EAI Endorsed Transactions
on Energy Web
Research Article EAI.EU

Pedestrian Perception Tracking in Complex Environment of Unmanned Vehicles Based on Deep Neural Networks

Ruru Liu[1,2*], Feng Hong[2,3] and Zuo Sun[3]

[1]Merchant Marine College, Shanghai Maritime University, No. 1550, Harbour Avenue, Pudong New Area, Shanghai
[2]Intelligent Perception and Computing Research Center of Chizhou University, Chizhou University No. 1, Education Park, Chizhou City, Anhui Province, China
[3]Anhui Research Center of Semiconductor Industry Generic Technology

Abstract

INTRODUCTION: In recent years, machine learning and deep learning have emerged as pivotal technologies with transformative potential across various industries. Among these, the automobile industry stands out as a significant arena for the application of these technologies, particularly in the development of smart cars with unmanned driving systems. This article delves into the extensive research conducted on the detection technology employed by autonomous vehicles to navigate road conditions, a critical aspect of driverless car technology.
OBJECTIVES: The primary aim of this research is to explore and highlight the intricacies of road condition detection for autonomous vehicles. Emphasizing the importance of this key component in the development of driverless cars, we aim to provide insights into cutting-edge algorithms that enhance the capabilities of these vehicles, ultimately contributing to their widespread adoption.
METHODS: In addressing the challenge of road condition detection, we introduce the TidyYOLOv4 algorithm. This algorithm, deemed more advantageous than YOLOv4, particularly excels in pedestrian recognition within urban traffic environments. Its real-time capabilities make it a suitable choice for detecting pedestrians on the road under dynamic conditions.
RESULTS: The application of the TidyYOLOv4 algorithm in autonomous vehicles has yielded promising results, especially in enhancing pedestrian recognition in urban traffic settings. The algorithm's real-time functionality proves crucial in ensuring the timely detection of pedestrians on the road, thereby improving the overall safety and efficiency of autonomous vehicles.
CONCLUSION: In conclusion, the detection of road conditions is a critical aspect of autonomous vehicle technology, with implications for safety and efficiency. The TidyYOLOv4 algorithm emerges as a noteworthy advancement, outperforming its predecessor YOLOv4 in pedestrian recognition within urban traffic environments. As companies continue to invest in driverless technology, leveraging such advanced algorithms becomes imperative for the successful deployment of autonomous vehicles in real-world scenarios.

Keywords: YOLOv4, Driverless Vehicles, Complex scene perception

Received on 27 December 2023, accepted on 09 April 2024, published on 15 April 2024

doi: 10.4108/ew.5793

1. Introduction

With the development of AI profound learning technology, in a variety of disciplines, including medical diagnostics [1-4], natural language processing [5-8], unmanned driving [9-12], and others, it has demonstrated substantial application benefits. With the continued expansion of the automotive business, driverless cars have came to the fore as one of the key areas of research and advancement. The driver was less car adopts the integration of multiple technologies, including front-end environment perception,

*Corresponding author. Email: liururu@czu.edu.cn

path planning and control. The front-end environment perception and detection are the core stage of unmanned driving technology, and the perception stage is mainly visual detection. Environmental perception is an important prerequisite for the realization of the unmanned system planning and decision-making, and plays a very important role in the entire unmanned driving [13].

