Pedestrian detection algorithm based on improved multi-scale feature fusion

Hui Liu 1,2, Keyang Cheng 1*,2

1 School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2 Research Center of New Telecommunication Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
*Corresponding author’s e-mail: s190131093@stu.cqupt.edu.cn

Abstract. Aiming at the problem of false detection and missed detection of small targets and occluded targets in the process of pedestrian detection, a pedestrian detection algorithm based on improved multi-scale feature fusion is proposed. First, for the YOLOv4 multi-scale feature fusion module PANet, which does not consider the interaction relationship between scales, PANet is improved to reduce the semantic gap between scales, and the attention mechanism is introduced to learn the importance of different layers to strengthen feature fusion; then, dilated convolution is introduced. Dilated convolution reduces the problem of information loss during the downsampling process; finally, the K-means clustering algorithm is used to redesign the anchor box and modify the loss function to detect a single category. The experimental results show that the improved pedestrian detection algorithm in the INRIA and WiderPerson data sets under different congestion conditions, the AP reaches 96.83% and 59.67%, respectively. Compared with the pedestrian detection results of the YOLOv4 model, the algorithm improves by 2.41% and 1.03%, respectively. The problem of false detection and missed detection of small targets and occlusion has been significantly improved.

1. Introduction

The application of target detection technology in the fields of national security, social security, and people's production and life has become more and more extensive, and it has also become one of the basic research topics. Pedestrian detection is an important task in target detection, and the research in the field of unmanned driving and intelligent video surveillance is getting deeper and deeper. In the usual sense, the target detection network mainly analyzes the geometric characteristics of the target to be detected, then locates the area of interest in the video or picture and can accurately determine the specific category, and finally gives each target a bounding box to determine its Location[1]. The pedestrian detection network has two main purposes: to realize the identification of pedestrians and complete the positioning of pedestrians. Traditional pedestrian detection methods focus on feature extraction and classifier classification to improve detection accuracy, mainly using artificially designed Histogram of Oriented Gradients[2], wavelet transform (Haar)[3], aggregation Feature extractors such as Aggregated Channel Features[4] are used to extract pedestrian features, and classifiers such as Support vector machines[5] and Adaptive Boosting[6] are used to determine whether the area has a goal. Among them, the representative detection methods: HOG+SVM, Haar+AdaBoost, etc. Traditional pedestrian detection algorithms separate feature extraction and
classification training into two independent processes, which are often limited by specific environmental conditions (such as small-scale transformation), set low-level features (such as texture features), and are different. The application of features and classifiers is different, resulting in insufficient feature expression ability, poor separability, and poor portability.

