YOLO-DSRF: An Improved Small-Scale Pedestrian Detection Algorithm Based on Yolov4

Runjie Liu, Shuguang Li * , Yipeng Duan and Lei Shi

National Supercomputing Center in Zhengzhou, Zhengzhou University, Zhengzhou 450001, Henan, China

Abstract

Although existing object detection algorithms have achieved good results, it is still a challenge to effectively detect small-scale pedestrians in real time. Aiming at the problems of complex structure, large number of parameters and high missed detection rate of small targets in existing pedestrian detection algorithms, the YOLO-DSRF pedestrian detection algorithm is proposed. On the basis of YOLOv4, the depth separation convolution was first introduced to significantly reduce the amount of parameters and computation of the model, and the channel attention mechanism was introduced into the network to improve the influence of important channel features on the network, a feature fusion module was designed in the backbone network by merging deep and shallow features to effectively extract target semantic information and location information, and introduce a receptive field module in the detection head to simulate the human receptive field to enhance feature extraction capability for small targets. For training and verification on the Caltech dataset, compared with the original algorithm, the number of parameters is reduced by 65.2%, and the running speed on the GPU is increased by 20%. The AP is roughly the same. The algorithm proposed in this paper can effectively reduce the model complexity while ensuring accuracy. Thereby increasing the running speed.

Key words

Pedestrians Detection, YOLOv4, Depthwise Separable Convolution,SE,Receptive Field Block, Feature Fusion Module

1.Introduction

Object detection has always been a hot research direction in the field of computer vision. Its task is to accurately identify and locate all objects in an image or video. Pedestrian detection is an important branch of object detection, which is widely used in re-identification, and intelligent monitoring[1]. The current mainstream pedestrian detection algorithms are based on deep learning, which includes one-stage algorithm and two-stage algorithm. The two-stage algorithm is based on candidate region extraction. The main algorithms include RCNN[5], Faster-RCNN[6], Mask-RCNN[7], etc. Although the accuracy is high, it's running slow. The one-stage algorithm has a simple structure and can directly detect the image output results, so the calculation efficiency is high and the running speed is fast. The main algorithms include SSD series[8] and YOLO series[9]. However, the accuracy of the single-stage target detection algorithm is low. Due to the low resolution of small-scale pedestrians and less feature information, it is easy to cause false detection and missed detection due to the influence of the surrounding environment and image noise. How to accurately and quickly detect small target pedestrians has become a hot research problem in the field of pedestrian detection.

The feature extraction network of the pedestrian detection algorithm has a low downsampling factor and a small receptive field, and pays more attention to small-scale pedestrian targets. However, due to the weak ability to represent semantic information, it is easy to detect false object. Deep features have large downsampling multiples and large receptive fields, which may easily lead to misse a lot of small-scale pedestrian targets. To solve the above problems, some scholars proposed FPN (Feature Pyramid

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China

*corresponding author's e-mail: iesgli@163.com (Shuguang Li *)

Networks)[10], which fuses the features of deep and shallow layers to improve the detection effect, but the detection speed of this kind of algorithm is low and cannot be detected in real-time. The YOLOv3[11] proposed by Redmon et al. uses the DarkNet-53[12] residual network for feature extraction and combines the FPN network to significantly improve the small target detection performance. Bochkovskiy A improved on YOLOv3 and proposed YOLOv4[13]. Using CSPDarkNet[14]as the backbone network, using SPP+PAN to fuse feature maps of different sizes, greatly improving the accuracy of model detection, but its complex structure and large number of parameters are difficult to deploy on mobile terminals or embedded devices.

In order to reduce the complexity of the YOLOv4 model and improve the network running speed. Wang et al.[2]replaced the YOLOv4 feature extraction network with a lightweight MobileNet and improved the general problem of small target detection by adding a shallow detection head.Li et al. [3]constructed a feature extraction network with reference to ShuffleNet and channel attention mechanism, which improved speed and ensured accuracy.

In order to make yolo more effective in detecting small targets, Gao et al. [4] introduced a threshold attention module (TAM) and embedding CAM into BiFPN as a feature pyramid network. Wei [17] introduced ASFF on the basis of FPN. Different from the layer-by-layer fusion of PANet, ASFF adaptively learns weight coefficients for all feature maps participating in the fusion, and multiplies the feature maps of each layer by the corresponding weights and then fuses them, which improves the detection accuracy of the network for small targets.