Because when driverless cars don't accurately detect pedestrians on the road, it can endanger pedestrians' lives. Therefore, it is crucial to guarantee that pedestrian detection is accurate. The identification of road pedestrians has been significantly enhanced due to the advancement and enhancement of deep learning algorithms, however fur-ther advancement is required for practical applications. There are two primary issues: (1) Deep learning algorithms for vision need plenty of space and processing power. A pedestrian's body part may be locked by other vehicles or traffic signs, for example, making it difficult to store and run to extract complete feature information. As a nutshell, in order to assess the characteristics of the target, one must depend on the limited information that has been collected. At the present, its main purpose is to examine and validate the server's detection performance. Machine vision is mostly divided into two categories: (1) Classification and localization challenges are combined to create target detection, with the objectives being to (1) characterize the image's element information and (2) locate its object information. In order to assess the target position and classification, the first target detection method first pulls the target information from the picture using a sliding window. The detection effect is not satisfactory. It has become one of the nice and warm professional fields in the field of sight well before incorporation of R-CNN prompted the enthusiasm of a fundamental number of investigators [14-16]. Now based on R-CNN, more excellent target detec-tion algorithms have been developed, such as R-CNN[15], Fast R-CNN[16], Mask-RCNN, R-FCN[14], SSD[17]],YOLO[18], YOLOv2[19], YOLOv3[20], YOLOv4[21], etc. According with their various network frameworks, the target detection algorithms of supervised learning are principally broadly classified into two categories: one is a two-stage target detector represented by R-CNN and Fast R-CNN, which is composed of three modules: the regional recommendation module, the backbone network, and the detection head. First, the region detection module of the two-stage object detector generates suggested regions based on the region of interest, and the detection head classifies the information based on the regions provided by the suggested regions. Finally, position regression will be used to pinpoint the target item precisely. With regional suggestion, the two-stage object detector achieves great detection accuracy. However, its ongoing operation demands more than just a lot of processing power and running memory degradation. This also results in sluggish real-time object recognition. In a different class, single-stage object detectors, such as the YOLO series and SSD, set k a priori boxes, tightly covering every specified area of the picture at every position of the feature map, without making use of a region proposal like the branch network. Therefore, the single-stage detector has a great improvement in real-time performance over the two-stage detector in terms of inference. The YOLOv4 object proposed methodology astonishes a correct compromise between sensitivity and precision amongst single-stage object detectors. It outperforms in detection accuracy and facilitated detection capability.

Reduce the vision based algorithm's space volume and processing power density to try and encourage a realis-tic widespread adoption of the target detection algorithm on the self - driving technology semiconductor. This is what will stem the tide that the deep learning objective detection method based can indeed be made available to the autonomous sensor. Pedestrians test the efficacy of the improved algorithm in the presence of partially occluded data sets to increase the detection effect of occluded pedestrians. Due to strong uncertainty in the feature infor-mation of legs, hands and bodies, it is selected to use the algorithm with significant features. The head is used as the object of the annotation, and the probability of the head being occluded on the road is relatively low. We probably partially lead to irritation the human body in the head annotation dataset whereas the particular piece of annotation is completely disconnected from widely used programming language pedestrian datasets. The preliminary results clearly demonstrate that TidyYOLOv4, an optimization framework based on this dataset, is better suited than YOLOv4 for pedestrian identification of unmanned aerial vehicles in urban traffic circumstances.

2. Network Model

2.1. Network Optimization

Based on the YOLO method, YOLOv4 is a sophisticated algorithm that is continually tuned and enhanced [21]. The YOLOv4 algorithm has significantly increased speed and accuracy and is mostly based on YOLOv3 supplemented with current sophisticated optimization methodologies. On the basis of backbone-53, CSP-Darknet-53, which has been managed to develop by fusing the ideas of YOLOv4 and CSP-Net (Cross Stage Partial Network) [20], considerably improves the transmission effectiveness of the network algorithm, while Neck It combines the advantages of SPP (Spatial Pyramid Pooling) and PAN, and head leverages the YOLOv3 detection techniques to enhance the purification of deep networks.

An SPP module was managed to add between the fifth and sixth convolution operation in the vicinity of the third detection head of YOLOv4 to significantly boost the extraction of features of the deep structure in the experiment and paired with YOLOv4-SPP1. As illustrated below:

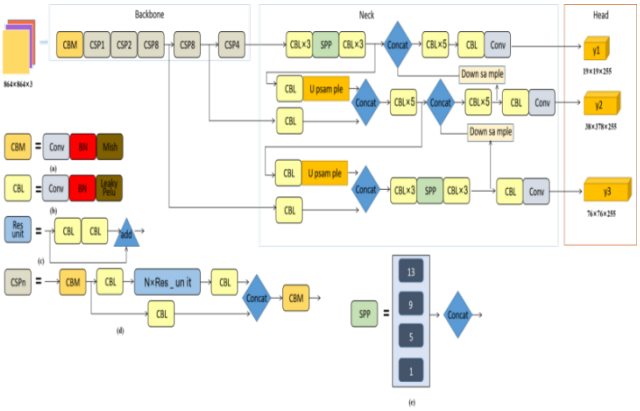

Figure.1 Architectural style of TidyYOLOv4 is depicted by a schematic.

2.2. Network Process

This paper uses the YOLO process, and on this basis uses the tensorflow method to train human detection and find human detection based on this.