With the continuous development of deep learning, in the field of target detection, the RCNN two-stage target detection algorithm model represented by Faster-RCNN (Faster Region-based Convolutional Neural Networks)[7] and SSD (Single Shot Detection)[8]. The one-stage target detection algorithm model represented by the YOLO (You Only Look Once)[9] series has made unprecedented development in pedestrian detection tasks. Compared with traditional pedestrian detection algorithms, algorithms based on deep learning avoid the complicated process of feature extraction and data classification and reconstruction and can learn more expressive and richer high-level features. PURKAIT P et al.[10] constructed SPPNet by adding the SPP (spatial pyramid pooling) structure to the convolution end of RCNN, and the designed network is more adaptable to input images of any size. At the same time, to reduce the amount of calculation and avoid repeated extraction of features, Fast RCNN[11] unified the category classification task and the candidate frame regression task to design a multi-task loss function and generated fixed-scale features through ROI pooling. The process of generating candidate frames is the main cause of its performance degradation, and it takes about 2 seconds to process a picture. To solve the time-consuming problem of generating candidate frames, Faster-RCNN uses a region proposal network to replace the original selective search method to generate a proposal window, which improves the speed of two-stage target detection. The two-stage target detection algorithm needs to obtain the candidate proposal box first, and then classify and regress the candidate proposal box, which leads to the disadvantages of the two-stage detection algorithm such as slow convergence speed and longer reasoning time. The shortcomings of the second stage led to the generation of the first stage target detection algorithm. By directly predicting the detection frame through regression, the algorithm convergence speed can be accelerated and the reasoning time can be reduced. YOLOv1 designed by Redmon[12] et al. improves the detection efficiency of the target detection model to 45FPS. In YOLO9000[13], the use of the backbone feature extraction module Darknet19 increases the algorithm detection speed to 67FPS, and the algorithm detection accuracy can be increased from 63.4% (VOC data set) to 76.8% by using methods such as multi-scale training and anchor boxes. At the same time, Liu[14] and others designed the SSD one-stage detection model, which uses hierarchical features to detect. While taking into account the detection accuracy, the detection speed has also been greatly improved, with the detection accuracy and detection speed reaching 72.1% and 58FPS respectively. In the YOLOv3[15] model proposed later, the backbone feature extraction module Darknet53 has the advantages of stronger feature expression ability and faster reasoning speed. The detection accuracy on the MS COCO data set reaches 57.9%, and the detection speed reaches 20+FPS. And the latest YOLOv4 model is improved based on the YOLOv3 model, from the aspects of data processing, backbone feature extraction network, feature fusion, and loss function, etc. The improved method is to summarize almost all target detection techniques, and then filter. And permutation and combination, and finally conduct ablation experiments to select effective detection methods. Because the environment where the pedestrian target is located in a crowded area such as a street, a shopping mall, an airport, or a station, these environments often have mutual occlusion, so in the process of pedestrian detection, it is easy to cause small targets and occluded targets to be missed and wrongly detected problem. This paper analyzes and improves the universal target detection model YOLOv4. Because there is a certain semantic gap between different layers of the backbone feature extraction network, but the multi-scale feature fusion module PANet[18] of YOLOv4 does not consider the relationship between adjacent layers, so PCNet replaces PANet to reduce adjacent layers. In the process of feature fusion, down-sampling will lead to loss of feature information, and the introduction of hole convolution to reduce the problem of small targets and occluded target loss; for only detecting a single The goal is to use the k-means clustering algorithm to redesign and modify the loss function of the anchor box. Through the above modifications, this paper proposes an improved pedestrian detection algorithm. After conducting
comparative experiments, the improved pedestrian detection algorithm’s detection accuracy compared with the YOLOv4 model improves the detection accuracy of small targets and occluded targets, and reduces missed and false detections of pedestrian targets.

2. YOLOv4 model

YOLO is a convolutional neural network target detection algorithm based on a non-regional proposal. It converts the target detection problem into a bounding box regression problem. It directly uses the image as input and outputs the bounding box position information and category information of the target to be detected in the last layer. To achieve end-to-end detection. YOLOv4 is the fourth version of the YOLO series, and its overall network architecture is shown in Figure 1. Compared with the two-stage target detection model, YOLOv4 has lower algorithm complexity and faster detection speed; compared with the same-stage target detection algorithm SSD, YOLOv4 has higher detection accuracy and faster detection speed. Based on the original YOLOv3 target detection architecture, the YOLOv4 model has been improved in data processing, backbone feature extraction network, network training, activation function and loss function, etc, so that the detection speed and accuracy of the model reach the best match. The basic idea of the YOLOv4 model is first to scale the size of the input image to a size of 416×416, then divide it into cells of the same size as S×S according to the size of the feature map, and finally have three outputs of different scales, 13 respectively. 13×13, 26×26, 52×52 size, each cell predicts 3 candidate frames.

The backbone feature extraction network CSPDarknet53[15] is the core network of the YOLOv4 model, which is used to extract the feature information of the target to be detected. CSPDarknet53 draws on the CSPNet network to not only enhance the learning ability of CNN but also achieve lightweight capabilities. CSPDarknet53 adds CSP (Cross Stage Partial) to each large residual block of Darknet53 and divides the feature map of the base layer into two parts. And then merge through the cross-stage hierarchical structure to reduce the amount of calculation while ensuring accuracy. The activation function of CSPDarknet53 uses the Mish[16] activation function, and the network behind YOLOv4 uses the Leaky_relu activation function. Experiments show that this setting is more accurate in target detection.

YOLOv4 uses multi-scale prediction in the prediction part. The input image is divided into S×S cells. Each cell can predict 3 boxes. Each box is represented by a five-tuple, which represents the center point coordinates of the rectangular box, (w, h) represents the width and height of the rectangular frame, and c represents whether the current prediction iframe contains the detection object and its confidence. The output dimension of each prediction box is S×S×3×(4+1+C), where 4 represents the bounding box coordinates of the prediction target, C represents the number of object categories, and 1 represents the confidence of the prediction target. The loss function of YOLOv4 consists of three parts: boundary regression loss, confidence loss, and classification loss. The calculation formula is:

\[ \text{Loss} = \sum_{i} (E_{\text{coord}i} + E_{\text{conf}i} + E_{\text{class}}) \]  

