Pedestrian Detection Based on Modified YOLOv5

Ruopeng Pei[1]

Computer and Math Teaching Department, Shenyang Normal University
Shenyang 110034, China

Received Feb. 6, 2025; Revised and Accepted Feb. 27, 2025

Abstract. In the pedestrian detection scenario, the detection algorithm usually misses obscured and distant fuzzy pedestrians, and at the same time cannot take into account the detection accuracy and speed. In this paper, we propose a modified YOLOv5 model for pedestrian detection. Firstly, the backbone network uses the SPD-GCONV module constructed by the combination of SPD (Space-to-Depth) module and Ghost convolution for down-sampling to reduce the loss of fine-grained feature information. Secondly, the multi-scale detection ability of the model is enhanced by adding a small size detection layer. Then, the original CIoU loss function is replaced by α-EloU loss function to improve the accuracy of pedestrian target location. According to the experiments on WiderPerson data set, the average detection accuracy is improved by 2% compared with other pedestrian detection algorithms on the premise of ensuring the detection speed. Experimental results show that the improved algorithm can significantly improve the detection performance.

Keywords: Pedestrian detection, Space-to-Depth module, Ghost convolution, α-EloU.

1. Introduction

Pedestrian detection is one of the research hotspots in the field of computer vision [1], which refers to the computer processing images or videos to automatically filter out pedestrian targets. The traditional pedestrian detection algorithm usually includes three steps: image preprocessing, manual feature extraction and classification. The purpose of image preprocessing is to extract more effective features [2,3]. The manual extraction process is usually detected by a sliding window method, which uses a rectangular box to scan the image in full from the top left to the bottom right corner and obtain one or more specific features (e.g., edge features, image blocks, wavelet coefficients, etc.) from the input image. The traditional pedestrian detection has the advantages of fast detection speed, but the single feature information leads to low recognition accuracy [4-6].

With the development of deep learning, various excellent detection algorithms have emerged, which effectively solve the problem of low recognition rate of traditional pedestrian detection [7]. Mainstream target detection algorithms are divided into two categories: two-stage target detection algorithms represented by RCNN [8], Fast RCNN [9], Faster RCNN [10], and single-stage target detection algorithms represented by YOLO [11] and SSD [12]. The former first obtains the candidate region with the target, and then carries on the regression prediction to the size and position of the obtained candidate frame. The latter generates several candidate frames and performs regression predictions. Pedestrian detection tasks present challenges, such as complex weather conditions and severe blocks. In addition, existing object detection networks have problems such as large model parameters and slow inference speed. In order to solve these problems, many researchers have proposed many solutions. Kim et al. [13] took YOLOv4-tiny as the baseline, combined Ghost module and extended convolution. This greatly reduced the capacity of the model. However, it limited the model's ability to learn advanced features and was less friendly to smaller objects. The pedestrian detection algorithm based on YOLOv4 proposed by Fang et al. [14] used ShuffleNet instead of YOLOv4, and used depth-separable convolution to reduce the model size. As the network became more lightweight, its feature extraction capability decreased. In complex detection scenarios, this could lead to a significant drop in recall rates. Liu [15] could improve the detection performance of the network in the case of obstructions by using YOLOv4-tiny as the baseline, combining multi-spectral methods and reducing delay. In the YOLOv5s-G2 network proposed by Guang et al. [16], α-CIoU loss function was used to improve occlusion and small target recognition in pedestrian detection tasks. Wang et al. [17] proposed a lightweight detection model based on YOLOv5, which combined the MD-SILBP operator and five-frame difference method to enhance the ability of contour feature extraction, and used the non-maximum suppression of Distance-IoU to reduce the missing rate in detection. Single-stage target detection is fast, but the accuracy is low, two-stage target detection is more accurate but slower, the above algorithm can not take into account both speed and accuracy, it is difficult to meet the detection needs.

In this paper, an improved pedestrian detection algorithm based on YOLOv5s network model is proposed under the framework of deep learning. First, in the backbone network, SPD-GConv is used to replace the conventional

$Conv$ to increase the capacity of the model and reduce the loss of fine-grained features during down-sampling. Secondly, a small scale detection layer is added to improve the feature extraction ability of long-distance small target pedestrians and improve the performance of the network model. Finally, α-EIoU loss function is used as the loss function of pedestrian detection to improve the accuracy of boundary box regression.