Table1 Experimental verification

Algorithm 1: Adaptive BMS
■ A set of boolean operations B={ };
■ A set of attention MapsA={ }; $\bar{A} \leftarrow 0$
■ The input feature map A is generated from the thermal image$\emptyset(I)$
■ Use equations (3) and (4) to calculate the initial threshold t_1 and sampling step size for i=1 to N//
■ For $\theta = t_1 \ to \ 255$
$B_i = THRESH(\emptyset(I), \theta)$
$\bar{B}_i = INVERT(B_i)$
■ Open to the morphology of $B_i$and $\bar{B}_i$
B_i and $\bar{B}_i$ add to **B**
End
End
(1) For $i = 1$ to N //
If all pixels of $B_i(x, y)$ are connected to the image boundary, set $A_i(x, y) = 0$
Morphological expansion of A_i
(2) Normalization A_i
$\bar{A} \leftarrow \bar{A} + A_i$
End
(3)$\bar{A} \leftarrow \bar{A}/maxiA_i$ //
(4)$S \leftarrow Post \ Proces \ Sin(\bar{A})$ //

3. Experimental results

3.1. Dataset

First, it contains more than 1,000 hours of training data along a single route, rather than focusing on an expansive city. Second, it provides HD scene data with bounding boxes and class probabilities. Third, high-resolution aerial images are available in the dataset [15]. The second figure depicts the dataset's digital audio duration, which wanted to run from November 2019 to March 2020.

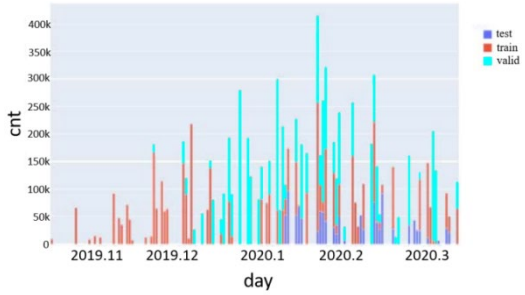

Figure. 2 Training effective test data set distribution

For three reasons, the Lyft dataset is distinct from the ones described above. First, rather of concentrating on a large metropolis, it has more than 1,000 hours of training data along a single path. Second, rather than only providing pure centroids, it also includes bounding boxes and class probabilities for HD scene data. Third, the file includes high-resolution aerial photos. The statistical distribution of the dataset is displayed in Figure 2. Figure 3 displays the dataset data, which was collected from November 2019 to March 2020.

3.2. Training optimization

For the deep learning model to be optimized, the displacement error of the path actor is introduced for the horizon Z ∈ {1, ... ,Z} at time t_p , which is the Euclidean distance between each of the agent's trajectory's estimated annual and inherent geographic areas. In equation (1), φ defines the parameters of the model, (ap(q+r), bp(q+r) define the available states S_q at a practical given time) and maps M as an input.

$$D_{p(q+r)} = ((a_{p(q+r)} - \hat{a}_{p(q+r)}(S_q, M, \varphi))^2 + (b_{p(q+r)} - bpq + rSq, M, \varphi 212 \quad (1)$$

The loss function applied in this instance at least is the Mean Squared Error (MSE), which is roughly comparable to the mean squared stress - strain error of the speed and direction points.

$$L_{pq} = \frac{1}{z} \sum_{z=1}^{Z} D_{p(q+r)}^2 \quad (2)$$

The objective is to minimize the overall training loss while optimizing the loss function in equation (2) across all participants and time steps.

$$\varphi^* = \arg\min_{\varphi} L = \arg\min_{\varphi} \sum_{q=1}^{T} \sum_{p=1}^{Nq} L_{pq} \quad (3)$$

After 30k iterations of each model training, the average training loss and average validation loss are determined. Experiments indicate that the training and validation losses are optimized through iterations, as seen in Figure 3. The ADE (Average Displacement Error) was computed as an average over several forecast horizons. The future (a,b) coordinates are predictions of the previous situation and position of various infrastructure traffic agents (cars, cyclists, and pedestrians) in the scene within 5 seconds, with a history size and shape of zero seconds, i.e. information about their actual state and position is not actually produced at the time of prediction, after the model's training data.

