Improved Pedestrian Fall Detection Model Based on YOLOv5

Yuhua Feng[1] , Yi Wei[1],Kejiang Li[2], Yuandan Feng[3] , Zhiqiang Gan[4]
1. College of Automation, Wuhan University of Technology, Wuhan, China
2. Taiyuan University of Technology College of Modern Science and Technology
3. College of Finance, Hubei College of Economics, Wuhan, China
4. Lotus Holding Group, Wuhan, China
2664399884@qq.com,546247830@qq.com, 404957876 @qq.com,2878179500@qq.com,853824938@qq.com
Corresponding Author: Kejiang Li Email: 404957876@qq.com

2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) | 978-1-6654-5864-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IAEAC54830.2022.9930104

***Abstract*—Falls are an important factor in the death and injury of workers in complex operating environments. In view of the problems of missed detection and wrong detection of the original YOLOv5 network, this paper proposes a pedestrian fall detection model based on YOLOv5. The self-built pedestrian data set is used for fall detection research. In order to weaken the interference of complex background on network feature extraction, an improved SENet attention mechanism is proposed, which helps the network to pay more attention to the fall posture. In addition, in order to reduce the missed detection rate, Soft-NMS is introduced to replace the original NMS of YOLOv5. The results show that the mAP of the fall pedestrian detection training set of the improved model is increased from 97.62% to 98.33%, which proves that the improved model can better meet the requirements of pedestrian fall detection than the unimproved YOLOv5.**

Keywords—Fall Detection; YOLOv5; Improved SENet; Soft-NMS

I. INTRODUCTION

Real-time detection of fall behavior in elderly people and workers in complex operating environments is a major issue of life safety and has important research significance. With the development of computer vision, deep learning algorithms have become one of the mainstream techniques for pedestrian pose detection. 2012, Krizhevsky and Hinton introduced AlexNet[1], which solved the problems of gradient disappearance and overfitting in network training. Since then, a boom of deep learning algorithms has been set off. Throughout these years, the development of deep learning target detection algorithms has been roughly divided into two major schools, one is Two-Stage algorithm: first generating candidate regions and then performing CNN classification, and the iconic models are mainly R-CNN (Region CNN)[2], Fast R-CNN (Fast Region-Based CNN)[3], and Faster R-CNN (Faster Region-Based CNN)[4], etc. Another class is One-Stage algorithm: the algorithm is applied directly to the input image and outputs the category and the corresponding localization, whose representative models are SSD (Single Shot MultiBox Detector)[5], YOLO (You Only Look Once) [6]series, etc. Although the detection accuracy of CNN series algorithm is getting higher and higher, but there is always a bottleneck of slow speed, for fall detection, CNN series algorithm is lacking in the detection of real-time[7], and the detection effect and accuracy of SSD for small targets does not surpass YOLOv5[8], so this paper selects YOLOv5 of One-Stage algorithm for the study of fall detection.

In recent years, YOLO series related algorithms have been improved, and Jie Hu et al [9]proposed the SENet (Squeeze-and-Excitation) attention module, which enables the network to give different "attention" to different channels, solving the problem of convolutional pooling process in the feature map Chen Yixiao[10] et al. used Res2Block to reconstruct the backbone network of YOLOv5 to improve the fine-grained feature fusion capability of the network; ZHOU LONG et al[11]. proposed a lightweight convolutional neural network LiraNet, which achieves good detection accuracy with less memory and computational cost.

The above research results make full use of the target depth feature to detect falling pedestrians, but the yolov5-based fall posture detection algorithm still has several problems that need to be solved urgently: when encountering structures similar to the posture of falling pedestrians, it is easy to be confused and cause errors. When the target overlaps in a large area, the phenomenon of missed detection will occur. In this regard, this paper proposes a series of improved methods based on the yolov5s network: an improved SENet attention mechanism[12] is proposed to enable the network to capture direction-aware and position-aware information; Soft-NMS is used to replace the NMS of the original network algorithm to reduce the missed detection rate. Experiments show that the improved model has a good detection effect, and can effectively improve the false detection phenomenon caused by the overlapping of falling targets and the inconspicuous features.

II. INTRODUCTION TO THE YOLOV5 NETWORK MODEL

As shown in Figure .1, the YOLOv5 model mainly consists of four modules: Input, Backbone, Neck, and Prediction. First, the image is input at the Input side, and

through algorithms such as resize, adaptive anchor frame algorithm, and Mosaic data enhancement, the image size is unified, the most suitable anchor frame size is found, and the image background is enriched to improve the localization accuracy of the fallen pedestrian target; then through the backbone feature extraction network (Backbone), it is aggregated at different image fine grained [13] to obtain three scale image features; through the feature fusion network (Neck), the structure of feature pyramid networks (FPN) combined with path aggregation network (PAN) is used to fuse the semantic features and location features of the network to enrich the network extracted Finally, the network has three detection layers for the detection of targets of different sizes, and the Non-Maximum Suppression (NMS) algorithm is used to remove the redundant prediction frames, generate the best detection frame and make the judgment of the target class.