2. YOLOv5 Network

The YOLOv5 network structure consists of four parts: Input, feature extraction (Backbone), feature fusion (Neck) and output (Head). The input side is the part of image preprocessing, including Mosaic data enhancement, adaptive anchor frame calculation and flipping up and down pictures. Backbone network is the part of feature extraction, which mainly includes CBS module, C3 module and SPPF module in a three-layer structure. The CBS module is composed of conventional convolution, and the activation function is SiLU function. The C3 module uses a residual structure that allows the network to pass gradients by skipping connections, ensuring model capacity while improving detection efficiency [18-20]. The SPPF module is an improvement of the Spatial Pyramid Pooling (SPP) module, which adopts the same size of small size pool kernel stacking connection mode to further improve the model running speed. Neck network is a part of feature fusion. It adopts the structure of FPN (feature pyramid networks)+PAN (path aggregation networks). FPN structures transfer semantic information from deep feature maps to shallow feature maps from top to bottom, and PAN structures transfer location information from shallow feature maps to deep feature maps. The combination of the two can achieve multi-scale feature fusion to obtain more semantic and location-rich feature representation. The output end contains three detection layers of different sizes to classify and predict the fused features. YOLOv5 uses CIoU_Loss as a border loss function.

3. Modified YOLOv5 Network

In order to improve the accuracy of pedestrian detection in long-distance and dense scenes, the YOLOv5 algorithm is used as the baseline model in this paper. In the backbone network, SPD-GConv is used to replace common convolution for downsampling, which reduces the loss of fine-grained feature information in downsampling and ensures the efficiency and performance of the model. The addition of a small scale detection layer enables the model to better capture and locate detailed information about distant pedestrians. The CIoU loss function is replaced by α-EloU to improve the prediction accuracy of the model. The improved YOLOv5 structure is shown in Figure 1.

3.1. Feature Extraction Network Improvement

In most target detection network models, convolution with step size of 2 is usually used for down-sampling, which will lead to the loss of fine-grained information of feature map for detecting pedestrian targets at long distances in images. SPD module is composed of conversion from space to depth layer and non-step convolution. It adopts the operation of space exchange depth, uses slice method to down-sample the feature map, and rearranges the channel dimensions. Then, the channel dimension information is fused by non-step convolution to retain more fine-grained information and improve the expressiveness and robustness of features.

Ghost convolution [21] is the use of convolution generated by the feature map many feature maps have a high degree of similarity, using less convolution to generate a few feature maps. Then the generated feature maps use cheap operation to obtain some similar feature maps, and finally the feature maps generated in the first two steps are spliced into the feature maps to be output.

Assuming that the input and output feature sizes are $W \times H \times C$ and the convolution kernel sizes are $K \times K$, the number of parameters required for a common convolution operation is $C_{params} = C \times C \times K \times K$, and the number of parameters for Ghost convolution is $G_{params} = C \times C/2 \times K \times K + C/2 \times K \times K$. It can be seen that Ghost convolution can reduce a large number of redundant parameters and achieve efficient feature extraction.

Although SPD can avoid the loss of feature information, it will increase the channel dimension and the number of parameters will also increase. To solve this problem, Ghost convolution is used to fuse channel information after SPD operation, which can reduce the parameter number of the model and improve the efficiency and performance of the model.

3.2. Multi-scale Detection Improvement

The Head part of the YOLOv5 network uses three different scales of detection layers. Assuming that the input image size is 640×640, three different sizes of feature maps of 80×80, 40×40 and 20×20 can be obtained