This paper proposes the YOLO-DSRF (Depthwise separable convolution & SE attention module & Receptive field module & Feature fusion module) algorithm by improving YOLOv4. First, replacing traditional convolution with depthwise separation convolution reduces the amount of parameters and improves the running speed. Second, in order to fully extract the features of small objects, the SE attention mechanism is introduced to suppress the influence of noise and enhance the learning of important channel features. Third, the FFM module is designed to integrate deep and shallow features, and to strengthen the influence of shallow features on geometric information. and semantic information representation ability. Finally, adding a receptive field module to the detection head of small-scale targets to enhance feature extraction for small-scale targets. Finally, the effectiveness of the proposed algorithm is verified on the Caltech dataset.

2.Improved Algorithm

The proposed YOLO-DSRF algorithm is improved by YOLOv4, and the improved part is the dashed box in Figure 1. Depth separable convolution is introduced to replace the 3*3 convolution kernel in the original network, and SE attention mechanism, RFB module, and top-down feature fusion module are added

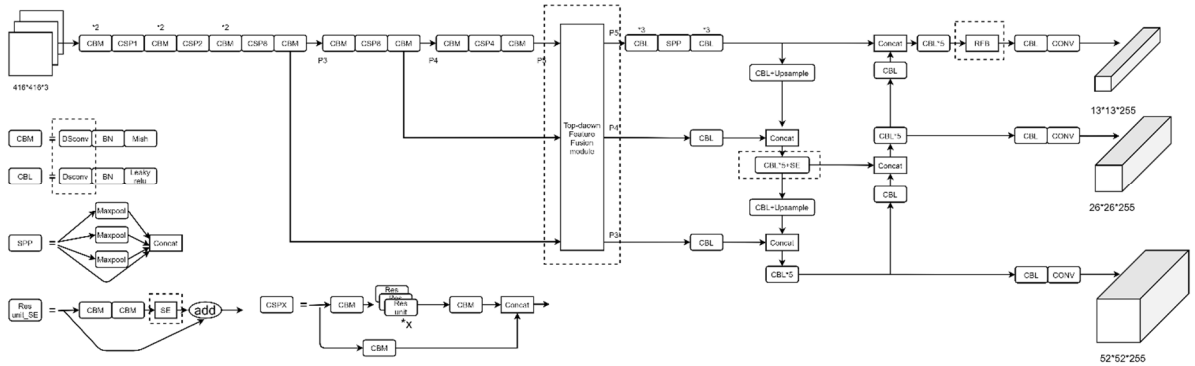

Figure 1. Overall of YOLO-DSRF network

2.1.Depthwise Separable Convolution

The depthwise separation convolution is a lightweight method for embedded devices proposed by Sandler et al. in MobileNet [15] in 2017. As shown in Figure 2, the depthwise separation convolution

consists of Depthwise convolution and Pointwise convolution. The input size The feature map of H*W*C, the Depthwise convolution uses C convolution kernels of size n*n*1 to convolve with the 1-layer feature map respectively to output a feature map. The independent operations do not effectively utilize the feature information of different channels in the same spatial position. Pointwise convolution uses C' convolution kernels of size 1*1*C to convolve with the output feature map. The output feature map of size H'*W'*C'. The parameter amount and calculation amount of depthwise separable convolution are about $1/n^2$ of ordinary convolution. In order to improve the running speed of the network to meet the real-time requirements.The convolution kernel with a size of 3*3 in YOLOv4 is replaced with a depthwise separation convolution, and its parameter amount and calculation amount are about 1/9 of that of ordinary convolution. After the replacement, the model parameter amount and calculation amount are greatly reduced, which is conducive to deployment in embedded devices.