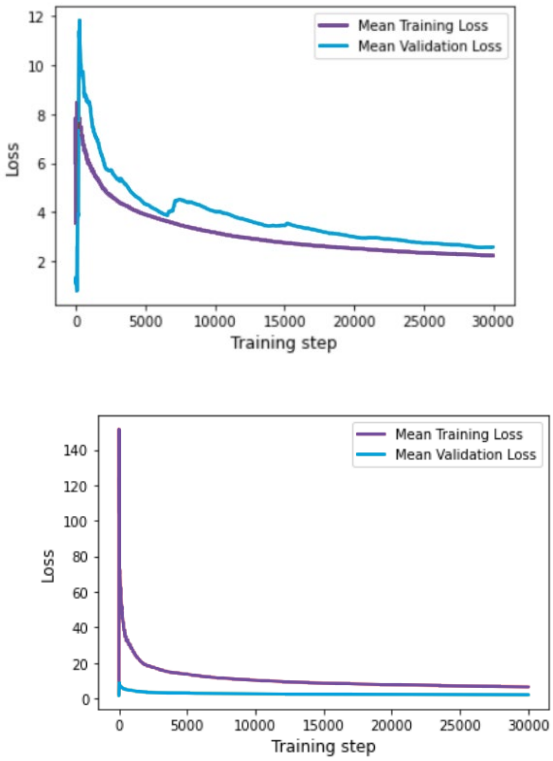

Figure.3 (a) EfficientNet (b) ResNet loss and training steps diagram

To decide on the maximum possible classification and recognition model, the regression coefficients of the basic model and the model is trained are reviewed and approved utilising variety of multi - objective optimization problems (YOLOv4-SPP1-X, SPP1 denotes that spatial Pyra-mid pooling has been included, X indicates to prune X%), according to YOLOv4.

To decide on the maximum possible classification and recognition model, the regression coefficients of the basic model and the model is trained are reviewed and approved utilising variety of multi - objective optimization problems (YOLOv4-SPP1-X, SPP1 denotes that spatial Pyra-mid pooling has been included, X indicates to prune X%), according to YOLOv4.

Table 1 Experimental verification

Model	Input size	Precious	Recall	mAP	Total BFLOPS	Inference Time (ms)	Parameters(M)	Volume (MB)
YOLOv3	416	82.28	71.98	75.08	66.27	27.98	59.50	240.38
	864	85.48	87.08	87.88	284.63	87.18		
YOLOv4	416	85.38	80.68	80.38	67.65	29.68	62.65	254.58
	864	88.58	89.88	90.48	285.25	90.38		
YOLOv4-SPP1	416	85.89	81.58	82.88	73.15	34.94	64.73	256.78
	864	89.88	91.68	93.28	312.00	91.66		
YOLOv4-SPP1-95	864	87.48	87.78	89.18	16.08	17.98	0.80	3.90
YOLOv4-SPP1-96	864	86.78	83.38	86.48	13.18	14.68	0.66	2.70
YOLOv4-SPP1-97	**864**	**84.58**	**75.18**	**81.48**	**11.97**	**13.98**	**0.45**	**1.80**

Table 1 displays the experimental outcomes of many groups with varying degrees of trimming. The Input size column in the table shows that when the size of the network input picture rises from 416416 to 864864, the evaluation indicators of YOLOv3, YOLOv4, and YOLOv4-SPP1 improve dramatically, with the mAP of YOLOv3 increasing by 12.8 and Recall increasing by 15.1. YOLOv4's mAP has increased by 10.1, and its recall has increased by 9.2. The mAP of YOLOv4 SPP1 has increased by 10.4 and the recall has increased by 10.1. As a potential outcome, a network structure with an order to fill size of 864×864 is financially supported for refinement and elimination of redundant with completely separate hierarchical clustering rates.

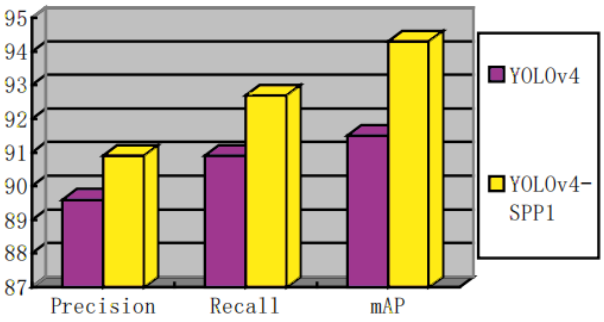

Fig.4 Evaluation indexes of YOLOv4 andYOLOv4-SPP1 (Table 4, Input size is 864).

Figure 4 shows a comparison of the evaluation indicators for YOLOv4 and Yolov4-SPP1. As can be highlighted, the feedback and discussion added to the spatial pyramid have an obvious influence, demonstrating

that that by incorporating the spatial pyramid, the feature extraction module period leading up to the YOLOv4 detection head may be efficaciously significantly improved. YOLOv4-SPP1 is therefore selected as the pruning network.

