ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

Lightweight Pedestrian Detection Algorithm Combined with Data Enhancement

Lifen Li[1,*], Minghe Ma[2]

[1]School of Computer Science, North China Electric Power University (Baoding), Baoding, Hebei, 071000, China

[2]School of Computer Science, North China Electric Power University (Baoding), Baoding, Hebei, 071000, China

Abstract

YOLOv3 has become a commonly used target detection algorithm in the industrial field due to its fast detection speed and high detection accuracy. But the disadvantage is that the network model is too large to be easily deployed on small terminals. In order to solve the problem of YOLOv3 big network model and further improve its detection speed and detection accuracy, a lightweight pedestrian detection algorithm combined with data enhancement is proposed. Through using bilinear interpolation, Mosaic data enhancement and other methods to optimize the pedestrian image, and use the lightweight network MobileNetV3 to replace the backbone network of YOLOv3, and select the Kmeans++ algorithm to replace the original Kmeans algorithm used to obtain the anchor boxes. Experiment on the current mainstream pedestrian data set. Experimental results show that the size of the improved algorithm model is 1/20 of the original YOLOV3 algorithm model, while the detection speed is increased to 91FPS, and the detection accuracy has also been improved to a certain extent.

Keywords

Pedestrian detection; Deep learning, YOLOv3; MobileNetV3; Mosaic data enhancement.

1. INTRODUCTION

At present, deep learning shines brightly in the field of target detection, and various excellent algorithm models emerge one after another [1]. For example, the famous RCNN[2][3][4] series of algorithms, YOLO[5][6][7] series of algorithms, SSD[8] series of algorithms and so on. Compared with traditional target detection algorithms, these algorithms have been greatly improved in detection accuracy and speed. This also promotes deep learning to become one of the main research methods in the field of computer vision.

As deep learning technology continues to enter all walks of life in society. Autonomous driving, robotics and other fields have attracted more and more attention and have become a hot research direction. Among them, pedestrian detection technology, as an important part of the above fields, has attracted great attention of scholars. Pedestrian detection algorithms are mainly divided into two categories, namely traditional pedestrian detection algorithms and pedestrian detection algorithms based on deep learning. The traditional pedestrian detection algorithm usually uses an exhaustive strategy, that is, a sliding window method is used, that is, a window is selected, and the region of interest in the image is selected at one time by sliding the window on the image, and then HOG[9], SIFT[10] and other methods to extract the features, and finally use the classifier to classify the pedestrian target, and the representative algorithm

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

is HOG+SVM[9] and so on. However, the traditional target detection algorithm has the disadvantages of low accuracy, slow speed and large amount of parameter calculation.

In recent years, lightweight neural networks have gradually become a popular research direction in the field of deep learning due to their small model size and easy portability and deployment. Although the detection accuracy and speed of the YOLO series network are relatively good, the scale of the YOLO series network model is large, and the requirements for the equipment to be transplanted and deployed are also relatively high. In order to ensure the detection accuracy and speed, the YOLO model can be easily and effectively deployed on small terminal devices, this paper proposes a target detection algorithm that integrates the lightweight neural network MobileNetV3[11] and YOLOv3[7] algorithms. , and use data augmentation to improve the generalization ability of the model. Compared with the original YOLOv3 network model, the improved YOLOv3 has less parameter calculation, and has a certain improvement in detection speed and accuracy.

2. METHODOLOGY

2.1. YOLOv3 Network Model

Along with the movement of the target, the sink node timely notifies the sensor nodes in the relevant detection area to join in the process of target tracking. Figure 1 is the flow chart of the moving target tracking process.