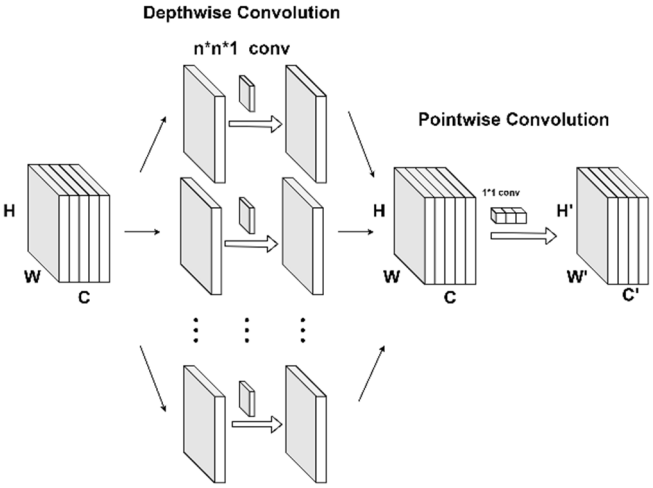

Figure 2. Depthwise separable convolution

2.2.SE Attention Mechanism

Paying attention to important parts and ignoring irrelevant parts is the attention mechanism. SENet [16] adopts the squeeze-and-excitation module to collect full set information, capture the relationship between channels, and improve the representation ability. As shown in Figure 3, SE attention mechanism includes squeeze module and excitation module. The squeeze module uses global average pooling to collect global spatial information. The excitation module uses a fully connected layer and a nonlinear layer to capture the channel relationship. The output attention vector is multiplied by each channel of the input feature for scaling. SE plays an important role in suppressing noise while strengthening. function of the channel. Since small-scale pedestrians contain few pixels, the backbone network can only extract fewer features, in order to promote the network to learn more important features and improve the detection accuracy. In this paper, SE is added to the Res unit of YOLOv4 to form the Res unit-SE module in Figure 1 to enhance the ability of the backbone network to extract important features. SE is also introduced after the convolution in the dashed box in Figure 1, which enhances the learning of the important channel features of P4, and then fuses with the P3 and P5 feature maps. The introduction here can better exert the effect of the SE module.

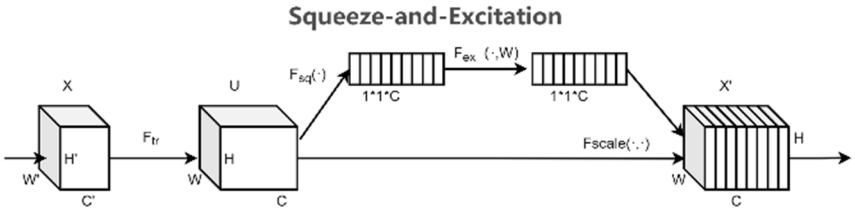

Figure 3. SE attention mechanism

2.3.Receptive Field Module

The receptive field module was proposed by Liu et al. [19] in RFBNet. Inspired by the receptive field structure in the human visual system, it simulates the group receptive field in the human superficial retinal image. As shown in Figure 4, its construction is similar to the Inception structure. A similar multi-branch convolution module that captures multi-scale information. In addition, dilated convolution is introduced to expand the sampling range to extract finer features of the target. Since there are few features of small-scale pedestrians, in order to enhance the network's feature representation for small objects, this paper introduces RFB in YOLOv4 to improve the network's detection effect on small objects.

2.4.Feature Fusion Module

The shallow features extracted by the network have low downsampling times and high resolution for detecting small-scale pedestrian targets. They have strong geometric representation capabilities, but weak semantic representation capabilities. The deep features extracted by the network have small downsampling multiples and low resolution to detect large-scale pedestrian targets. They have strong semantic representation ability, but weak geometric representation ability. In order to enhance the feature representation ability of the network to achieve accurate pedestrian detection, this paper designs a top-down FFM as shown in Figure 5(a) and a bottom-up FFM as shown in Figure 5(b). Strengthen the fusion of deep features and shallow features in the backbone network, and enhance the semantic information representation ability of shallow features to improve the detection effect of small-scale pedestrian targets. Check the effect. The module uses P3, P4, and P5 in Figure 1 as input. It can be seen from Figure 5 that the top-down FFM first upsamples P5 and stacks P4, and then performs convolution to achieve feature integration and output P4' to replace the original network. P4; the convolution output P3' performed after stacking the resulting P4' upsampling with P3 replaces P3 in the original network. Bottom-up FFM is to downsample the P3 feature map with a convolution with a stride of 2 and stack it with P4, and then perform the convolution output to integrate the output P4', stack the resulting P4' downsampling with P5, and convolve the output P5 ', both operations strengthen the fusion of shallow and deep features in the backbone network, and enhance the geometric and semantic representation capabilities of features, thereby improving the accuracy of network detection.