In equation (1), \( E_{\text{coord}} \) represents the loss of the bounding regression box, \( E_{\text{conf}} \) represents the loss of confidence; \( E_{\text{class}} \) represents the loss of category.
2.1. Improved pedestrian detection algorithm

There is a difference between the universal target detection network YOLOv4 and the pedestrian detection network. The general target detection network needs to consider the differences between classes and scales. For example, the width of a car is greater than the height, and the height of a pedestrian is greater than the width, which results in the network's ability to handle a certain category is not outstanding. Therefore, it is necessary to redesign the target detection network for pedestrians, and make the following 4 improvements to the YOLOv4 target detection network:

1. Improvement of multi-scale feature fusion module. The YOLOv4 multi-scale feature fusion module uses SPP (Spatial Pyramid Pooling) [17] structure and PANet (Path Aggregation Network) [18] structure. Because the PANet structure does not consider the neighbor relationship between scales, the PCNet (Pyramid Convolutional Network) module is used to replace the PANet module, and the attention mechanism is introduced to learn the importance of different layers, strengthen feature fusion, and reduce the difference between different layers. Semantic gap.

2. The introduction of dilated convolution. Down-sampling can expand the receptive field, but it will cause small targets and occluded targets to be lost in this mapping process. Therefore, hole convolution is used in the down-sampling process to reduce the loss of feature information.

3. Reclustering of anchor box. The anchor box of YOLOv4 is obtained by learning the PASCAL VOC and COCO data sets. It is necessary to consider the differences between multiple categories and the differences in scales. The pedestrian detection network is no longer applicable, and a new anchor box needs to be redesigned.

4. Improvement of a loss function. The pedestrian detection network only needs to detect the pedestrian target. By simplifying the original loss function, it can reduce the output dimension and reduce the amount of calculation.

2.1. Improvement of multi-scale feature fusion module.

The down-sampling process in the backbone feature extraction network will lead to the expansion of the feature map receptive field of the deep network and the enhancement of the representation ability of semantic information, but it will reduce the resolution of the feature map and weaken the...
representation ability of geometric information (lack of spatial geometric feature details); the receptive field of the feature map of the shallow network is smaller, and the representation ability of geometric information is enhanced. Although the resolution of the feature map is increased, it will cause the semantic information representation ability to weaken, and finally cause the loss of feature information in the downsampling process. A multi-scale feature fusion module is an important means to solve this kind of problem. Many researchers directly scale and add feature maps to enhance feature fusion, thus reducing the semantic gap between feature maps of different layers, as shown in Figure 2. The FPN (Feature Pyramid Network) in the left picture of Figure 2 is top-down, passing down the strong semantic feature information of the deep feature map, and enhancing the entire multi-scale feature fusion module, but only the semantic information is enhanced, and there is no positioning information. Transfer: The PANet (Path Aggregation Network) in the right figure of Figure 2 adds a bottom-up structure after the FPN to transfer the strong positioning features of the shallow feature map, and enhance the transfer of feature information by repeatedly extracting the features. But none of them considered the intrinsic properties of scales well. To solve this problem, this paper improves the multi-scale feature fusion module PANet of YOLOv4 and introduces the attention mechanism into the feature fusion module to associate feature maps of similar layers and mine the interaction between scales.

The multi-scale feature fusion module of YOLOv4 is composed of SPP structure and PANet structure. The SPP structure can greatly increase the receptive field and isolate the most significant context features; PANet enhances the transfer of feature information by repeatedly extracting features.

The SPP structure is connected to the last convolutional layer of the CSPDarknet53 module. After performing three convolutions on the last feature layer of CSPDarknet53, four layers of different scales of maximum pooling are used for processing, and the maximum pooling core size is $1 \times 1$ ($1 \times 1$ means no processing), $5 \times 5$, $9 \times 9$, $13 \times 13$.

This article is to modify the PANet structure in YOLOv4 and replace PANet with PCNet. As shown in Figure 3, PCNet is a three-dimensional convolution across scales and spatial dimensions. There is a scale mismatch between different feature layers. The size decreases with the rise of the feature layer, using 3-D convolution to correlate feature maps of similar feature layers, across scales and spatial dimensions, and mining the interaction between scales. PCNet can represent 3 different 2-D convolution kernels, and the PCNet formula is shown in equation (2). On the bottom layer ($l=1$), the last term of equation (2) is unnecessary. At the top level ($l=L$), the first term of equation (2) can be ignored. When fusing different input features, most previous studies just add the features without distinction, as shown in Figure 2. However, because different input feature maps have different resolutions, their contribution to the fused output features is often Unequal, to solve this problem, a simple attention mechanism is introduced for weighted feature fusion, and learnable weights are introduced to learn the importance of different input features. Before performing weighted feature fusion, it is necessary to adjust the output features of different layers to the same size and the same number of channels, so it is necessary to up-sampling or down-sampling the features of different layers and adjust the number of channels. For the up-sampling layer, the high-level features need to be adjusted to the same size as the bottom-level features, using a convolution kernel with a stride of 0.5; for the down-sampling layer, a hole convolution with a stride of 2 [19] is used for downsampling.