After the YOLO and YOLOv2 network models were proposed and achieved excellent results in the field of target detection, in 2018, the YOLOv3 network model was born. Compared with the previous generation YOLOv2 network model, YOLOv3 has upgraded its main network structure from Darknet19 to Darknet53. Darknet53 includes 53 convolutional layers, which can extract richer feature information than the previous generation network model. At the same time, it absorbs the idea of residual structure proposed in ResNet [12], and introduces shortcut connection, which effectively solves the problems of difficulty in training, gradient explosion and degradation in deep neural networks. At the same time, YOLOv3 draws on the FPN [13] network structure to fuse low-level semantic features and high-level semantic features to obtain better feature maps. Compared with the previous generation network, the detection speed and accuracy have been significantly improved. The network structure of YOLOv3 is shown in Figure 1.

At the same time, the concept of Anchor Box proposed in Faster RCNN [4] is introduced into the YOLOv3 network. The initial Anchor Box is generated by Kmeans clustering algorithm. In the Darknet53 network, the pooling layer of the previous generation network is abandoned, and the convolution step size is set to 2 to realize the downsampling operation of the picture. After the picture enters Darknet53, after several Darknet convolution modules, a target frame of 13×13 is obtained, which is used to detect large targets. After that, upsampling and feature fusion are used to generate 26×26 and 52×52 respectively. The target frame of the specification is used to detect medium and small targets. The final result contains the coordinate information, category information and confidence score of the target, and finally the final target detection result is obtained through the non-maximum value suppression algorithm.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

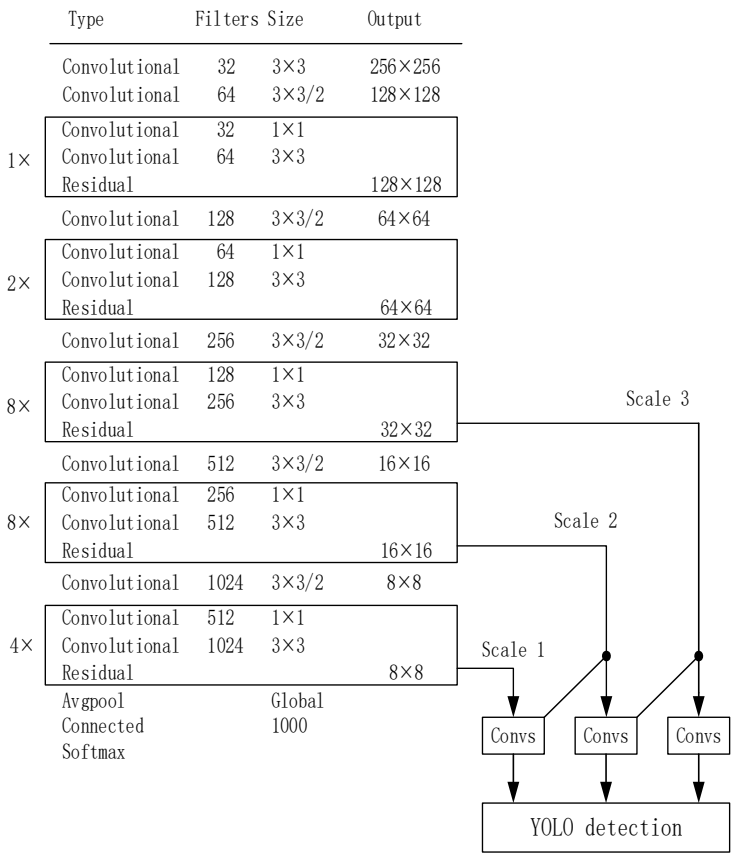

Figure 1. YOLOv3 network structure

2.2. MobileNetV3 Network Model

MobileNetV3 is a lightweight convolutional neural network improved from the previous two generations MobileNetV1 [14] and MobileNetV2 [15]. In MobileNetV1, the main innovation is the use of depthwise separable convolution. The depthwise separable convolution essentially divides the standard convolution into two steps, namely depthwise convolution and pointwise convolution. The depth convolution selects a separate filter for each input channel of the image to perform the convolution operation, and obtains an output feature map that is consistent with the number of channels in the input feature map. Fusion to get the final result. In terms of parameter quantity, using the parameters of 3×3 depthwise separable convolution can reduce the parameters to 1/9 of the standard convolution.

