Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 1, 2022

Small pedestrian target detection based on YOLOv5

Ziyi Zhang [1], Xuewen Ding [1,2,*]

[1] School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China
[2] Tianjin Yunzhitong Technology Co., Ltd., Tianjin 300350, China

[*] **Corresponding author**: Xuewen Ding (Email: dingxw1@126.com)

Abstract: YOLOv5s is the network with the smallest depth and feature map width and the fastest image inference, but when applied to small pedestrian target detection in complex scenes, the detection still suffers from wrong and missed detections. To address this problem, an improved model based on YOLOv5s is proposed with the addition of a new convolutional neural module, SPD-Conv, which improves the accuracy of the network in detection tasks of low-resolution images or smaller objects. The improved YOLOv5s-SPD model obtained better detection results compared with the original network model, with an average accuracy improvement of 3.9% and an increase in mAP value of about 9.9%.

Keywords: YOLOv5s; Pedestrian detection; YOLOv5s-SPD; Small pedestrian target.

1. Introduction

Target detection is a hot research topic in the field of machine vision [1], and pedestrian detection [2], which is one of the important components of the target detection task, has a higher research and commercial value. It is a prerequisite for pedestrian segmentation [3] and pedestrian re-identification [4], and drives the development of other target detection tasks.

As pedestrian detection in realistic environments is unavoidably affected by the environment, e.g. exposure and shadow surfaces due to strong daylight exposure; blurred pedestrian features caused by foggy[5] and rainy[6] weather; the distance of pedestrians from the camera in surveillance scenes, which can lead to differences in scale spanning; and the small pedestrian problem caused by the dense pedestrian flow in special scenes such as high-speed railway stations, airports, and public gathering places[7] can all affect the effectiveness of detection. To address the problems of low detection accuracy of obscured pedestrian targets and small pedestrian targets in real scenes, Zou Ziyin and Li Jinyu[8] proposed a series of solutions to further improve the detection accuracy and optimise the performance of the model. However, for small pedestrian targets, the features extracted by the model contain a large amount of redundant background information, and the detection accuracy still needs to be improved. To address these problems, this paper proposes an improved SPD-Conv (Space-to-depth layer and non-strided Convolution layer) algorithm based on YOLOv5s[9] for the detection of small pedestrian targets in complex scenes[10], which improves the detection capability of the network for small pedestrian targets.

2. YOLOv5 algorithm and improvements

2.1. YOLOv5 network structure

The YOLOv5 model was proposed in June 2020 by Glenn Jocher from the Ultralytics team, who updated YOLOv5 after studying YOLOv3. The initial version of YOLOv5 is very fast, efficient and easy to use. YOLOv5s is the smallest network in terms of depth and width of the feature map, and the fastest inference speed of 0.007s. The network structure of YOLOv5s consists of four main components, namely the input, the reference network, the Neck network, and the Head output.The network structure of YOLOv5s is shown in Figure 1.

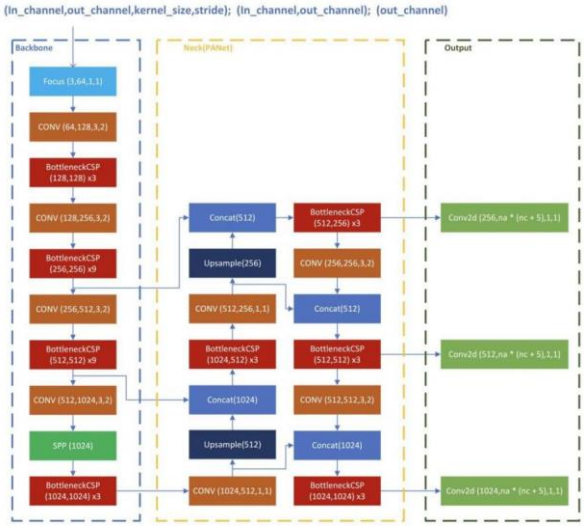

Figure 1. YOLOv5s model structure

2.2. YOLOv5 effect demonstration

The test results of the different versions of the YOLOv5 detection algorithm on the MS COCO dataset without using any other datasets or pre-trained weights are shown in Figure 2. Where the grey dash is the EfficientDet model and the remaining four are different network models of the YOLOv5 family.

