development program (No. 2021BAA042), open project of engineering research center of Hubei province for clothing information (No. 2022HBCI01), Wuhan applied basic frontier research project (No. 2022013988065212), MIIT's AI Industry Innovation Task unveils flagship projects (Key technologies, equipment, and systems for flexible customized and intelligent manufacturing in the clothing industry), and Hubei science and technology project of safe production special fund (Scene control platform based on proprioception information computing of artificial intelligence).

References

1. Ali, S.G., et al.: Experimental protocol designed to employ Nd: YAG laser surgery for anterior chamber glaucoma detection via UBM. IET Image Process. **16**(8), 2171–2179 (2022)
2. Antwi-Afari, M.F., Li, H.: Fall risk assessment of construction workers based on biomechanical gait stability parameters using wearable insole pressure system. Adv. Eng. Inform. **38**, 683–694 (2018)
3. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
4. Cao, J., Lyu, J., Wu, X., Zhang, X., Yang, H.: Fall detection algorithm integrating motion features and deep learning. J. Comput. Appl. **41**(2), 583 (2021)
5. Chen, Y., Li, W., Wang, L., Hu, J., Ye, M.: Vision-based fall event detection in complex background using attention guided bi-directional LSTM. IEEE Access **8**, 161337–161348 (2020)
6. Fanez, M., Villar, J.R., de la Cal, E., Gonzalez, V.M., Sedano, J., Khojasteh, S.B.: Mixing user-centered and generalized models for fall detection. Neurocomputing **452**, 473–486 (2021)
7. Hussain, F., Hussain, F., Ehatisham-ul Haq, M., Azam, M.A.: Activity-aware fall detection and recognition based on wearable sensors. IEEE Sens. J. **19**(12), 4528–4536 (2019)
8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR (2015)
9. Lee, J.S., Tseng, H.H.: Development of an enhanced threshold-based fall detection system using smartphones with built-in accelerometers. IEEE Sens. J. **19**(18), 8293–8302 (2019)
10. Li, J., et al.: Automatic detection and classification system of domestic waste via multimodel cascaded convolutional neural network. IEEE Trans. Industr. Inf. **18**(1), 163–173 (2021)
11. Li-zhan, W., Xia-li, W., Qian, Z., Wei-hao, W., Chao, L.: An object detection method of falling person based on optimized yolov5s. J. Graph. **43**(5), 791 (2022)
12. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
13. Montanini, L., Del Campo, A., Perla, D., Spinsante, S., Gambi, E.: A footwear-based methodology for fall detection. IEEE Sens. J. **18**(3), 1233–1242 (2017)
14. Pandya, B., Pourabdollah, A., Lotfi, A.: Fuzzy logic web services for real-time fall detection using wearable accelerometer and gyroscope sensors. In: Proceedings

of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 1–7 (2020)
15. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
16. de Sousa, F.A.S.F., Escriba, C., Bravo, E.G.A., Brossa, V., Fourniols, J.Y., Rossi, C.: Wearable pre-impact fall detection system based on 3D accelerometer and subject's height. IEEE Sens. J. **22**(2), 1738–1745 (2021)
17. Tong, Z., Chen, Y., Xu, Z., Yu, R.: Wise-IoU: bounding box regression loss with dynamic focusing mechanism. arXiv preprint arXiv:2301.10051 (2023)
18. Wei, H., Zhang, Q., Qin, Y., Li, X., Qian, Y.: YOLOF-F: you only look one-level feature fusion for traffic sign detection. Vis. Comput. 1–14 (2023)
19. World Health Organization: Falls (2022). https://www.who.int/news-room/fact-sheets/detail/falls
20. Zhang, J., Wu, C., Wang, Y.: Human fall detection based on body posture spatio-temporal evolution. Sensors **20**(3), 946 (2020)
21. Zhao, S., Li, W., Niu, W., Gravina, R., Fortino, G.: Recognition of human fall events based on single tri-axial gyroscope. In: 2018 IEEE 15th International conference on networking, sensing and control (ICNSC), pp. 1–6. IEEE (2018)
22. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets V2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9308–9316 (2019)
23. Zhu, Y., Zhang, Y., Li, S.: Fall detection algorithm based on deep vision sensor and convolutional neural network. Opt. Tech. **47**(1), 56–61 (2021)