However, due to the use of the ReLU activation function in MobileNetV1, the loss of image feature information will be caused during training, and when the number of image channels is less, the loss of image information will be more serious. In order to solve the above problems, MobileNetV2 mainly made two improvements while retaining the depthwise separable convolution in the V1 version. The first point is to adopt a spindle-like structure. First, the 1×1 convolution layer is used to increase the number of channels of the feature map, and then a 3×3 depth separable convolution kernel is used to perform convolution and extract features. Finally, the convolution of 1×1 is used to restore the number of channels of the feature map, so that the image channels are first expanded and then compressed. The second point is to remove the ReLU function used in the original V1 version and use linear output. This can effectively ensure the diversity of features, and also improve the expressiveness of the network.

In order to further reduce the computational delay and overhead of the model, Google proposed the MobileNetV3 network after a year. Compared with the previous generation V2, there are two major innovations. The first innovation is the use of two automated learning techniques in the network, namely, automatic mobile neural architecture search (MnasNet) [16]

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

and platform-aware algorithms for mobile applications (NetAdapt) [17]. In MobileNetV3, MnasNet is used to perform a rough search on the model parameters to obtain a model with good recognition accuracy and speed, and then NetAdapt is used to fine-tune the searched model structure to obtain the best network configuration.

The second innovation is to improve the structure of the V2 version, introducing the network structure of SENet [18]. The core idea of SENet is to obtain the importance of each feature channel of the feature map through learning, and then according to the result of the importance level, the weight of the features that are useful to the current task is increased, and the weight of the features that have little effect on the current task is reduced. The SE module and the spindle structure are combined to form the bottleneck module (Bottleneck) of MobileNetV3. At the same time, a 1×1 convolutional layer before the average pooling layer in the V2 version tail network is moved to the back of the average pooling layer. First use average pooling to reduce the size of the feature map from 7×7 to 1×1, and then use 1×1 convolution to restore it.

In the MobileNetV3 network, the activation function of some modules adopts the h-swish activation function. The formula for this activation function is as follows:

$$h-swish[x]=x\frac{\mathrm{Re}\,LU(x+3)}{6} \tag{1}$$

The h-swish function is a nonlinear function. When the depth of the neural network gradually deepens, the use of the nonlinear function will reduce the computational cost of network parameters, which can make the model better reduce the parameters while ensuring the accuracy.

3. PEDESTRIAN DETECTION MODEL

3.1. Improved YOLOv3 Network Model

The current YOLO series networks have achieved good results in the field of target detection, but for pedestrian detection, due to the characteristics of pedestrians of different sizes and shapes, the YOLO network is not adaptive to pedestrian features. , and the original YOLOv3 model is too large to be deployed in small devices. Therefore, the target detection model needs to be optimized. This paper proposes a pedestrian detection model that fuses MobileNetV3 and YOLOv3 algorithms, called M-YOLOV3. In M-YOLOV3, MoblienetV3 is replaced by Darknet53 as the backbone network of the modified YOLOv3. The main part of MobileNetV3 is mainly composed of many different bottleneck modules (Bottleneck). In some bottleneck modules, SE modules are selectively added to increase the weight of features useful for training.

This paper uses the MobileNetV3 network to replace the original Darknet53 network. The network structure of MobileNetV3 is mainly composed of multiple bottleneck modules. Each bottleneck module mainly consists of three parts, which are a 1×1 convolutional layer, a 3×3 or 5×5 depthwise separable convolutional layer, and a 1×1 convolutional layer. Two 1×1 convolutional layers are used as the input channel and output channel of the feature map, respectively. The depthwise separable convolutional layer inherits the spindle-shaped structure of MobileNetV2, and enlarges the feature channel of the image according to the convolution step size. The specific network structure is shown in Figure 2.

