Improved Pedestrian Detection Algorithm of Yolov4 Network Structure

Xiujun Zhu
School of Information Science & Engineering, Yunnan University, Kunming, China
12019100757@mail.ynu.edu.cn

Yujie Bai
School of Information Science & Engineering, Yunnan University, Kunming, China
baiyujie@mail.ynu.edu.cn

Yijian Pei∗
School of Information Science & Engineering, Yunnan University, Kunming, China
yndxpyj@163.com

ABSTRACT

When the YOLOV4 network detects pedestrians alone, the small target pedestrians will be missed, resulting in the reduction of *P (Precision)* and *AP (Average Precision)* values. This paper improves the YOLOV4 network structure. In order to improve the feature extraction capability of the network for small targets, a shallower feature layer is added to the original three output feature layers of the YOLOV4 backbone network to build PANet (Path Aggregation Network) together. And two SPP (Spatial Pyramid Pooling) structures are added to expand the receptive field. The channel attention mechanism module is added and some convolutional layers of the original network are deleted. Finally, transfer learning is used to make the detection effect better. The *P* value of the pedestrian on the PASCAL VOC data set increased from 84.43% to 91.37%, and the *AP* value increased from 74.78% to 87.39%, and the *P* value on the commonly used pedestrian detection data set INRIA (INRIA Person Dataset) increased from 93.20% increased to 98.02%, *AP* value increased from 91.08% to 94.02%. Experimental results show that the network has a better effect on pedestrian detection, and the accuracy and average precision are improved.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision problems**; • **Object detection**;

KEYWORDS

YOLOV4, Pedestrian detection, PANet, Receptive field

ACM Reference Format:
Xiujun Zhu, Yujie Bai, and Yijian Pei∗. 2021. Improved Pedestrian Detection Algorithm of Yolov4 Network Structure. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487088

1 INTRODUCTION

As an important branch of target detection, pedestrian detection is a hot spot in the field of computer vision. Pedestrian detection tasks are usually implemented through traditional machine learning methods and deep learning-based methods. Before deep learning emerged, pedestrian detection was achieved through traditional machine learning methods. In recent years, with the deepening of deep learning research, there are more and more researches on the use of deep learning methods to achieve pedestrian detection, and good results have been achieved. Pedestrian detection based on traditional machine learning methods is mainly realized by using feature extraction and classifiers. The extracted features mainly include the target's grayscale, contour, texture, color, gradient histogram and other information. Dalal et al. proposed the HOG (Histogram of Oriented Gradient) feature as a pedestrian feature description operator [1, 2], and combined with the SVM (Support Vector Machine) classifier to achieve excellent results in pedestrian detection. The accuracy rate reached 90% on the INRIA pedestrian database; Wu et al. proposed the Edgelet feature to describe the local contour direction of pedestrians [3]. This method can effectively deal with the problem of pedestrians blocking each other; the LBP (Local Binary Pattern) feature proposed by Ojala et al. It can be used to extract the texture information of pedestrians [4]. Although these traditional machine learning methods can achieve pedestrian detection, they still have some shortcomings and are susceptible to interference from the external environment, such as light, pedestrian size, posture, and density [5-7]. In recent years, researchers have used deep learning methods to do target detection better than traditional machine learning algorithms. The detection results are better than traditional machine learning algorithms. Girshick et al. proposed Region-Convolutional Neural Networks (R-CNN) [8], which applied the deep learning method to target detection for the first time, which greatly improved the accuracy of target detection; Then Girshick et al. made some improvements on the basis of R-CNN, and proposed the Fast R-CNN network and the Faster R-CNN network [9, 10]. Compared with the R-CNN network, they have faster speed and accuracy, but this series of algorithms are two-step algorithms, which are time-consuming and cannot achieve real-time performance. In response to this phenomenon, Redmon et al. proposed the end-to-end target detection algorithm YOLO (You Only Look Once) [11] series and the SSD (Single Shot Detecto) algorithm proposed by Wei Liu et al. The end-to-end target detection [12] algorithm is faster and more accurate than the two-stage target detection algorithm, and can achieve real-time performance.