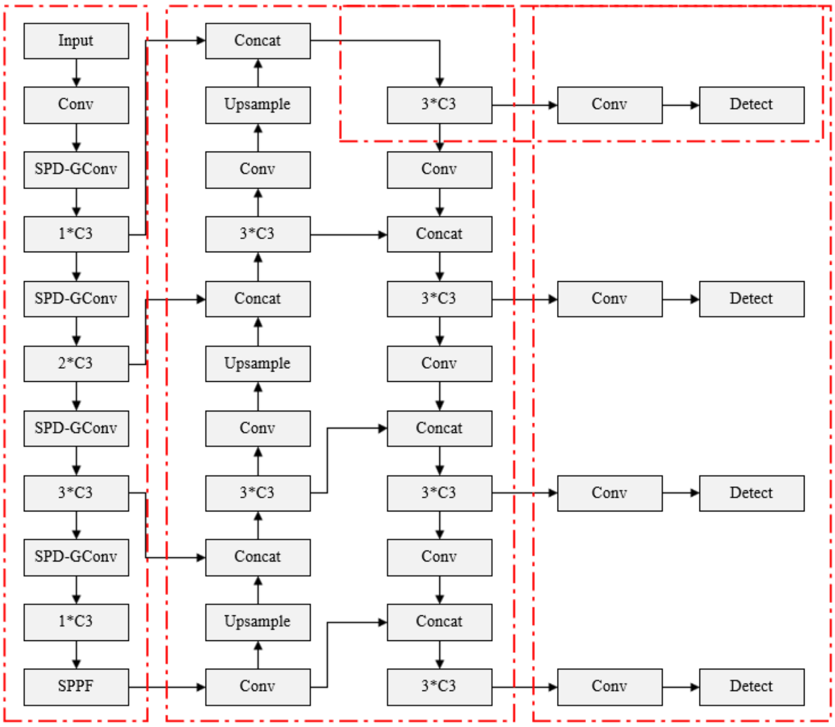

Fig. 1. Improved YOLOv5 structure diagram

through the feature extraction of the backbone network, which can be used to detect large, medium and small targets. However, there are often many long-distance pedestrian targets in the actual surveillance, and these targets are usually smaller in the image and video. The original YOLOv5 algorithm will miss the detection of such small targets. In order to improve the detection accuracy of the network for small target pedestrians, a P2 detection layer with a scale of 160 × 160 is improved on the basis of the original three detection layers, and a higher resolution feature map is introduced into the network to better capture and locate the details of long-distance pedestrians, thus reducing the case of missing detection. The improved network is shown in Figure 2.

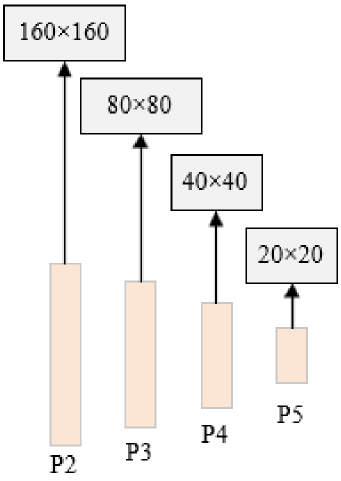

Fig. 2. Multi-scale detection structure

3.3. Improved Loss Function

The original YOLOv5 uses CIoU_Loss as the boundary loss function, and the calculation formula of CIoU is as follows:

$$L_{CIoU} = 1 - IoU + \frac{p^2(b, b^{gt})}{C^2} + av. \quad (1)$$

Where, IoU is the ratio of intersection and union of prediction boundary and GT boundary. p is the Euclidean distance between the two central points. b and b^{gt} are the center points of the real box and the real box respectively. C is the minimum diagonal length of the external rectangle of the two frames. a indicates the trade-off parameter. The calculation formula is as follows:

$$a = \frac{v}{(1 - IoU) + v}. \quad (2)$$

In the formula, v is used to measure the consistency of the aspect ratio of the two frames, and the calculation formula is as follows:

$$v = \frac{4}{\pi}(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h})^2. \quad (3)$$

Although CloU introduces the center point distance and aspect ratio by adding an influence factor, v in the formula reflects the difference in aspect ratio and cannot reflect the real difference between the width and height position and the confidence degree, which may lead to slow convergence and inaccurate regression. Therefore, this article uses the α-EIoU loss function to replace the CIoU loss function used by YOLOv5. EloU divides the aspect ratio to calculate the width and height respectively, taking into account not only the loss of the center point, but also the real difference in width and height between the target and the anchor frame. The EloU loss function is divided into IoU loss, distance loss and width and height loss. The calculation formula is as follows:

$$L_{EIoU} = L_{IoU} + L_{dis} + L_{asp}. \quad (4)$$

$$L_{EIoU} = 1 - IoU + \frac{p^2(b, b^{gt})}{C^2} + \frac{p^2(\omega, \omega^{gt})}{C_\omega^2} + \frac{p^2(h, h^{gt})}{C_h^2}. \quad (5)$$

Where C_ω and C_h are the width and height of the minimum external rectangle of the predicted box and the real box respectively. On the basis of this loss function, a unified exponentiated α-IoU loss function is added. By adding a $power$ parameter to the IoU loss function, the IoU loss and gradient can be weighted adaptively and dynamically adjusted to achieve the effect of improving the regression accuracy at different levels. The calculation formula of the improved α-EIoU is as follows:

$$L_{\alpha-EIoU} = 1 - IoU^\alpha + \frac{p^{2\alpha}(b, b^{gt})}{C^{2\alpha}} + \frac{p^{2\alpha}(\omega, \omega^{gt})}{C_\omega^{2\alpha}} + \frac{p^{2\alpha}(h, h^{gt})}{C_h^{2\alpha}}. \quad (6)$$

4. Experiment and Result Analysis

This paper verifies the proposed target detector on Ubuntu 18.04.4 LTS system. Trained and tested on four graphics processing units, NVIDIA GeForce RTX 3090 (24GB), using Intel(R)Xeon(R)Silver 4210 CPU2.40GHz and Python 3.8. CUDA uses version 11.4 and PyTorch uses version 1.8.0. During model training, the input image size is set to 640 × 640, using a gradient descent optimizer. The initial learning rate is set to 0.01, the learning rate factor to 0.1, the momentum to 0.937, and the total number of training iterations is 300.