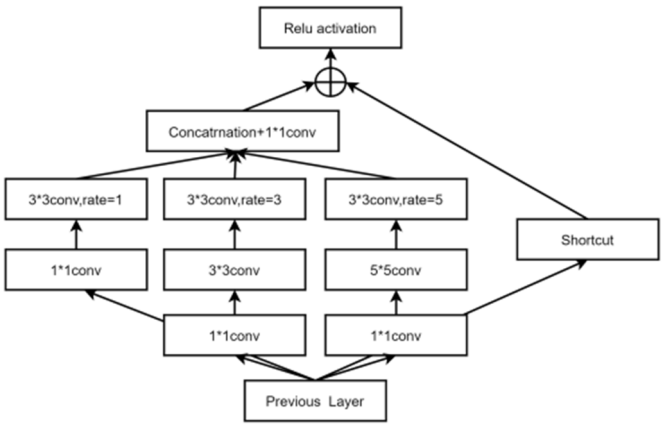

Figure 4 Receptive field module

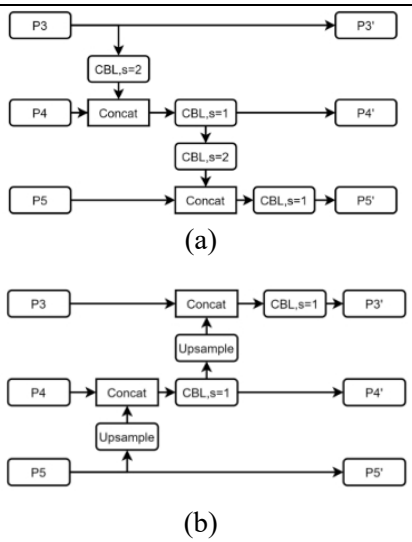

Fig 5 Feature fusion module

3.Experiment

3.1.Dataset

The experiment uses the Caltech[18] pedestrian dataset, which is collected from the street view road and collected from the conventional street view road, including 350,000 pedestrian annotation boxes and 2,300 different pedestrians. Small-scale pedestrian detection dataset. After the data set is filtered, it is divided into training set, validation set and test set according to the proportion.

3.2.Experimental Environment And Parameter Configuration

The experimental environment in this paper is the GNU operating system, the graphics card is Tesla V100s, the CUDA version is 11.0, the Pytorch 1.8 deep learning framework, and the compiled language is Python3.6.8. The experiments use the Adam optimizer with a batch size of 8 and a cosine annealing learning rate with an initial learning rate of 0.001.

3.3.Measurement indicators

This paper uses five indicators of P (Precision), R (Recall), AP(Average Precision), PARAM (Parameters), FPS (Frame Per Second) five indicators to evaluate network performance, The calculation formulas of P , R and AP are shown in formula (1-3).

$$P = \frac{TP}{TP+FP} * 100\% \quad (1)$$

$$R = \frac{TP}{TP+FN} * 100\% \quad (2)$$

$$AP = \int_0^1 P\ (R)\ dR \quad (3)$$

TP (True Positive) is the number of targets correctly detected by the model, FP (False Positive) is the number of falsely detected targets, and FN (False Negative) is the number of false detections and missed detections. AP is the area under the P-R curve, which is used to measure the detection capability of the network; the parameter quantity is used to measure the complexity of the network; FPS is the number of pictures processed per second, which is used to measure the computing power of the network.

Table 1. Performance comparison of each algorithm.

Model	PAMRAM(M)	P(%)	R(%)	AP(%)	FPS
YOLOv4	245.24	92.15	76.04	84.96	25
YOLOv4-D	68.38	90.99	72.19	82.46	39
YOLOv4-D-SE	82.49	91.41	73.41	83.39	35
YOLOv4-D-RFB	68.73	91.63	74.18	83.56	38
YOLOv4-D-FFM1	80.98	91.62	74.46	83.93	37
YOLOv4-D-FFM2	73.42	91.65	74.47	83.98	37
YOLOv4-D-SE-RFB	82.93	91.99	74.85	84.28	35
YOLOv4-D-SE-FFM1	95.09	92.12	75.16	84.34	33
YOLOv4-D-SE-FFM2	87.52	92.18	75.54	84.69	33
YOLO-DSRF	87.87	92.22	75.71	84.81	30

3.4.Results and Analysis

In order to verify the effectiveness of the algorithm proposed in this paper, an ablation experiment is designed. First, YOLO-D replaces the 3*3 convolution in the network with a depthwise separable convolution, and then introduces the SE attention mechanism. , RFB, and the two FFMs designed in this paper. The results are shown in Table 1. The introduction of depthwise separable convolution greatly reduces the amount of network parameters, which is only 27.9% of YOLOv4. 2.5%, and Recall decreased by nearly 4%. The introduction of the SE attention mechanism increases the number of parameters by 14.11M, the AP increases by 0.93%, and the Recall increases by 1.22%, which proves that the SE attention mechanism can enhance the feature extraction ability. The introduction of RFB in the network detection head only increases the amount of parameters by 0.35M, the Recall is increased by about 2%, and the AP is increased by 1.1%, which proves that RFB can more effectively extract the features of small targets. YOLO-D-FEM1 introduces the bottom-up feature fusion module designed in this paper, the parameters are increased by 12.6M, the AP is increased by 1.52%, and the Recall is increased by 2.27%. YOLO-D-FFM2 introduces the top-down feature fusion module designed in this paper. The number of parameters is increased by 5.04M, the AP is increased by 1.47%, and the Recall is increased by 2.28%. It is proved that the two modules designed in this paper can improve the representation ability of semantic information and geometric information by fusing deep and shallow features in the backbone network. In addi-tion, based on YOLO-D-SE, RFB, bottom-up FFM, and top-down FFM are introduced respectively. Compared with YOLO-D-SE, AP increased by 0.89%, 0.95%, 1.3%; Recall increased by 0.89%, 0.95%, 1.3%. Finally, for the algorithm YOLO-DSRF proposed in this paper, compared with YOLO-D, AP, Recall, and Precision have been improved by 2.35%, 3.52%, and 1.23 respectively; compared with YOLOv4, the difference in AP is only 0.15%; the difference in Recall is only 0.33 %; increased by 0.07%. Compared with YOLO, the number of parameters is increased by 19.49M, which is only 35.8% of YOLOv4. Runs 20% faster on GPU