The research of this paper is to improve the YOLOV4 algorithm to make it more suitable for pedestrian detection, and it has a good recognition effect for pedestrians with small targets. The algorithm adds SPP structure and channel attention mechanism modules to the YOLOV4 network structure [13-16], and adds a shallow feature

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487088

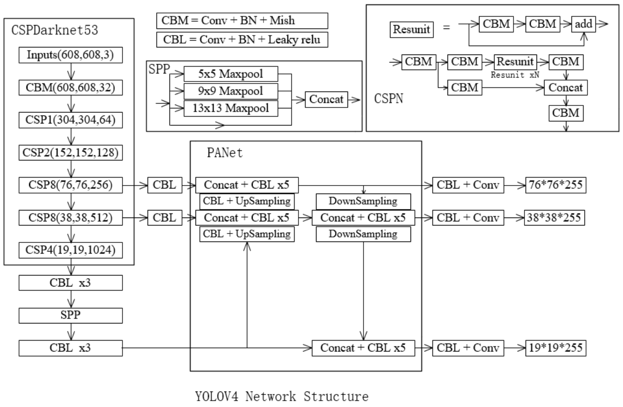

Figure 1: YOLOV4 Network Structure.

layer during feature fusion, and finally uses transfer learning to make the model training better [17, 18]. The SPP structure can increase the receptive field of the network. The larger the receptive field, the better the detection effect of the network; the channel attention mechanism module improves the representation ability of the network by modeling the dependence of each channel, and can perform features channel-by-channel adjustment, so that the network can learn to use global information to selectively enhance features containing useful information and suppress useless features, thereby improving the detection ability of the network; the information of small targets will decrease as the number of network layers deepens, Therefore, the introduction of shallow features can strengthen the detection ability of small targets; the use of transfer learning can make the feature extraction ability of the backbone network stronger.

Aiming at the phenomenon that YOLOV4 algorithm is prone to missed detection of small target pedestrians in pedestrian detection, the improved algorithm in this paper improves the detection ability of small target pedestrians, reduces pedestrian missed detection, and has better performance in the field of pedestrian detection. The improved algorithm in this paper can not only be used in the field of pedestrian detection, but also provide ideas to the improvement of target detection tasks in other fields.

2 YOLOV4 DETECTION ALGORITHM

In 2020, Redmon, the author of the YOLO series, issued a statement that out of ethical considerations, he has since retired from the CV world. In the same year, Alexey Bochkovskiy et al. obtained Redmon's consent and named their research YOLOV4. YOLOV4 is based on the traditional YOLO, combined with a large number of predecessor research techniques, added some very practical skills and carried out appropriate innovative algorithms, so as to achieve the best balance of detection speed and accuracy. The network of the YOLOV4 model is mainly shown in Figure 1 [19, 20]. Among them, CSPDarknet53, as the backbone network of the YOLOV4 model, has powerful feature extraction capabilities; the SPP module is used to increase the receptive field and separate important context features; the PANet module is used for feature fusion and extraction More effective features.

The YOLOV4 algorithm inputs a picture with a size of 608*608*3 into the CSPDarknet53 backbone network, and then passes through 5 large residual blocks CSPN. Each large residual block contains 1, 2, 8, 8, and 4 small residuals. Connect a feature output layer after the 3rd, 4th, and 5th large residual blocks respectively, named feat1, feat2, and feat3. The size of feat1 is 76*76*256, the size of feat2 is 38*38*512, and the size of feat3 is 19*19*1024. The SPP network divides the input features into 4 branches, one of which is not processed, and the other 3 are respectively subjected to the maximum pooling of the convolution kernel size of 5*5, 9*9, and 13*13, and then they are spliced together. The purpose of the SPP network is to increase the receptive field of the network and make the performance of the network better. The PANet network uses upsampling to splice and fuse deep and shallow features, and then splice and fuse shallow and deep features through downsampling. The PANet network shortens the distance between the shallow features and the deep features, Enriching the content of each feature layer.

The YOLOV4 network has three outputs with different scales, the sizes are 76*76, 38*38, 19*19. Since the deeper the network, the more information is lost in the picture, the 76*76 feature output layer is mainly used to detect small targets, and the 19*19 feature output layer is mainly used to detect large targets. Each feature layer divides the picture into S*S grids, and each grid is responsible for detecting the target whose center position falls into it. In the output network of YOLOV4, each grid is responsible for predicting the confidence of 3 bounding boxes, the 4 parameters of the prediction box, and the probability of C categories, so the size of each output feature layer of YOLOV4 is S*S* (3*(1+4+C)), in the COCO data set, there are a total of 80 categories, so the output size is S*S*255, and in the pedestrian detection task, there is only one category, so each feature output layer size is S*S*18.