4.1. Data Sets and Evaluation Indicators

The experiment used the WiderPerson data set, a diverse and dense pedestrian detection data set with rich foreground and background images, as well as rich crowd scenes with many pedestrians highly blurred. The WiderPerson data set divides pedestrians into five categories. The first category is complete pedestrians. The second category is people who ride electric bicycles or bicycles. The third category is partially visible pedestrians, all of whom are shielded to varying degrees. The fourth category ”neglected areas” consists mainly of objects that look like people but are not. The fifth category is the densely populated population. Since the neglected areas and groups of people are not people, we remove the labels for both categories and merge pedestrians, riders, and partially visible people into the people category for the experiment. Since the test data and real frame labels of the original WiderPerson data set are not disclosed, 90% of the original training set is used as our training set, 10% of the original training set is used as our validation set, and the original validation set is used as our test set in this experiment.

In the experiment, Precision (P), Recall (R) and average accuracy (AP) are used as evaluation indicators [22-25]. The specific formula is as follows.

$$P = \frac{TP}{TP + FP}. \tag{7}$$

$$R = \frac{TP}{TP + FN}. \tag{8}$$

$$AP = \int_0^1 P(r)dr. \tag{9}$$

Where TP represents the correct prediction of the model. FP stands for incorrect prediction of the model. FN stands for incorrectly identifying a positive example in the sample as a negative example. mAP0.5 is the average accuracy mean for all classes with an IoU=0.5.

4.2. Results

In this paper, the improved algorithm is compared with A, B, C and D on the data set, and the detection results are shown in Table 1. As can be seen from Table 1, the algorithm in this paper has the highest mAP0.5. Compared with algorithms A, B, C and D, mAP0.5 of the proposed algorithm is improved by 24.8%, 26.6%, 21.9% and 2% respectively.

Table 1. Comparison of experimental results

Method	P/%	R/%	mAP0.5/%
A	53.2	46.5	48.8
B	51.7	43.8	47.0
C	25.5	56.0	51.7
D	77.8	64.2	71.6
Proposed	77.1	64.8	73.6

We use the WiderPerson data set to conduct ablation experiments on the model, and add the improved methods mentioned in this paper one by one or in combination to verify the effectiveness of the improved methods in this paper. All experimental settings have the same parameters and are conducted in the same environment. The experiment adds various improvements to YOLOv5, including adding SPD and α-IoU. As shown in Table 2, experimental results show that each improvement increases P, R and mAP0.5. Compared with the basic model, P, R and mAP0.5 after adding SPD increase by 0.9%, 0.8% and 1.2%, respectively. When α-IoU is added, P, R and mAP0.5 increase by 0.4%, 0.6% and 0.8%, respectively. Finally, the performance of the proposed model is better than the above algorithm models.

Table 2. Ablation experiment

Method	P/%	R/%	mAP0.5/%
YOLOv5	76.3	63.4	71.6
YOLOv5+SPD	77.2	64.2	72.8
Proposed	77.6	64.8	73.6

5. Conclusion

This paper introduces the pedestrian detection method based on YOLOv5s, in order to improve the detection accuracy of the model and strengthen the ability to extract dense and long-distance pedestrian targets. SPD-GConv is used to replace the common convolutional module of the backbone structure, and a small scale detection head is added, and the loss function is modified. The experimental results show that compared with other algorithms, the evaluation index of the improved YOLOv5 algorithm is greatly improved, and the problem of missing and false detection in pedestrian detection is reduced. Subsequent work will investigate how to design lightweight network structures to increase detection speed while maintaining detection accuracy, so that the model can run on devices with limited computing resources.

6. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. None.