$$y^l = \alpha * w_i * s_{0.5} x^l + \beta * w_0 * s_i x^l + \gamma * w_2 * s_2 x^l$$

In equation (2), $l$ represents the pyramid feature layer; $w_i, w_0, w_2$ represent three independent 2-D convolution kernels, and $x$ represents the input feature map. * $s_2$ represents a convolution kernel with stride of 2. * $s_{0.5}$ represents a convolution kernel with a stride of 0.5, which is composed of a normal convolution with a stride of 1 and a continuous bilinear interpolation up-sampling layer. $\alpha, \beta, \gamma$ is the weight parameter, the calculation method is as shown in equation (3)(4)(5):
In equations (3)(4)(5), $\varphi \geq 0$ is achieved by using the Relu function after each $\varphi$, $\varepsilon = 0.0001$, which is a small value to avoid numerical instability.

2.2. The introduction of dilated convolution

Downsampling can effectively increase the receptive field of the high-level network [20], enhance the ability in the semantic information table, and enable a single pixel to have a wide range of feature information, but the feature information of small targets and occluded targets will gradually be in this mapping process disappear. Therefore, it is proposed to use dilated convolution in the downsampling process to solve the problem of feature information loss. Compared with ordinary convolution, the hole convolution has an extra parameter with a dilated rate of $r$, which can increase the receptive field while maintaining the image resolution and without adding additional calculations. Since targets of different scales correspond to different sizes of receptive fields, it is proposed that three kinds of hole convolutions with different hole rates are paralleled for down-sampling so that after convolution, different ranges and sizes of information around the target can be obtained, and information loss is reduced. Then perform feature fusion on the down-sampled feature maps of each branch. The specific operation is shown in Dilated Convocation in Figure 3. The three convolution kernels are $3\times3$ and the stride is 2 dilated convolution for down-sampling, and the hole ratios are 1, 2, and 3 respectively, and the input feature map is processed at the same time. And then use the concat method to fuse the convolved feature maps, and finally use $1\times1$ convolution for dimensionality reduction processing, so that the feature layer used for prediction has the same number of channels as the input feature map of the module. The formula of the processing process is as equation (6):

$$ F = D_{3,1} + D_{3,2} + D_{3,3} $$

Equation (6), $D_{k,r}(x)$ represents the dilated convolution, $k$ represents the size of the convolution kernel, $r$ represents the size of the convolution, and $F$ represents the fused feature.

2.3. Reclustering of the Anchor box

The anchor box size in the original YOLOv4 algorithm is obtained by training the COCO data set and the PASCAL VOC data set clustering. There are 80 types of targets in the COCO data set, and 20
types of targets in the PASCAL VOC data set. These targets have different sizes, so the anchor boxes obtained by clustering have different shapes. The improved target detection network mainly detects pedestrians. The shape of most anchor boxes should be thin and tall, that is, the height of the anchor box is greater than the width. To make the improved target detection network predict the position of the target more accurately, the K-means algorithm is used to re-cluster the pedestrian data set to obtain a more accurate and representative anchor box.

Algorithm 1 K-means clustering algorithm:

Input: the number of cluster center points N, the data points in the data set form a set S
Output: cluster anchor

process:
- Randomly select N points from the set S as the initial center
- Calculate the respective distances between each data point in S and the N centers, and assign them to the nearest cluster according to the principle of minimum distance
- Recalculate the center of each cluster
- Until the cluster center no longer changes
- End, get N clusters

In the K-means algorithm, the Euclidean distance is usually used as the calculated measurement distance, but in the target detection algorithm, the area overlap IoU(T, P) of the prediction frame, and the real frame is used as the measurement distance and the new metric calculation formula for:

\[ d(T, P) = 1 - IoU(T, P) \] (7)

In equation (7), T represents the set of true bounding boxes of the object, and P represents the set of cluster center boxes. Clustering the crowded pedestrian dataset WiderPerson, 9 groups of Anchors are obtained as (3,7, 5,16, 8,31, 13,51, 19,23, 21,77, 21,77, 32,129, 53,208, 101,341).