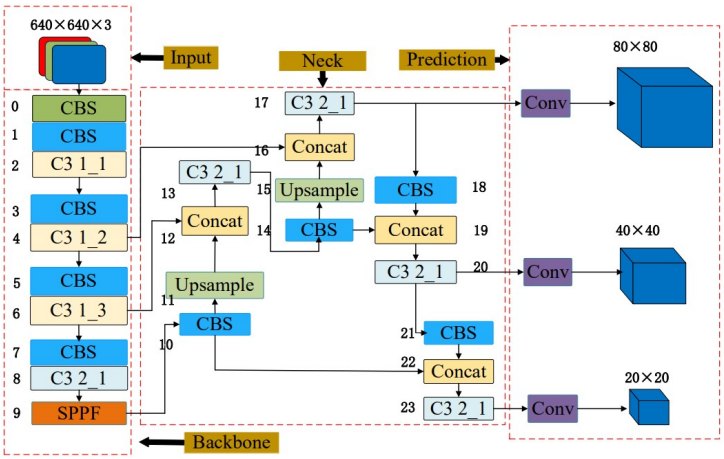


Fig. 1. YOLOV5 network model structure

III. IMPROVED YOLOV5 ALGORITHM

A. *Improving SENet*

2019 SENet (Squeeze-and-Excitation) attention mechanism emerged to enable the network to focus on important features and suppress unnecessary features.The SE block can be decomposed into two steps: Squeeze and Excitation , for global information embedding and adaptive Re-weight of channel relationships, respectively.In the input of condition, the Squeeze step corresponding to the channel can be expressed as:

$$Z_C(W) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{1}$$

The input comes from a convolutional layer with a fixed kernel size, which can be seen as a collection of local descriptors. the Sqeeze operation makes it possible for the model to collect global information. the purpose of Excitation is to completely capture the dependencies between channels, by which the network can learn the importance of each channel, which can be formulated as follows.

$$\begin{cases} \tilde{X} = X \bullet \sigma\left(\tilde{Z}\right) \\ \tilde{Z} = T_2(\mathrm{ReLU}(T(Z))) \end{cases} \tag{2}$$

SE Block only considers the importance of each channel by modeling channel relationships to remeasure the importance of each channel, ignoring the location information, which is crucial for generating spatially selective attention maps. Therefore, in this paper, the SENet is improved so that the network considers the relationship between channels while focusing on the location information in the feature space, and the schematic diagram of the improved SENet is shown in Figure.2.

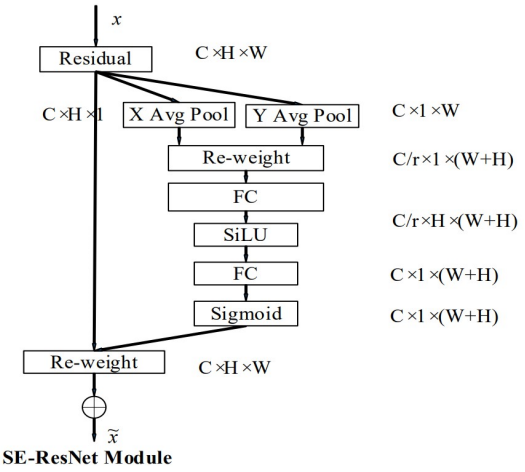


Fig. 2. Improved SE module schematic

The channel attention is decomposed into two one-dimensional features along two directions, capturing the remote dependencies along one of the spatial directions and preserving the position information along the other spatial direction. Given an input X, each channel is first encoded along the horizontal and vertical coordinates using a pooling kernal of dimension (H, 1) or (1, W), respectively. Thus, the output of the cth channel with

height H and width W can be expressed as follows, respectively.

$$Z_c^H(H) = \frac{1}{w} \sum_{0 \le i \le W} x_c(H, i) \tag{3}$$

$$Z_c^W(W) = \frac{1}{H} \sum_{0 \le i \le H} x_c(W, i) \tag{4}$$

The above two transformations aggregate features along two spatial directions respectively to obtain a pair of direction-aware feature maps. This is different from the original algorithm that yields a suitable single feature vector, and these two transformations not only allow the attention module to capture long-term dependencies along one direction, but also preserve precise location information along the other spatial direction, which helps the network to locate the fall features more accurately.