The M-YOLOv3 algorithm utilizes convolutional neural networks to extract image features, where the use of depthwise separable convolutions can effectively reduce the number of parameters of the model. Adjust the IOU (intersection and union ratio) and the learning rate of the convolutional neural network to further optimize the training process of the model. In YOLOV3, the Kmeans clustering algorithm is selected to obtain Anchor Boxes, which are the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

initial prediction boxes of pedestrian targets. In order to obtain a better pedestrian target prediction frame, choose to use the Kmeans++ algorithm to replace the original Kmeans algorithm, so that a better Anchor Box can be obtained. At the same time, it will also improve the accuracy and recall of the model. In the final size of the target frame, the target frame of 13×13, 26×26, 52×52 is also output. The target frame of size 13×13 has the largest receptive field of each unit, so it is used to detect large size target. 26×26 is used to detect medium-sized targets, while 52×52 is used to predict small-sized targets because the cell has the smallest target receptive field.Moreover, since MobileNetV3 uses depthwise separable convolution, the amount of network parameters is much smaller than that of the Darknet53 network that originally used standard convolution, so the model trained by M-YOLOV3 will be much smaller than the model trained by the original YOLOV3. Because M-YOLOv3 has very few network parameters, the training speed of the model will also be faster than the original model.

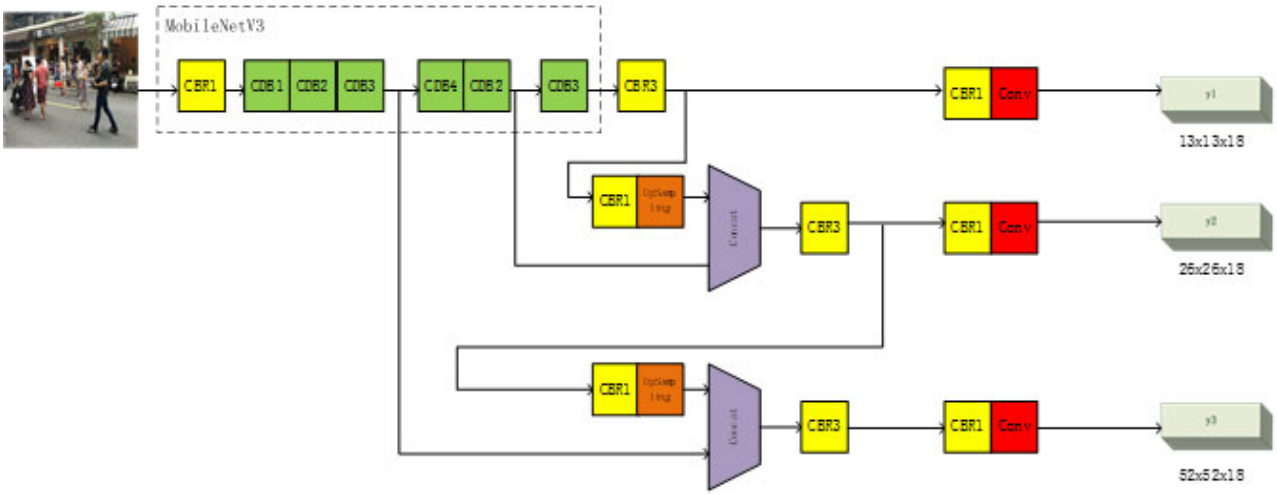

Figure 2. M-YOLOv3 network structure

3.2. Data Enhancement

In order to make the pedestrian image more expressive, and in order to allow the neural network to better extract the pedestrian image features, it is decided to optimize the image input to the neural network.

The Mosaic data augmentation algorithm is an improved data augmentation algorithm based on the cutmix[20] algorithm. The principle of Mosaic data enhancement is to select four images, randomly segment each image and scale the segmented parts according to a certain ratio, and then integrate the segmented four parts to obtain a single image containing multiple A picture of the target. With this method, the detection targets can be greatly enriched and the detection dataset can be effectively expanded. At the same time, since one image contains the detection targets of four images, the training speed of the model can be effectively improved, Reduced hardware requirements for model training.