References

1. Ghari B, Tourani A, Shahbahrami A, et al. Pedestrian detection in low-light conditions: A comprehensive survey[J]. Image and Vision Computing, 2024: 105106.
2. Park S, Kim H, Ro Y M. Robust pedestrian detection via constructing versatile pedestrian knowledge bank[J]. Pattern Recognition, 2024, 153: 110539.
3. Teng L, Qiao Y, Yin S. Underwater image denoising based on curved wave filtering and two-dimensional variational mode decomposition[J]. Computer Science and Information Systems, 2024, 21(4): 1765-1781.
4. Liu Q, Ye H, Wang S, et al. YOLOv8-CB: dense pedestrian detection algorithm based on in-vehicle camera[J]. Electronics, 2024, 13(1): 236.
5. Zhang Y, Zeng W, Jin S, et al. When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 430-448.
6. Yin S, Wang Q, Wang L, et al. Multimodal deep learning-based feature fusion for object detection in remote sensing images[J]. Computer Science and Information Systems, 2025, 22(1): 327C344.
7. Kim T, Shin S, Yu Y, et al. Causal mode multiplexer: A novel framework for unbiased multispectral pedestrian detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 26784-26793.
8. Tashk A, Alavianmehr M A. Enhanced Pedestrian Detection and Tracking Using Multi-Person Pose Extraction and Deep Convolutional LSTM Network[C]//2024 IEEE International Conferences on Internet of Things (iThings). IEEE, 2024: 386-391.
9. Gao S. Exploration and evaluation of faster R-CNN-based pedestrian detection techniques[J]. Applied and Computational Engineering, 2024, 32: 185-190.
10. Lei S, Yi H, Sarmiento J S. Synchronous End-to-End Vehicle Pedestrian Detection Algorithm Based on Improved Y-OLOv8 in Complex Scenarios[J]. Sensors, 2024, 24(18): 6116.
11. Wang Q, Liu F, Cao Y, et al. LFIR-YOLO: Lightweight Model for Infrared Vehicle and Pedestrian Detection[J]. Sensors, 2024, 24(20): 6609.
12. Yin S, Li H, Laghari A A, et al. An Anomaly Detection Model Based on Deep Auto-Encoder and Capsule Graph Convolution via Sparrow Search Algorithm in 6G Internet of Everything[J]. IEEE Internet of Things Journal, 2024, 11(18): 29402-29411.
13. Kim T, Chung S, Yeom D, et al. Mscotdet: Language-driven multi-modal fusion for improved multispectral pedestrian detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024.
14. Fang Y, Pang H. An improved pedestrian detection model based on YOLOv8 for dense scenes[J]. Symmetry, 2024, 16(6): 716.
15. Liu S, Cao L, Li Y. Lightweight pedestrian detection network for UAV remote sensing images based on strideless pooling[J]. Remote Sensing, 2024, 16(13): 2331.
16. Guang J, Hu Z, Wu S, et al. RPEA: A residual path network with efficient attention for 3d pedestrian detection from LiDAR point clouds[J]. Expert Systems with Applications, 2024, 249: 123497
17. Wang B, Li Y Y, Xu W, et al. VehicleCpedestrian detection method based on improved YOLOv8[J]. Electronics, 2024, 13(11): 2149.
18. Wang L, Shoulin Y, Alyami H, et al. A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images[J], vol. 11, no. 3, pp. 237-251, 2024. https://doi.org/10.1002/gdj3.162.
19. Jiang Y, Yin S. Heterogenous-view occluded expression data recognition based on cycle-consistent adversarial network and K-SVD dictionary learning under intelligent cooperative robot environment[J]. Computer Science and Information Systems, 2023, 20(4): 1869-1883.

20. Teng L, Qiao Y, Shafiq M, et al. FLPK-BiSeNet: Federated learning based on priori knowledge and bilateral segmentation network for image edge extraction[J]. IEEE Transactions on Network and Service Management, 2023, 20(2): 1529-1542.
21. Lu X, Yang R, Zhou J, et al. A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(5): 1755-1767.
22. Meng X, Wang X, Yin S, et al. Few-shot image classification algorithm based on attention mechanism and weight fusion[J]. Journal of Engineering and Applied Science, 2023, 70(1): 14.
23. Jisi A, Yin S. A new feature fusion network for student behavior recognition in education[J]. Journal of Applied Science and Engineering, 2021, 24(2): 133-140.
24. Wang X, Yin S, Sun K, et al. GKFC-CNN: Modified Gaussian kernel fuzzy C-means and convolutional neural network for apple segmentation and recognition[J]. Journal of Applied Science and Engineering, 2020, 23(3): 555-561.
25. Yu J, Li H, Yin S L, et al. Dynamic gesture recognition based on deep learning in human-to-computer interfaces[J]. Journal of Applied Science and Engineering, 2020, 23(1): 31-38.

Biography

Ruopeng Pei is with the Shenyang Normal University. Research direction is computer application and AI.