B. Improvements to NMS

The IoU-NMS used by the original Yolov5 network, IoU is the only factor considered in the algorithm, the prediction frames of the same category are sorted by confidence, the frame with the highest score is set as the reference frame, and the remaining prediction frames are traversed. If the IoU (Intersection over Union) value of the prediction box is greater than the set threshold, it will be removed, otherwise it will be retained. Repeat this cycle until all prediction frames are processed, and the optimal prediction frame obtained is the detection result. In the actual detection scene, when the distance between two objects of the same category is very close, the IoU value is large, and the algorithm may filter out one of the boxes when dealing with this situation, resulting in missed detection.

To solve this problem, this paper introduces the Soft-NMS algorithm into the network. After the reference frame is selected, other prediction frames are traversed, and the prediction frame whose IoU value is higher than the threshold is not deleted, but its confidence is reduced by the decay function, and the detection frame with the highest confidence is retained, and the detection frame with the second highest is used as the benchmark. Press the cycle to perform the second and third attenuation. Finally, the selected frame is comprehensively screened to obtain the optimal prediction frame. The Soft-NMS formula is as follows:

$$s_i \begin{cases} s_i, & IoU - R_{DIoU}(\mu, B_i) < \varepsilon \\ s_i(1 - IoU(\mu, B_i)), & IoU - R_{DIoU}(\mu, B_i) \ge \varepsilon \end{cases} \tag{5}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Construction of the experimental data set

Experimentally, 8000 daily fall pictures were collected, and part of the data set is shown in Fig. 8. In order to improve the detection ability of the target detection model, we try to ensure the diversity of the collected pictures of fall.

To improve the detection capability of the target detection model, the diversity of fall poses in the collected images is ensured as much as possible. In addition, the image set also includes small targets, target stacking, and occlusion.

The collected fall samples are screened and sorted to ensure the high quality of the dataset. In this study, the training model adopts .xml dataset format, and the open source tool LabelImg is used in the Python environment to label the area where the target object is located in each image, and the label set for the fallen pedestrian in this experiment is down. the input image size of the model is 640×640, the number of channels is 3, and the images in the dataset are original images without any pre-processing such as clarification.

B. Experimental environment and protocol design

The experimental environment of this paper is based on Windows 10 64-bit system, 16GB RAM, GPU version is NVIDIA GeForce RTX 3060 6GB Laptop, GPU acceleration library is Cuda11.3, Cudnn10.0, and the software used includes Anaconda, Pycharm, etc., to build deep learning Pytorch framework to implement the training of YOLOv5 target detection model. The input image resolution is 640×640, and the Mosaic data enhancement method is used in the training in order to enhance the model's anti-interference ability. The batch size is 16, the number of training rounds is 100 rounds, the momentum coefficient is 0.937, the weight decay coefficient is 0.0005, and the initial learning rate is 0.01.

C. Evaluation indicators

To validate the performance of YOLOv5s improved algorithm, this paper uses generic target detection evaluation metrics, Precision (P), Recall (R), and mean Average Precision (mAP) to evaluate the model. Two metrics, P and R, are usually used to measure the goodness of the model, and mAP can measure the performance of the whole model .

$$P = \frac{T_P}{T_P + F_N} \times 100\% \tag{6}$$

$$R = \frac{T_P}{T_P + F_P} \times 100\% \tag{7}$$

$$mAP = \frac{1}{c} \sum_{j=1}^{c} AP_j \tag{8}$$

D. Training results

To more intuitively reflect the performance of the improved algorithm, the improved algorithm was compared with Faster R-CNN, SSD, YOLOv4, YOLOv5s and other target detection mAP, P(100%), R(100%), and Params(M) are used to evaluate and compare each

mainstream detection algorithm, and the results of the comparison experiments are shown in Table I.

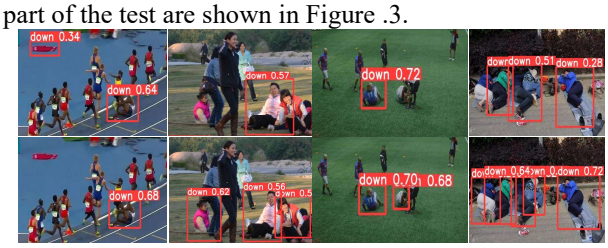
TABLE I. TABLE I. VALIDATION SET RESULTS FOR FIVE NETWORK MODELS

models	Improved Algorithm	YOLOv5	YOLOv4	SSD	Faster R-CNN
mAP(100%)	98.33	97.62	95.63	94.23	92.56
P(100%)	97.65	95.42	93.23	89.64	87.65
R(100%)	98.12	96.27	94.11	90.41	88.51
Params(M)	7.06	7.01	42.22	4.96	6.23

Comparing the training results of the five network models in Table 1, it can be seen that by improving the network of YOLOv5, the mAP of the improved algorithm reaches 98.33%, which is 0.71 percentage points higher compared with the original YOLOv5 network,and P, R, mAP indicators are better than the comparison network.Without significantly increasing the model complexity, the improved algorithm has significantly increased the detection accuracy and has certain advantages over other mainstream algorithms.