3.3. KMeans++ Algorithm

In the YOLO series of algorithms, the concept of Anchor Box proposed by Fast RCNN is used for reference. In YOLOv2 and YOLOv3, the Kmeans clustering algorithm is selected to extract the target box, so that the extracted target box is compared with that of the original selective search algorithm. The target box is more accurate, but the Anchor Box is still not accurate enough, which may cause the target to be missed. For this reason, this paper decided to use the Kmeans++ algorithm to replace the original Kmeans algorithm. The Kmeans++ algorithm is an

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

improvement on the Kmeans algorithm. The IOU is also used for calculation, and the formula is as follows:

$$d(b,c) = 1 - IOU(b,c) \tag{2}$$

According to the K value in the Kmeans++ mean clustering algorithm and the average IOU value in each image, the K value that is most suitable for the target detection model can be selected. The curves of K value magnitude and average IOU are shown in Figure 3.

It can be seen from the curve in Figure 6 that when K=6, the slope of the image gradually decreases and becomes stable. In the selection of K value, it is necessary to ensure the speed of network detection, but also to ensure the accuracy of the algorithm model to detect pedestrians. According to the image curve, K=9 is selected as the K value in the Kmeans++ algorithm, that is, the number of Anchor Boxes is 9.

The difference from the Kmeans algorithm is that the Kmeans++ algorithm selects the center points one by one, unlike the Kmeans algorithm, the number of all center points should be selected at the beginning. This will weaken the influence of the initial selection center on the Anchor Box regression results, improve the accuracy of target box selection, and further improve the accuracy and recall of pedestrian target detection results.

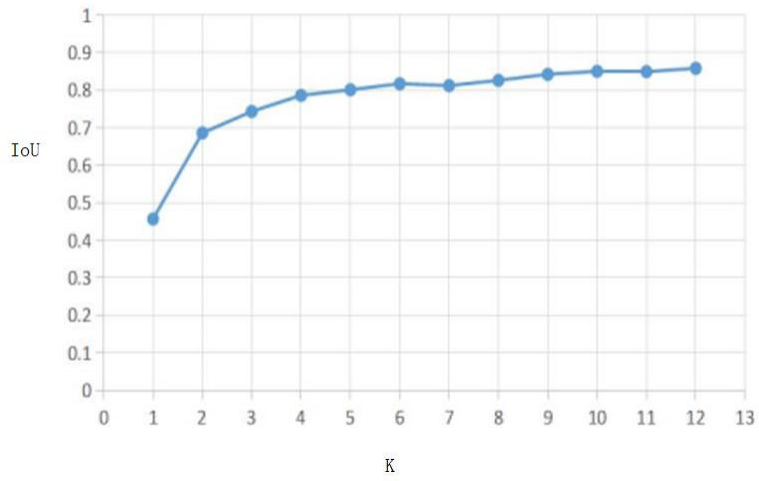

Figure 3. K value and average IOU curve

3.4. Pedestrian Detection Process

The algorithm flow of the whole experiment is mainly divided into two parts, namely the training part and the detection part. During the training phase, data augmentation is first performed on the images in the used pedestrian detection dataset. The main method is to perform bilinear interpolation and Mosaic data enhancement, which can enhance the generalization ability of the model while preventing overfitting of the model. Then these images are input into M-YOLOv3, the backbone network will perform feature extraction on these images, and classification and regression will get the final training model. Then enter the detection section. After training, each cell on the image results in 9 object detection boxes. 3 of them are used to detect large targets, 3 are used to detect medium targets, 3 are used to detect small targets, and the confidence and category of objects in each target detection frame. Afterwards, these target boxes are subjected to non-maximum suppression according to the IOU (intersection and union ratio), and the final detection box for judging pedestrians is selected.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

4. EXPERIMENT AND ANALYSIS

4.1. Experimental Platform and Dataset

As a hot research technology, moving target tracking technology has been widely used in various fields. With the help of low cost, low power consumption, self-organization and high error tolerance of wireless sensor networks, moving target tracking based on wireless sensor networks also has broad application prospects.