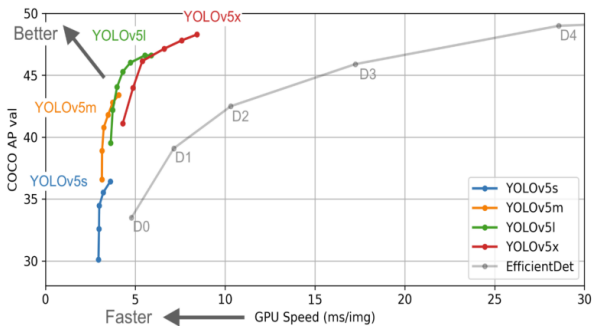

Figure 2. Testing of the YOLOv5 weighting file

2.3. Improvements to the YOLOv5 algorithm

Convolutional neural networks (CNN) have achieved great success in computer vision tasks such as image classification and target detection. However, the loss of fine-grained information caused by the convolutional and pooling layers of the convolutional neural network itself and the low feature extraction capability lead to a rapid degradation of the network's detection accuracy in low-resolution images or detection tasks of smaller objects. Therefore, this paper adds a new convolutional neural module, SPD-Conv, which is composed of a space-to-depth (SPD) layer and a non-strided convolution (Conv) layer. Where space_to_depth means superimposing the dimensions on the length and width to the depth, which is equivalent to the pooling layer, but pooling is choosing one of all sizes, whereas this method takes one of the sizes and superimposes the rest to the depth direction, thus preserving the low latitude features, the figure below shows when block_size=2 (block_size is the pooled size of the block in the pooling) the schematic diagram of SPD-Conv.

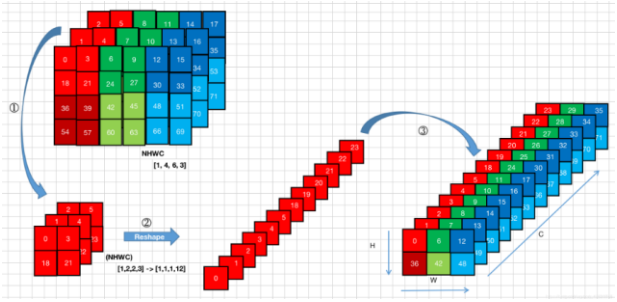

Figure 3. Schematic diagram of SPD-Conv with block_size=2

The YOLOv5s-SPD model after adding SPD-Conv simply replaces the YOLOv5 stride-2 convolution with SPD-Conv, which is structured as follows.

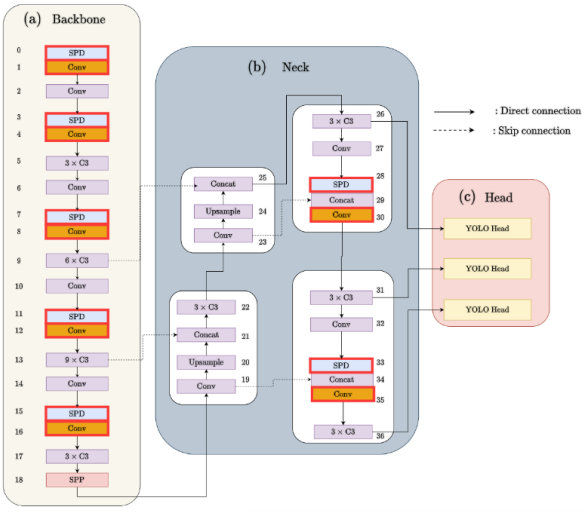

Figure 4. YOLOv5s-SPD model structure

3. Experiments and analysis of results

3.1. Description of the relevant data sets

In this paper, we use the Caltech Pedestrian dataset released by the California Institute of Science and Technology in 2009, which consists of the training set + test set data in. seq format and the pedestrian label data in.ebb (video bounding box) format. For training with YOLO, we need .jpg images and .txt annotated data, so we need to convert the. seq. vbb data into the corresponding images and annotated data.

3.2. Comparison of detection accuracy of different models

In this experiment, 2000 images from Caltech Pedestrian were randomly selected as the training set and 400 images were trained 300 times as the validation set. The evaluation effect of the datasets of YOLOv5s and YOLOv5s-SPD are shown in Figure 5 and Figure 6.

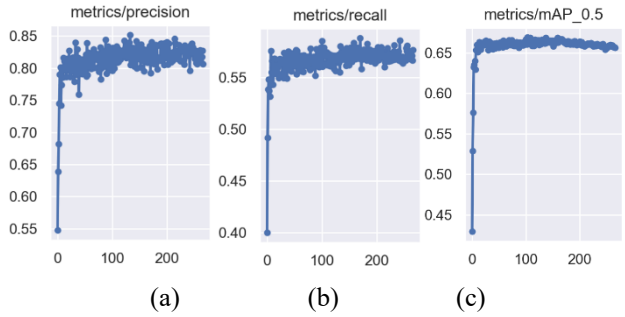

Figure 5. YOLOv5s training results

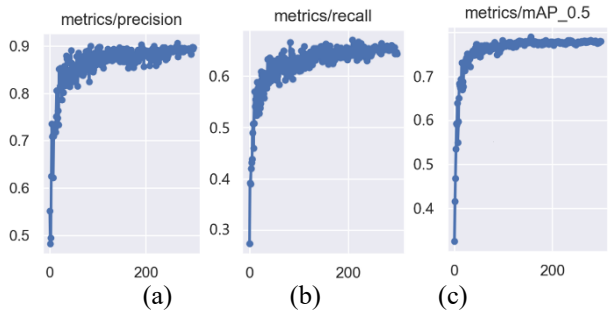

Figure 6. YOLOv5s-SPD training results

The specific values evaluated for the YOLOv5s and the modified YOLOv5s-SPD datasets are shown in Table 1 below.