3 IMPROVED YOLOV4 NETWORK STRUCTURE

3.1 Increase SPP Network

In the original YOLOV4 network, only the feat3 feature layer is followed by the SPP network, and the feat3 feature layer is mainly used to detect large targets. The detection capability of small targets has not been greatly enhanced, and missed detections are still prone to occur. In order to solve this problem, the improved algorithm introduces the SPP structure after the feature layer of feat1 and feature layer of feat2. The feature layers of feat1, feat2, and feat3 all go through the SPP network to expand the receptive field, and then enter the PANet network for path aggregation after enhancing the features.

3.2 Introducing Shallow Feature Layers

While the SPP structure enhances the network's receptive field, it will be accompanied by a reduction in image resolution, which will lose the information of small targets to a certain extent. In order to make the feature layer contain more information about small targets, the improved algorithm connects a feature output layer feat0 after the second residual block in the CSPDarknet53 backbone network, and the feature layer enters the PANet network after passing through a shallow residual network. Compared with

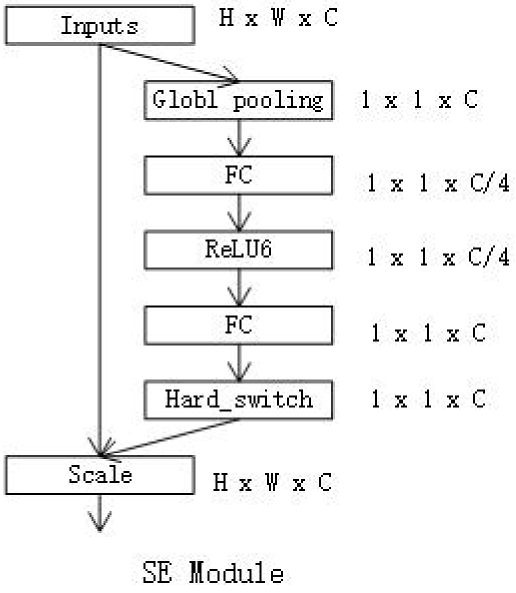

Figure 2: SE Module.

feat1, feat0 has the same size, feat0 is obtained from a shallower network and contains more small target information [21].

3.3 Channel Attention Mechanism

Each channel of the feature layer in the neural network contains different content of information, and the importance of these content is different, while the original network feature layer defaults to the same importance of each channel of the feature layer. Therefore, the improved algorithm joins the SE (Squeeze-and-Excitation) network after PANet, so that the output network focuses on channel information. The SE network model is shown in Figure 2. Assuming that the input is a h*w*c feature layer, first perform global average pooling on it to obtain a feature layer with a size of 1*1*c; Then enter the two fully connected layers, the number of neurons in the first fully connected layer is c/4, and the number of neurons in the second fully connected layer is c, and the weight of 1*1*c is obtained. The activation function followed by one fully connected layer is ReLU6, and the activation function followed by the second fully connected layer is Hard_switch; finally, the weight of 1*1*c is multiplied by the input feature layer to realize the channel attention mechanism.

3.4 Transfer Learning

Transfer learning is a machine learning method that uses the weight of the trained model of task A as the initial weight of task B. The model of task A is usually obtained from a large amount of data training, and task B is similar to task A but not exactly the same. At this time, transfer learning can be used in the training process of task B. Transfer learning can save training time, accelerate network convergence, and in most cases, it does not require too many training samples to get better performance.

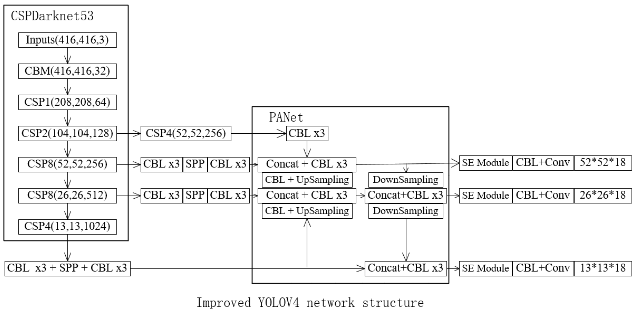

Figure 3: Improved YOLOV4 Network Structure.

3.5 Improved Network Structure

The improved network structure is shown in Figure 3. In addition to adding the above modules, the size of the input image is changed to 416*416*3, so the size of the three output feature layers is 52*52*18, 26*26* 18, 13*13*18.