This experiment uses the Windows10 operating system, and the deep learning framework chooses the Tensorflow framework, the version number is 1.15.0, and the Keras framework, the version number is 2.2.4. In the experimental hardware environment configuration, the CPU is Intel Core i7-9700k, the GPU is GeForce RTX 2080Ti, and the video memory size is 11GB.

This experiment decided to use the INRIA pedestrian dataset and the CUHK01 dataset, plus a total of 5,000 images after data enhancement. Among them, the INRIA dataset is a set of images of people marked with standing, walking, etc. poses, collected from NaVneet Dalal's research work on detecting upright pedestrians in images and videos. In the INRIA data, the training set in the dataset contains 614 positive samples, 1237 pedestrians, and 128 negative samples. The test set contains 288 positive samples, a total of 589 pedestrians, and 453 negative samples. Most of the pedestrian objects marked in the dataset are standing poses and the height of the individual pedestrian is greater than 100 pixels. The image also has a high definition. The CUHK01 dataset has a total of 3884 images, including a total of 971 pedestrian targets, which are captured by two cameras.

In order to enhance the generalization ability of the network model, some images are randomly selected in the training data set for bilinear interpolation and flip transformation operations to improve the expressiveness of pedestrian target images. At the same time, the Mosaic data enhancement method is used to randomly select some images for splicing to improve the training speed and training effect of the model.

4.2. Network Training

In this experiment, the two data sets are mixed, and the experimental data is divided into training set, validation set and test set according to the ratio of 7:1:2. At the same time, some data are randomly selected to expand the experimental data by mirror transformation, linear interpolation and Mosaic data augmentation. First, through the Kmeans++ algorithm, 9 anchor boxes are selected, with sizes (30,75), (44,117), (60,158), (74,208), (98,253), (120,412), (131,413), (141,413), (152, 413). Among them, the first three anchor boxes are used to detect small objects, the middle three anchor boxes are used to detect medium-sized objects, and the last three anchor boxes are used to detect large objects. In terms of network training, the Adam optimizer is used for network training, the training round is 200 rounds, the learning rate is set to 0.001, and the size of each batch is set to 16. The network training loss curve is shown in Figure 8. The network is trained for 200 rounds, and the loss of training set and validation set shows a decreasing trend. When the training reaches 200 epochs, the loss value drops to 1.079 and val_loss drops to 1.473.

4.3. Experimental Results and Analysis

4.3.1 Lightweight comparison of pedestrian detection models

In order to verify the lightweight results of the M-YOLOV3 model, this experiment compares and analyzes the size of the network model and the amount of parameters for model training. The network models compared with M-YOLOV3 are YOLOv3 and Tiny-YOLOv3. The specific parameters of the three models are shown in Table 1. It can be seen from the table that the model training parameters of M-YOLOV3 are 1/20 of the training parameters of the YOLOv3 model and 1/3 of the training parameters of the Tiny-YOLOv3 model. At the same time, in terms

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

of model size, the size of M-YOLOV3 is 13MB , which is 1/20 of the training size of the YOLOv3 model and about 1/3 of the Tiny-YOLOv3. Compared with the two original YOLOv3 models, the M-YOLOV3 model trains with fewer parameters and a smaller model, which has a good effect on network lightweighting.

Table 1. Comparison of network models

Model	Model parameters	Model size
YOLOv3	61576342	263MB
Tiny-YOLOv3	10062523	42MB
M-YOLOv3	3011721	13MB

4.3.2 Accuracy Test

In order to verify the detection effect of the model, the mean precision and recall rate are used as the evaluation indicators in this experiment. mAP represents the average value of the average precision of detecting various types of targets in the case of n types.