Table 1. Evaluation of results for the Caltech Pedestrian dataset

Model type	Precision	Recall	mAP@0.5
YOLOv5s	0.816	0.564	0.652
YOLOv5s-SPD	0.855	0.626	0.751

From the experimental results it can be seen that the improved YOLOv5s has improved accuracy by 3.9%, recall has improved by 6.2% and mAP values have improved by approximately 10%. This shows that the improved YOLOv5s does provide a good improvement on the small pedestrian target detection problem.

After training, 267 images were randomly selected from Caltech Pedestrian's test set for testing. A comparison of the specific test results is shown in Figure 7. Where (a) is the test result of the YOLOv5s model and (b) is the test result of the YOLOv5s-SPD model.

(a)YOLOv5s test results (b)YOLOv5s-SPD test results

Figure 7. Test results

The small pedestrian target person in Fig.7 (a) is not detected; whereas the small target pedestrian person in the improved test result (b) is correctly detected with an accuracy of 83%. Thus YOLOv5s-SPD improves the problem of missed and false detections and poor detection, but the model

can be optimised by expanding the dataset and performing more training sessions. In summary, the YOLOv5s-SPD algorithm improves the network's ability to detect small pedestrian targets with better accuracy, reduced miss detection rates and improved detection accuracy.

4. Conclusion

In this paper, an improved pedestrian detection model for complex scenes based on the YOLOv5s model is proposed to address the problems of missed detection, false detection and poor detection results in complex scenes using YOLOv5s detection. A new convolutional neural module, SPD-Conv, is added to improve the accuracy of the network in detection tasks with low-resolution images or smaller objects. The improved YOLOv5s-SPD model yields relatively good detection results compared to the original network model: the average accuracy improvement is increased by 3.9% and the mAP value is increased by about 9.9%.

Acknowledgment

This work was supported by TianJin Science and Technology Commissioner Project, grant number 20YDTPJC01110.

References

[1] Chollet F.Xception:Deep learning with depthwise separable convolutions[C] //Proceedings of the IEEE conference on computer vision and pattern recognition.2017:1251-1258.W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Anides Esteban,Garcia Luis,Sanchez Giovanny,Avalos Juan Gerardo,Abarca Marco,Frias Thania,Vazquez Eduardo,Juarez Emmanuel,Trejo Carlos,Hernandez Derlis. A biologically inspired spiking neural P system in selective visual attention for efficient feature extraction from human motion[J]. Frontiers in Robotics and AI,2022,9.B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.

[3] Zhang Jianlong,Liu Chishuai,Wang Bin,Chen Chen,He Jianhui,Zhou Yang,Li Ji. An infrared pedestrian detection method based on segmentation and domain adaptation learning[J]. Computers and Electrical Engineering,2022,99.J. Wang, “Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),” IEEE J. Quantum Electron., submitted for publication.

[4] Zhang Renjie,Fang Yu,Song Huaxin,Wan Fangbin,Fu Yanwei,Kato Hirokazu,Wu Yang. Specialized re-ranking: A novel retrieval-verification framework for cloth changing person re-identification[J]. Pattern Recognition,2023,134.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces (Translation Journals style),” IEEE Transl. J. Magn. Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[5] Zhang H,Patel V M. Densely Connected Pyramid Dehazing Network[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3194-3203.J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility (Periodical style),” IEEE Trans. Electron Devices, vol. ED-11, pp. 34–39, Jan. 1959.

[6] Zhang H,Patel V M.Density-aware Single Image De-raining Using a Multi-stream Dense Network[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 695-704.R. W. Lucky, “Automatic equalization for digital communication,” Bell Syst. Tech. J., vol. 44, no. 4, pp. 547–588, Apr. 1965.

[7] Dong X W, Han Y, Zhang Z, et al. Metro Pedestrian Detection Algorithm Based on Multi-scal -e Weighted Feature Fusion Network[J]. Journal of Electronics & Information Technology, 2 021,43(7): 2113-2120.

[8] Li J Y, Yang J, Kong B, et al. Multi-scale vehicle and pedestrian decection algorithm based on attention mechanism [J]. Optics and Precision Engineering,2021, 29(6): 1448-1458.

[9] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, IEEE conference on computer vision and pattern recognition. 2016, pp. 779–788.

[10] Jingwei Cao,Chuanxue Song,Silun Peng,Shixin Song,Xu Zhang,Yulong Shao,Feng Xiao. Pedestrian Detection Algorithm for Intelligent Vehicles in Complex Scenarios[J]. Sensors,2020,20(13).J. G. Kreifeldt, “An analysis of surface-detected EMG as an amplitude-modulated noise,” presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.