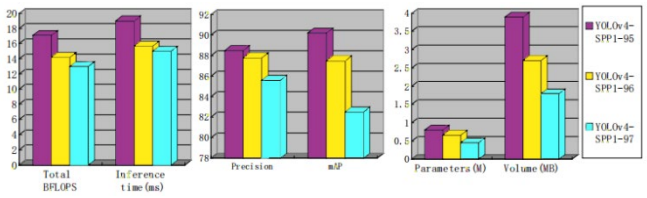

Fig.5 The evaluation indexes of YOLOv4 and YOLOv4-SPP1 trimmed by 95%, 96% and 97% respectively (Table 4, Input size is 864).

In the trials, pruning to various degrees was done on YOLOv4-SPP1. The graphic shows that when the model's pruning rate is raised to 96%, the best outcomes are attained. As shortly afterward as the hierarchical clustering rate surpasses 97%, the detection model's performance of the company begins to deteriorate. Figure 9 demonstrates and contrasts the independent review conclusive evidence of YOLOv4-SPP1-95, YOLOv4-SPP1-96, but rather YOLOv4-SPP1-97. YOLOv4-SPP1-96 is miraculously chosen as the most favorable hierarchical model TidyYOLOv4.

Analysis of detection effect: As shown in Table 2, under the identical network settings and experimental environment, TidyYOLOv4 has 272.07 fewer BFLOPS and Inference time than YOLOv4, 75.70ms fewer Inference time, 63.01M fewer parameters, and 252.9MB fewer Inference time than YOLOv4. Figures 6 and 7 indicate that there is no significant difference in detection performance between YOLOv4 and TidyYOLOv4. Matter of fact, the inference time for each frame is declined significantly by 75.70 ms, greatly reducing the amount of effort required for determining benchmarks and designed to allow for ever more time for localization and mapping and computation.

Fig.6 Visual detection effect of YOLOv4.

Fig.7 Visual detection effect of TidyYOLOv4.

The goal of implementing homogeneous deep neural networks is to observe how sliding changes in architecture affect the performance of the network. Is therefore evident from The above table 3 that as the parameters are adjusted, deep networks perform at a higher level for both EfficientNet and ResNet. Only certain experimental ResNet 50 results of the study are displayed in [6]. As can be seen from Table 2, ResNet-101 and ResNet-152 give better compared appears to result to ResNet-50. The same applies to EfficientNet. EfficientNetB7 achieves better results than other members of the family.

The model in this paper first finds the road area and based on this check if the human is walking through the area, so they check the human model of the human walking, and based on this decide the positioning of the human body Finally they give the final result about the human on the road, Figure 8 time comparison analysis of three baseline methods for this method and the current process. The analysis revealed that our methodology is significantly more effective and night before going to bed than that of other alternatives.

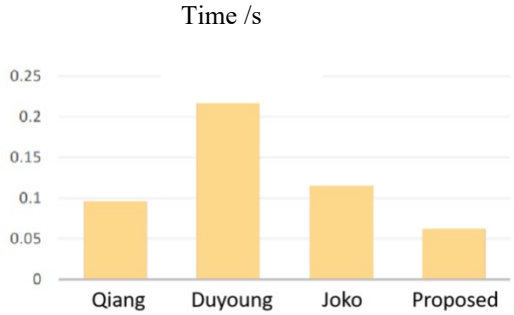

Figure 8. Comparative analysis

3. Results

The target identification algorithm TidyYOLOv4 is improved in this experiment to make it acceptable for unmanned urban traffic roadways. Furthermore, as demonstrated in Figure 1, adding SPP before the third detection head of YOLOv4 in this paper improves network feature extraction. Second, the redundancy of the YOLOv4-SPP1 training model is trimmed by a combined layer and channel pruning technique, resulting in a more potent detection model. The pruning strategy is implemented, sparse L1 regularization is applied to the channel scale factor, and the appropriate scale factor is adjusted to prune the unimportant the network model's component elements in order to significantly improve the object detector's profitability. This completes the process of automatically identifying non-essential components of the training model. The TidyYOLOv4 model is optimized using this method based on the YOLOv4 original model (the input picture size for the network configuration is 864864). TidyYOLOv4 features a model space in addition to having a quicker detection speed and more detection accuracy than YOLOv3. The results demonstrate that TidyYOLOv4 is better appropriate for autonomous vehicles to identify pedestrians in urban traffic conditions than YOLOv4 since the volume is decreased by 99.05% when compared to YOLOv4.