The specific test results of the experiment are shown in Table 2. It can be seen that the mAP of M-YOLOv3 is 88.24%, which is 2.61% higher than that of the YOLOv3 algorithm. Compared with Tiny-YOLOv3, the improvement is large. In terms of recall rate, the recall rate of M-YOLOv3 is 3.15% higher than that of YOLOv3, and 17.38% higher than that of Tiny-YOLOV3. It can be seen that M-YOLOv3 has improved the accuracy and recall to a certain extent.

Table 2. Comparison of model accuracy and recall

Model	mAP	Recall
YOLOv3	85.63%	82.47%
Tiny-YOLOv3	66.47%	68.24%
M-YOLOv3	88.24%	85.62%

4.3.3 Experimental results

The experimental results are shown in Figure 4. It can be seen that the M-YOLOv3 network model can detect pedestrian objects in the image well.

Figure 4. Model test results

5. CONCLUDING REMARKS

This paper proposes a lightweight object detection model based on YOLOv3 and MobileNetV3. The model is trained on the public pedestrian dataset INRIA dataset and CUHK01 dataset, and

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

uses interpolation and other methods to process the images to enhance the generalization of the network model, and use the more effective dimensional clustering method to obtain more accurate results. Anchor Box. The YOLOv3 algorithm and the MobileNetV3 algorithm are integrated, and a new lightweight target detection model-M-YOLOv3 is proposed.

Compared with the current mainstream target detection algorithm and M-YOLOv3, the model obtained by M-YOLOv3 is smaller and lighter. While the model is smaller, it also has higher accuracy and faster detection speed, which ensures the real-time performance of target detection. And because of the small size of the model, it is also easy to transplant the model.

Although M-YOLOV3 has faster detection speed and higher accuracy than other models. However, in the face of complex situations, such as occlusion, shadows, and bad weather conditions, there are still problems such as inaccurate detection and poor results. Therefore, how to improve the detection precision and recall rate under complex conditions will be the main research focus in the future.

REFERENCES

[1] Feng Yuan,Li Jingzhao. Improved convolutional neural network pedestrian detection method [J]. Computer Engineering and Design, 2020,41 (05): 1452-1457.

[2] R. Girshick, J. Donahue, T. Darrell and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014:580-587.

[3] R. Girshick. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision(ICCV).2015:1440-1448

[4] S. Ren, K. He, R. Girshick and J. Sun.Faster R-CNN: towards real-time object detection with region proposal networks[C]//IEEE Transacti-ons on Pattern Analysis and Machine Intelligence.2017,39(6):1137-1149.

[5] J. Redmon, S. Divvala, R. Girshick and A. Farhadi. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2016:779-788.

[6] J. Redmon and A. Farhadi. YOLO9000: better, faster,stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).2017:6517-6525

[7] Redmon,Joseph and Ali Farhadi.YOLOv3:an incremental improvement[EB/OL].2018. https://arxiv.org/abs/1804.02767.

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

[9] Ji Mian, Zhang Xin, Xu Hai. Pedestrian detection based on improved hog feature and SVM classifier [J]. Software, 2020,41 (02): 70-74.

[10] Lowe, D.G. Distinctive image features from scale-invariant keypoints[C]//International Journal of Computer Vision,2004:91–110.

[11] A.Howard et al.Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV).2019:1314-1324.

[12] K. He, X. Zhang, S. Ren and J. Sun.Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:770-778.

[13] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017: 936-944.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0026

[14] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL],2017:https://arxiv.org/abs/1704.04861.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen.MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.2018:4510-4520

[16] M. Tan et al. MnasNet: patform-aware neural architecture search for mobile[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).2019:2815-2823.

[17] Yang, T., Howard, A.G., Chen, B., Zhang, X., Go, A.,Sze, V., & Adam, H. NetAdapt: platform-aware neuralnetwork adaptation for mobile applications[EB/OL].2018:https://arxiv.org/abs/1804.03230.

[18] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu.Squeeze-and-Excitation networks[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence,.2020,42(8):2011-2023.s