2.4. Improvement of loss function

The improved target detection network only needs to detect the single target of pedestrians, so there is no need to predict the probability of the category in the network prediction process. It only needs to predict the confidence and the bounding box information of the pedestrian, which can effectively reduce the output dimension and use a five-tuple. \( P(x, y, w, h, c) \) represents the final output, in which the coordinates of the center point of the rectangular box are represented by \( (x, y) \), the width and height of the rectangular box are represented by \( (w, h) \), the foreground or background The probability is denoted by c. The loss function of the improved pedestrian detection network is composed of two parts: the bounding regression box loss and the confidence loss, namely.

\[ \text{Loss} = \sum_i (E_{\text{coord}_i} + E_{\text{conf}_i}) \] (8)

In equation (8), \( E_{\text{coord}_i} \) represents the loss of the bounding regression box, \( E_{\text{conf}_i} \) and represents the loss of confidence.

3. Experimental results and analysis

3.1. Experimental environment and data set

To verify the effectiveness of the improved pedestrian detection algorithm under different congestion conditions, this paper uses the pedestrian detection data sets INRIA and WiderPerson [21] to conduct experimental comparative analysis. The INRIA training set has 614 pictures, including 1,237 pedestrians, and each picture has about 2.01 pedestrians. The WiderPerson data set has a total of 13,382 pictures with about 400,000 annotations, and each picture has about 29.9 pedestrians.

3.2. Experimental procedure

The experiment in this article is based on the Pytorch-1.5.0 framework, the programming language is Python 3.6, and CUDA 10.2 and cuDNN 7.6.5 are installed at the same time to support the use of GPU. The algorithm training adopts the stochastic gradient descent optimizer algorithm for optimization
iteration, in which the algorithm parameter Batch Size is set to 8, and the initial learning rate is set to.
In the matching process of the prior frame and the real frame, the IoU value is set to 0.5, and the value
greater than 0.5 is regarded as a positive sample, and vice versa.

3.3. Experimental results and analysis
After the network training is completed, it is tested using the INRIA and WiderPerson test data sets.
The algorithm performance evaluation indicators used in the experiment are Precision (accuracy),
Recall (recall rate), and AP (Average Precision).

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{9}
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \tag{10}
\]

In equations (9) and (10), TP (True Positive) represents the total number of samples in the test set
that can be correctly classified as positive samples, FP (False Positive) represents the total number of
samples in the test set that can be incorrectly classified as positive samples, FN (False Negative)
indicates the total number of samples in the test set that can be mistakenly classified as negative
samples. Precision represents the proportion of target frames that can be correctly predicted in all
prediction frames. The recall represents the proportion of target frames that can be correctly predicted
in all real frames. AP represents the detection accuracy rate of a certain type, which is obtained by
finding the highest accuracy under different recall rates. The larger the value of AP reflects the better
detection of the improved model.

3.3.1 INRIA data set
In the INRIA data set, each picture contains about 2.01 pedestrians, verifying the effectiveness of the
proposed innovation under non-crowded conditions. The detection results obtained by using the
original YOLOv4 model and the model in this paper on the INRIA data set are shown in Table 1. The
AP of YOLOv4 is 94.42% and the average detection time is 0.039s. The improved pedestrian
detection method has an AP of 96.83% and the average detection time is 0.038s. Compared with
YOLOv4, the method proposed in this paper improves the detection accuracy by 2.41% AP when the
detection speed is not much different.

| Method                        | Data Set | AP/%  | Average detection time/s |
|-------------------------------|----------|-------|--------------------------|
| YOLOv4                        | INRIA    | 94.42 | 0.039                    |
| Improved pedestrian           | INRIA    | 96.83 | 0.038                    |

Figure 4 shows the AP comparison between YOLOv4 and the improved pedestrian detection
network in the number of training iterations in the INRIA data set. In the first 1200 iterations,
YOLOv4 consistently surpassed the improved pedestrian detection network on AP; after 1200
iterations, improved pedestrians The detection network always exceeds YOLOv4 on the AP; after
4800 iterations, both methods gradually converged.