Network	Average Training Loss	Average Validation Loss	ADE
EfficientNet-B0	3.122	3.037	3.013
EfficientNet-B3	2.555	2.403	2.499
EfficientNet-B5	2.631	2.639	2.112
EfficientNet-B7	2.343	2.761	1.940
ResNet-34	6.437	1.960	2.830
ResNet-50	6.774	2.037	2.802
ResNet-101	7.559	2.181	2.307
ResNet-152	7.200	2.215	2.105

Acknowledgements

The research was supported by the National Fund Cultivation Project (No. CZ2021GP08) and Chizhou University Natural Key Project (No. CZ2021ZRZ10).

References

[1] Jin Q, Cui H, Sun C, et al.Domain adaptation based self-correction model for COVID-19 infection segmentation in CT images[J].Expert Systems with Applications,2021, 176.
[2] Li W, Raj A N J, Tjahjadi T, et al.Digital hair removal by deep learning for skin lesion segmentation[J].Pattern Recognition,2021, 117.
[3] Niehues S M, Adams L C, Gaudin R A, et al.Deep-Learning-Based Diagnosis of Bedside Chest X-ray in Intensive Care and Emergency Medicine[J].Investigative radiology,2021, 56 (8): 525-534.
[4] Owais M, Yoon H S, Mahmood T, et al.Light-weighted ensemble network with multilevel activation visualization for robust diagnosis of COVID19 pneumonia from large-scale chest radiographic database[J].Applied Soft Computing,2021, 108.
[5] Onan A, Tocoglu M a L P.A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification[J].Ieee Access,2021, 9: 7701-7722.
[6] Roh Y, Heo G, Whang S E.A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective[J].Ieee Transactions on Knowledge and Data Engineering,2021, 33 (4): 1328-1347.
[7] Wen S, Wei H, Yang Y, et al.Memristive LSTM Network for Sentiment Analysis[J].Ieee Transactions on Systems Man Cybernetics-Systems,2021, 51 (3): 1794-1804.
[8] Yang Z-L, Zhang S-Y, Hu Y-T, et al.VAE-Stega: Linguistic Steganography Based on Variational Auto-Encoder[J].Ieee Transactions on Information Forensics and Security,2021, 16: 880-895.
[9] Burnett K, Qian J, Du X, et al.Zeus: A system description of the two-time winner of the collegiate SAE autodrive competition[J].Journal of Field Robotics,2021, 38 (1): 139-166.
[10] Burnett K, Samavi S, Waslander S L, et al.aUToTrack: a lightweight object detection and tracking system for the SAE autodrive challenge arXiv[J].arXiv,2019: 8 pp.-8 pp.
[11] Samak T V, Samak C V, Ming X.AutoDRIVE Simulator: A Simulator for Scaled Autonomous Vehicle Research and Education arXiv[J].arXiv,2021: 8 pp.-8 pp.
[12] Wen J, Chen B, Tang W, et al.Harsh-Environmental-Resistant Triboelectric Nanogenerator and Its Applications in Autodrive Safety Warning[J].Advanced Energy Materials,2018, 8 (29).
[13] WangNa. Research on pedestrian detection algorithm and its security in unmanned driving [D],NanJing University.2020.
[14] Dai J, Li Y, He K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks,2016: arXiv:1605.06409.
[15] Girshick R, Donahue J, Darrell T, et al.: Rich feature hierarchies for accurate object detection and semantic segmentation,2014 Ieee Conference on Computer Vision and Pattern Recognition, New York: Ieee,2014: 580-587.
[16] Girshick R J a E-P. Fast R-CNN,2015: arXiv:1504.08083.
[17] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,2014: arXiv:1406.4729.
[18] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection,2015: arXiv:1506.02640.
[19] Redmon J, Farhadi A J a E-P. YOLO9000: Better, Faster, Stronger,2016: arXiv:1612.08242.
[20] Redmon J, Farhadi A J a E-P. YOLOv3: An Incremental Improvement,2018: arXiv:1804.02767.
[21] Bochkovskiy A, Wang C-Y, Liao H-Y M J a E-P. YOLOv4: Optimal Speed and Accuracy of Object Detection,2020: arXiv:2004.10934.