4 EXPERIMENTAL RESULTS AND ANALYSIS

The experimental data set used all the pictures containing pedestrians in PASCAL VOC2007 and PASCAL VOC2012, a total of nearly 13,000 pedestrian pictures, of which 10,000 pictures were used as the training set, and the rest were used as the test set. In the training process, the mosaic data enhancement method is used, the batch size is set to 8, and the initial learning rate is set to 0.001. If the loss of 4 consecutive rounds of training does not decrease, then the learning rate is reduced to half; if the loss of 10 consecutive rounds of training does not decrease, then end the training and save the weights obtained. This paper compares three experiments. Experiment one does not change the YOLOV4 network structure, and does not use transfer learning, and directly trains from scratch; Experiment 2 does not change the network structure, uses transfer learning, and loads YOLOV4 weights trained with the COCO data set before training; Experiment 3 changed the network structure according to the method in this paper, and also used transfer learning during training. Because the network structure after the backbone network has changed, only the weight parameters of the backbone network CSPDarknet53 are used. There are 80 categories in the COCO data set, and the categories contain pedestrians, so using the weights trained in the COCO data set as pre-training weights can speed up the network convergence speed and improve the detection ability of the network.

Before training, we first use the K-means algorithm to perform cluster analysis on the labels of the training set, and obtain 9 a priori box parameters that have a better effect on pedestrian detection. The evaluation indicators used in the experiment are *P (Precision)*, *R(Recall)* and *AP(Average Precision)*. The *P* and the *R* are contradictory. Increasing the *P* may reduce the *R*, and increasing the *R* may reduce the *P*. The *AP* value is the area enclosed by the *P-R* curve and the coordinate axis. The larger the *AP* value, the better the detection performance of the network.

$$R = \frac{Number\ of\ pedestrians\ correctly\ detected}{Total\ number\ of\ pedestrians\ in\ test\ set} \tag{1}$$

Table 1: VOC Data Set Experimental Results

Experiment	Precision	Recall	AP
Experiment One	84.43%	54.31%	74.78%
Experiment Two	84.12%	80.14%	85.51%
Experiment Three	91.37%	64.46%	87.39%

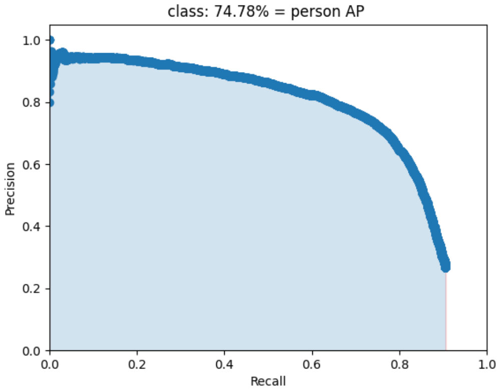

Figure 4: Experiment 1 P-R curve.

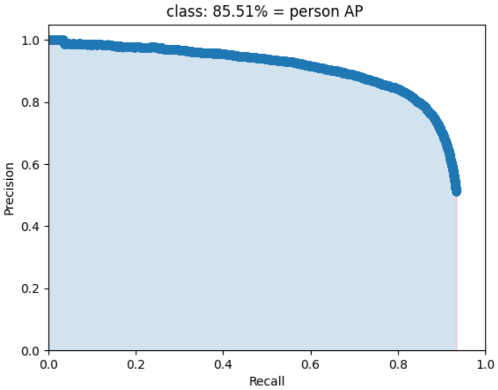

Figure 5: Experiment 1 P-R curve.

$$P = \frac{Number\ of\ pedestrians\ correctly\ detected}{Predicted\ total\ number\ of\ pedestrians} \tag{2}$$

4.1 VOC Test Set Test Results

In all the test experiments in this article, the threshold of IOU (Intersection over Union) is set to 0.5, and the threshold of confidence is set to 0.01. The models obtained from the three experiments are the test results of the VOC test set. As shown in Table 1. The compiled VOC test set pictures are about 3,000, with a total of 5861 pedestrians. Among these pedestrians, there are mutual occlusion, pedestrians of normal size and pedestrians with small targets, and the sharpness and light intensity of these pictures are not consistent, which can be represents pedestrians in various scenarios. It can be observed from the data in Table 1 that the recall rate of the network can be improved through transfer learning. The improved network structure in this paper improves the precision of the network, but reduces the recall rate of the network. And the *AP* value obtained in Experiment 3 is the highest, which means that the improved network detection performance in this paper is better than the previous network. The *P-R* curves of experiments 1, 2, and 3 are shown in Figures 4, 5, and 6, and the *AP* value is the green part of the figure.