4.Conclusion

In view of the complex structure of the YOLOv4 algorithm, the poor real-time performance of mobile devices and the insufficient extraction of small target pedestrian features, this paper adopts the depth separable convolution to replace the ordinary convolution, and introduces the SE attention mechanism, RFB and FFM designed in this paper. The designed YOLOv4-DSRF algorithm greatly reduces the amount of parameters and computation, and enhances the feature extraction capability of the network, especially the feature extraction capability for small targets. However, limited by the low resolution of the feature map extracted by the network, it is not conducive to the feature analysis of small objects. In the future research work, we will study the construction of a high-resolution lightweight network. Also, it is very important to study the appropriate feature fusion method to improve the detection of small objects. optimizing the feature fusion method to further improve the detection effect of small target pedestrians.

5.References

[1]. Li J, Liang X, Shen S M, et al. Scale-aware fast R-CNN for pedestrian detection[J]. IEEE transactions on Multimedia, 2017, 20(4): 985-996.

[2]. Wang H, Zang W. Research On Object Detection Method In Driving Scenario Based On Improved YOLOv4[C]//2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC). IEEE, 2022, 6: 1751-1754.

[3]. Li Y, Lv C. Ss-yolo: An object detection algorithm based on YOLOv3 and shufflenet[C]//2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2020, 1: 769-772.

[4]. Gao Y, Wu Z, Ren M, et al. Improved YOLOv4 Based on Attention Mechanism for Ship Detection in SAR Images[J]. IEEE Access, 2022, 10: 23785-23797.

[5]. GIRSHICK R，DONAHUE J，DARRELL T，et al.Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2014： 580-587.

[6]. REN S，HE K，GIRSHICK R，et al.Faster R-CNN：towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，39（6）：1137-1149.

[7]. HE K，GKIOXARI G，PIOTR D，et al.Mask R-CNN[C]// IEEE International Conference on Computer Vision，2017

[8]. LIU W，ANGUELOV D，ERHAN D，et al.SSD：single shot multibox detector[C]//European Conference on Computer Vision.Cham：Springer，2016：21-37.

[9]. REDMON J，DIVVALA S，GIRSHICK R，et al.You only look once：unified，real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2016：779-788.

[10].Lin T Y，Dollár P，Girshick R，et al.Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition，2017：2117-2125.

[11].REDMON J，FARHADI A.YOLOv3：an incremental improvement[J].arXiv：1804.02767，2018.

[12].KIM K J，KIM P K，CHUNG Y S，et al.Performance enhancement of YOLOv3 by adding prediction layers with spatial pyramid pooling for vehicle detection[C]// 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance（AVSS），2018：1-6.

[13].BOCHKOVSKIY A，WANG C Y，LIAO H Y M.YOLOv4： optimal speed and accuracy of object detection[J].arXiv： 2004.10934，2020.

[14].MAHTO P，GARG P，SETH P，et al.Refining YOLOv4 for vehicle detection[J].International Journal of Advanced Research in Engineering and Technology（IJARET）， 2020，11（5）：409-419.

[15].HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. Computer Vision and Pattern Recognition arXiv, Preprint arXiv: 1704. 04861, 2017.

[16].HU J, SHEN L, SUN G. Squeeze- and- excitation networks [C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18- 22, 2018. Washington: IEEE Computer Society, 2018: 7132-7141

[17].Wei Hongyu. Aircraft target detection method in remote sensing image based on YOLOv4 anti-occlusion [D]. China University of Mining and Technology, 2021.

[18].Ess A，Müller T，Grabner H，et al.Segmentation- based urban traffic scene understanding[C]//British Machine Vision Conference，London，Sep 7-10，2009

[19].Liu S，Huang D.Receptive field block net for accurate and fast object detetion[C]//European Conference on Computer Vision.Cham：Springer，2018：385-400.