![Figure 4. AP comparison under each number of training iterations in INRIA](image-url)
3.3.2 WiderPerson data set

In the WiderPerson data set, each picture contains approximately 29.9 pedestrians, verifying the effectiveness of the proposed innovation under more crowded conditions. The detection results obtained by using the original YOLOv4 model and the model in this paper to iterate on the WidePerson dataset are shown in Table 2. The AP of YOLOv4 is 58.64% and the average detection time is 0.040s. The improved pedestrian detection method has an AP of 59.67% and the average detection time is 0.042s. Compared with YOLOv4, the method proposed in this paper improves the AP by 1.03%, but the average detection time is increased by 0.002s.

| Method                        | Data Set      | AP/%  | Average detection time/s |
|-------------------------------|---------------|-------|--------------------------|
| YOLOv4                        | WiderPerson   | 58.64 | 0.040                    |
| Improved pedestrian detection method | WiderPerson   | 59.67 | 0.042                    |

Figure 5 shows the AP comparison between YOLOv4 and the improved pedestrian detection network in the WidePerson data set for each number of training iterations. In the first 50,000 iterations, YOLOv4 consistently surpassed the improved pedestrian detection network on AP; after 50,000 iterations, improved pedestrians The detection network always exceeds YOLOv4 on the AP; YOLOv4 gradually converges after 60,000 iterations, and the improved pedestrian detection network gradually converges after 90,000 iterations.

Figure 5. AP comparison under each number of training iterations in WiderPerson

4. Conclusion

This paper considers the relationship between scales to reduce the semantic gap between different layers, introduces the attention mechanism to learn the importance of different layers, introduces dilated convolution to reduce the problem of information loss in the downsampling process of small targets and occluded targets, and uses K-means The clustering algorithm redesigned the anchor box and modified the function to reduce the amount of calculation to improve YOLOv4, and proposed an improved pedestrian detection algorithm. The experimental results show that in the INRIA and WiderPerson data sets under different congestion conditions, compared with the YOLOv4 model, the modified human detection method proposed in this paper can increase the AP by 2.41% and 1.03%, respectively, while ensuring the detection speed, which can effectively improve Pedestrian detection accuracy. However, the real-time performance of the proposed method is not high. How to further simplify the network structure, further improve the detection real-time performance, improve the detection accuracy and reduce the false detection and missed detection is the next research direction.
References

[1] Fan Lili, Zhao Hongwei, Zhao Haoyu, Hu Huangshui, Wang Zhen. Survey of target detection based on deep convolutional neural networks [J]. Optical Precision Engineering, 2020, 28(05): 1152-1164.

[2] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE, 2005, 1: 886-893.

[3] Viola Paul, Jones M J. Robust real-time face detection [J]. Journal of Computer Vision, 2004, 57(2): 137-154.

[4] Dollar Piotr, Wojek Christian, Schiele Bernt, et al. Pedestrian detection: an evaluation of the state of art[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34: 743-61(10.1109/TPAMI.2011.155).

[5] Cortes, C. and Vapnik, V. Support-vector networks. Machine learning, 20(3), 1995, pp.273-297.

[6] Y Freund, R E Schapire. Adecision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Science, 1997, 55(1): 119-139.

[7] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]. neural information processing systems, 2015, 91-99.

[8] LIU W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]// Proceedings of European Conference on Computer Vision, 2016, 21-37.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of CVPR, 2015, 779-788.

[10] PURKAIT P, ZHAO C, ZACH C. SPP-Net: Deep Absolute Pose Regression with Synthetic Views[C]// Proc of the 28th British Machine Visio Conference. Imperial College London, United Kingdom: BMVC, 2017: 712-722.

[11] LI J, LIANG X, SHEN S M, et al. Scale-aware fast R-CNN for pedestrian detection[J]. IEEE Transactions on Multimedia, 2018, 20(4): 985-996.

[12] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// Proc of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 779-788.

[13] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii: IEEE, 2017: 7263-7271.

[14] REDMON J, FARHADI A. YOLOV3: An incremental improvement [EB/OL]. (2018). https://pjreddie.com/darknet.

[15] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A new backbone that can enhance learning capability of cnn. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop(CVPR Workshop), 2020.

[16] Misra D. Mish: A Self Regularized Non-Monotonic Neural Activation Function[J]. 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence( TPAMI), 37(9): 1904-1916, 2015.

[18] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pages 8759-8768, 2018.

[19] Li Y, Chen Y, Wang N, et al. Scale-Aware Trident Networks for Object Detection[J]. arXiv: Computer vision and pattern Recognition, 2019.

[20] F Yu, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. ICLR, 2016.

[21] Zhang S, Xie Y, Wan J, et al. WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild[J]. IEEE Transactions on Multimedia, 2019.