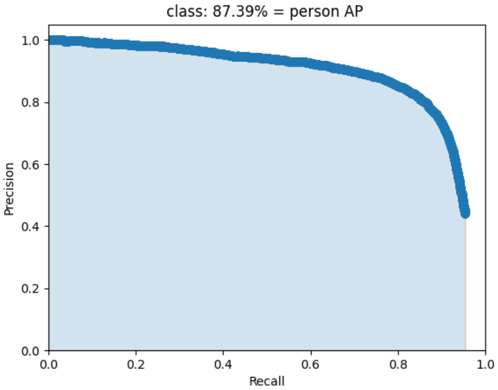

Figure 6: Experiment 1 P-R curve.

Table 2: INRIA Data Set Experimental Results

Experiment	Precision	Recall	AP
Experiment One	93.20%	75.71%	91.08%
Experiment Two	87.31%	85.26%	90.48%
Experiment Three	98.02%	66.33%	94.02%

4.2 INRIA Data Set Test Results

In addition to comparing the effects of the three experiments on the VOC data set, this article also compares the effects of these three experiments on the INRIA data set. The INRIA data set is currently the most widely used static pedestrian data set. The data set has a complex background, with large changes in pedestrian posture and mutual occlusion between pedestrians. The test set has a total of 288 pictures and 597 pedestrians. The experimental results are shown in Table 2. It can be observed from Table 2 that the improved network *AP* value has increased from 91.08% to 94.02%, and the *P* value is as high as 98.02%. Therefore, the improved network performance is

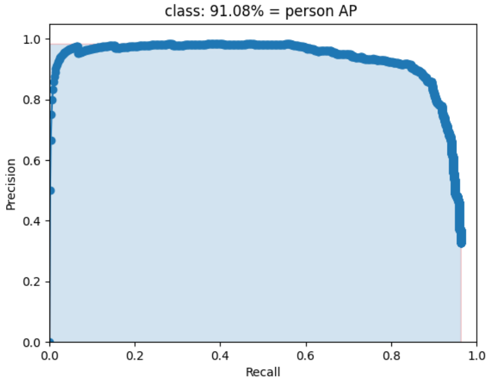

Figure 7: Experiment 1 P-R curve.

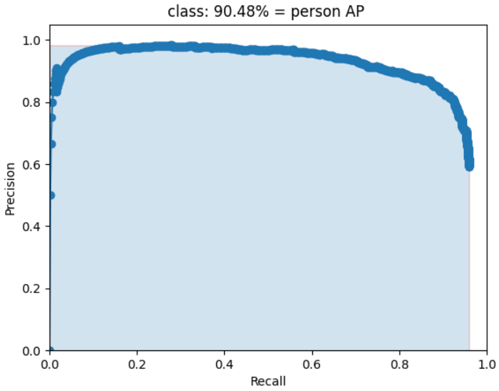

Figure 8: Experiment 1 P-R Curve.

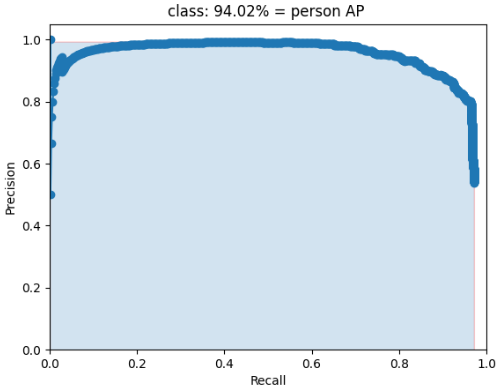

Figure 9: Experiment 1 P-R Curve.

significantly better than the previous network. The *P-R* curves of experiments 1, 2, and 3 are shown in Figures 7, 8, and 9, and the *AP* value is the green part of the figure.

5 CONCLUSIONS

The pedestrian detection algorithm proposed in this paper to improve the YOLOV4 network structure reduces the missed detection rate of small target pedestrians in pedestrian detection, thereby improving the precision and average precision of detection. The improved network introduces a shallow feature in the CSPDarknet53 backbone network for path aggregation, and then adds two SPP structures to expand the network's receptive field, and finally introduces a channel attention mechanism to make the network pay attention to channel information. The *P* value of the improved network is particularly improved, from 84.43% to 91.37% in the VOC test set, and from 93.20% to 98.02% in the INRIA test set, indicating that there are very few missed detections in the detection process, and after the improvement the *AP* value of the network is also improved, and the detection performance is improved compared with the original network.