E. Test results

In order to more clearly demonstrate the effect of the improvement of the network, the fallen pedestrian images were randomly selected from the test set for testing, and the results of the YOLOv5 and the improved algorithm part of the test are shown in Figure .3.



(a) (b) (c) (d)

Fig.3. Comparison of YOLOv5 and the improved algorithm result plots

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, we propose an improved pedestrian fall detection model based on YOLOv5 , design the improve SENet attention mechanism, which helps the network to focus on the fall pose more accurately. In order to reduce the missed detection rate, Soft-NMS is introduced to replace the original NMS of YOLOv5. Experiments prove that the improved algorithm has better detection accuracy and detection effect compared with the original YOLOv5 network, and is suitable for fallen pedestrian detection projects.

This work was supported by the National Natural Science Foundation of China under Grant No.51177114 and the Hubei Provincial Technology Innovation Major Project under Grant No.2019AAA016.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.61272518 and Research of the Information Access Technology for Complex Information System under Grant No.2013RC0208.

REFERENCES

[1] M.G.Guo,H.Gong,"Research on the improvement and optimization of AlexNet,"Computer Engineering and Applications,"vol. 56,issue 20, pp. 124-131,2020.

[2] ROBINSON JIMENEZ MORENO, LUIS A, RODRIGUEZ UMANA, JAVIER E. MARTINEZ BAQUERO,"Use of R-CNN for Driving Assistance System," International Journal of Applied Engineering Research,vol. 13,issue16 , pp. 12560-12569,2018.

[3] NEPH, R., HUANG, Y., YANG, Y.,Deep Learning MC: Fast CNN-Based Prediction of Monte Carlo Dose for MR-Guided Treatment Planning," Medical Physics,vol. 46,issue6,pp. 371, 2019.

[4] ALZRAIEE, HANI, LEAL RUIZ, ANDREA, SPROTTE, ROBERT,"Detecting of Pavement Marking Defects Using Faster R-CNN," Journal of Performance of Constructed Facilities,vol. 35,issue4 ,2021

[5] YUNDONG LI, HAN DONG, HONGGUANG LI, et al. "Multiblock SSD based on small object detection for UAV railway scene surveillance,"vol. 33,issue6,pp. 1747-1755,2020.

[6] ADIBHATLA, VENKAT ANIL, CHIH, HUAN-CHUANG, "Defect Detection in Printed Circuit Boards Using You-Only Look-Once Convolutional Neural Networks,"Electronics,vol. 9,issue9,pp. 1547,2020.

[7] Zhong Zhiqing, Chen Xindu, WU Lei, "A lightweight YOLO network for real-time estimation of 6D attitude," Combined Machine Tool and Automated Machining Technology, vol. 33,issue1,pp. 24-28,2022.

[8] LEE, JOO HWAN, ZHANG, HUI, LAGRANGE, VERONICA, " SmartSSD: FPGA Accelerated Near-Storage Data Analytics on SSD," IEEE computer architecture letters,vol. 19,issue2,pp.110-113,2020.

[9] J.Hu , L.Shen, G.Sun , Squeeze-and-Excitation Networks,IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[10] Y.X.Chen,W.L.Lin,X.Yuan."CA-YOLOv5for congested pedestrian detection," Compute Engineering and Applications,pp. 1-10,2022.

[11] L.Zhou, S.Y.Wei, Z.N.Cui, "Lira-YOLO: a lightweight model for ship detection in radar images," Systems Engineering and Electronics Technology, vol. 31,issue5,pp. 950-956, 2020.

[12] SUPASIT,KAJKAMHAENG,CHANTANA,CHANTRAPORNC H AI."SE-SqueezeNet: SqueezeNet extension with squeeze-and excitation block," International Journal of Computational Science and Engineering,vol. 24,issue2,pp. 185-199,2021.

[13] Intelligence and neuroscience,2021. [18] Y.M.Fang, Y.Zhong, J.B.Yan, "Automatic image cropping with deep attention based on aggregating fine-grained features ," Chinese Journal of Image and Graphics,vol. 27,issue2,pp. 586-601, 2022.