The improved algorithm in this paper improves the precision of pedestrian detection, but there are still some shortcomings. In model training, only factors such as light intensity and scale are considered, and the influence of some extreme weather, such as rain and fog, is not considered. Since there are no photos in extreme weather in the training set, the precision of pedestrian detection in these weathers may be reduced. This is an improvement direction for pedestrian detection algorithms in the future. In addition, the improved algorithm does not reduce the size of the model and is not suitable for use on mobile devices. In order to increase the practicality of the pedestrian detection algorithm in reality, future detection model algorithms will reduce the size of the model as much as possible.

REFERENCES

[1] N. Dalal, (2005). Histograms of oriented gradients for human detection. Proc of Cvpr.
[2] M. Nan, C. Li, , Jiancheng, H. , Qiuna, S and Guoping, Z. . (2019). Pedestrian Detection Based on HOG Features and SVM Realizes Vehicle-Human-Environment Interaction. 2019 15th International Conference on Computational Intelligence and Security (CIS).
[3] Wu, B. and Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. Tenth IEEE International Conference on Computer Vision. IEEE.
[4] Ojala, T. , Matti Pietikäinen and Topi Mäenpää. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Classification with Local Binary Patterns.
[5] Girshick, R. , Donahue, J. , Darrell, T. , & Malik, J. . (2013). Rich feature hierarchies for accurate object detection and semantic segmentation.
[6] Luo, H. W. , Zhang, C. S. , Pan, F. C. , & Ju, X. M. . (2020). Contextual-YOLOV3: Implement Better Small Object Detection Based Deep Learning. 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE.
[7] Lai, Y. , Sun, F. , & Liu, H. . (2020). Small Object Detection Base on YOLOv3 For Pedestrian Recognition. 2020 5th International Conference on Control and Robotics Engineering (ICCRE). IEEE.
[8] Nguyen, N. D. , Do, T. , Ngo, T. D. , & Le, D. D. . (2020). An evaluation of deep learning methods for small object detection. Journal of Electrical and Computer Engineering, 2020, 1-18.
[9] Girshick, R. . (2015). Fast r-cnn. Computer ence.
[10] Ren, S. , He, K. , Girshick, R. , & Sun, J. . (2017). Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(6), 1137-1149.
[11] Redmon, J. , Divvala, S. , Girshick, R. , & Farhadi, A. . (2016). You Only Look Once: Unified, Real-Time Object Detection. Computer Vision & Pattern Recognition. IEEE.
[12] Liu, W. , Anguelov, D. , Erhan, D. , Szegedy, C. , & Berg, A. C. . (2016). SSD: Single Shot MultiBox Detector. European Conference on Computer Vision. Springer International Publishing

[13] Bochkovskiy, Alexey & Wang, Chien-Yao & Liao, Hong-yuan. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. Proc of Cvpr.

[14] Zhu, Y. , Yang, J. , Deng, X. , Xiao, C. , & An, W. . (2020). Infrared pedestrian detection based on attention mechanism. Journal of Physics: Conference Series, 1634(1), 012032 (6pp).

[15] Yang, X. , Wang, Y. , & Laganiere, R. . (2020). A scale-aware YOLO model for pedestrian detection. International Symposium on Visual Computing (ISVC) 2020.

[16] Huang, Z. , Wang, J. , Fu, X. , Yu, T. , & Wang, R. . (2020). Dc-spp-yolo: dense connection and spatial pyramid pooling based yolo for object detection. Information Sciences, 522.

[17] Li, J. , Liao, S. , Jiang, H. , & Shao, L. . (2020). Box Guided Convolution for Pedestrian Detection. MM '20: The 28th ACM International Conference on Multimedia. ACM.

[18] Guo, Wei; Li, Weihong; Gong, Weiguo; Cui, Jinkai (2020). Extended Feature Pyramid Network with Adaptive Scale Training Strategy and Anchors for Object Detection in Aerial Images. Remote Sensing, 12(5), 784–.

[19] Deng, C., Wang, M., Liu, L., & Liu, Y. J. a. e.-p. (2020). Extended Feature Pyramid Network for Small Object Detection. arXiv:2003.07021.

[20] LIU L, ZHENG Y, FU D M.(2020). Occluded Pedestrian Detection Algorithm Based on Improved YOLOv3.Pattern Recognition and Artificial Intelligence, 2020, 33(6) : 568−574.

[21] YANG Yaru, DENG Hongxia, WANG Zhe, *et al*(2020). Deep network pedestrian detection guided by shallow feature fusion.Computer Engineering and Applications, 2020, 56(2): 196-200.