Improved YOLOv3-based Pedestrian Detection and Implementation

Jintao Wang，Yuqin Lu

Abstract—**Paper An enhanced YOLOv3 pedestrian recognition model is developed to address the problem of YOLOv3's easy to miss small objects and occlusion problem. The upgraded model replaces the original K-means clustering algorithm with the K-means++ clustering method to lessen the error effect caused by incorrect random selection of starting clustering centers. Meanwhile, the residual network module is introduced to the model to lighten it up, and the CBAM attention mechanism and the MHSA multi-head self-attention mechanism are incorporated into the framework. The algorithm's feature extraction capability is enhanced by efficiently allocating computational resources and gathering global information.Experiments testing the FLIR dataset on NVIDIA Xavier NX show that this al-gorithm greatly reduces the number of network model parameters with less loss of accuracy and improves the FPS.The experiment shows that the revised method performed well on the CUHK dataset, with the mAP value acquired during the experiment being 88.20%, 17.45% higher than the original algorithm. It has greater feature extraction capabilities, increases identification of small targets and blocked pedestrians, and has higher detection accuracy.**

Index Terms—**Object Detection;YoloV3;CBAM;multiplex self-attention mechanism;NVIDIA Xavier NX.**

I. Introduction

In recent years, pedestrian detection technology has been rapidly developed and achieved certain results, occupying an important position in the fields of intelligent face security, assisted driving system, and intelligent network connection. In the field of autonomous driving, pedestrian detection technology mainly uses vision sensors mounted on objects to analyze and process the collected area of interest to complete the recognition of pedestrians. The existing pedestrian detection techniques mainly design feature extraction methods by extracting geometric features [1][2][3]and motion information features [4]of the human body, and although the detection speed and accuracy are improved, there are still some problems that are not well solved. For example: the detection results are easily affected by light changes and occlusions, which reduces the pedestrian detection effect, and the presence of large deformation of pedestrians themselves increases the difficulty of effective feature extraction. In addition, there are problems such as complex target detection algorithms, long detection time for multi-target cases, poor real-time performance, and difficulty in achieving robustness.

At present, the algorithms used for pedestrian detection can be divided into two types: one is the detection algorithm that uses manually selected target features and then trains classifiers by machine learning; the other is the detection algorithm that uses deep learning to train network models.

In 2001, Viola et al.[5]extracted Haar features and trained an AdaBoost cascade classifier to detect the target faces. 2005, Dalal et al.[6]extracted HOG features to describe the edge features of pedestrians and found that the HOG descriptor is very suitable for the detection of human faces. found that HOG descriptors are very suitable for human detection, but the descriptor generation process is long, the real-time performance is poor, and it is difficult to deal with the problem that pedestrians are occluded.In 2015, Fei-Gang Tan et al.proposed a two-layer pedestrian detection algorithm combining binarized Haar features with multi-part verification, which improves the detection accuracy when pedestrians are partially occluded. Although traditional machine learning algorithms are also continuously optimized and updated, they are still difficult to meet the corresponding requirements.

At present, deep learning-based detection algorithms have gradually dominated the accurate detection and output when describing the target feature map.In 2014, Girshick et al.[7]proposed the R-CNN target detection framework, mainly by selectively searching for candidate regions that may contain detection targets and classifying each candidate box, and then using CNN to extract features, which has now become the most typical two-stage target detection algorithm. Subsequently, for the disadvantages of the large spatial model size and slow testing speed of the two-stage target detection algorithm, in 2015, Girshick et a.[8]proposed the Fast R-CNN algorithm based on bounding box and multi-task loss classification.In 2017, Ren et al.[9]proposed the Faster R -CNN algorithm by introducing a region suggestion network.

In the current situation, although the two-stage target detection algorithm is continuously optimized, it is still difficult to meet the real-time and robustness of the applicable scenarios of the target detection algorithm. In contrast, the single-stage target detection algorithm based on the idea of regression analysis has the characteristics of fast detection and high accuracy.In 2015, Redmon et al. proposed the YOLO detection algorithm [10], which places feature extraction, regression and classification in a single convolutional network to achieve end-to-end target detection by simplifying the network structure, but the accuracy of this algorithm for small-scale target detection and However, the

Manuscript received October 19, 2022.

Jintao Wang, School of Software, Tiangong University, Tianjin, China.

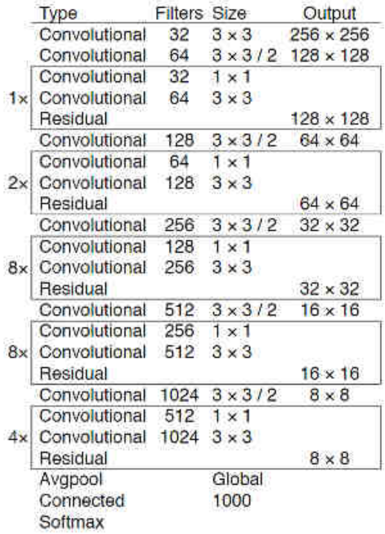
Yuqin Lu, School of Biological Science and Food Engineering, Chuzhou University, Chuzhou, China.

algorithm has low accuracy and recall for small-scale targets. To address this problem, in 2016, Liu et al.[11] proposed the SSD algorithm using the idea of hierarchical feature extraction and target prediction mechanism.In 2017, Jeong et al.[12] proposed the R-SSD algorithm based on the SSD algorithm by adding or subtracting the deconvolution module.Li et al.[13]proposed the R-SSD algorithm by fusing multiple feature layers and feature scales and generating Li et al.[13]proposed the F-SSD algorithm by fusing multiple feature layers and feature scales and generating a feature pyramid. However, the above algorithms still have the defects of poor detection of small targets and slow detection speed. Although the traditional target detection methods can basically meet the requirements of the object itself, there are still shortcomings in efficiency and accuracy that need to be improved. In this paper, we propose some improvements to the traditional YOLOv3 detection algorithm, and simulate the accuracy and recall of the improved algorithm on the CUHK datasets.

II. YOLOV3 NEURAL NETWORK ALGORITHM

Since 2015, several newer YOLO algorithms have been introduced in the academic community, and compared to the Fast R-CNN algorithm, the YOLO algorithm does not solve the detection result in 2 parts, but based on the idea of regression, the target location and its class are directly regressed in the output regression layer, which has better detection accuracy and detection speed.

In 2018, Joseph et al. proposed the YOLOv3 algorithm, which incorporates improvements to several parts compared to its predecessor. Mainly borrowing the idea of ResNet residual network, a better base feature extraction network Darknet-53 is used, which improves the detection speed to a certain extent compared with the previous network structure; meanwhile, a multi-scale fusion prediction method with three feature layers is used to improve the detection accuracy of the algorithm for small targets. Up to this point, the new cost function sigmoid is used to replace the original function Softmax in order to ensure the prediction accuracy of each target with multi-target label classification.



	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig.1 Darknet-53 structure

The YOLOv3 detection model mainly consists of two parts, the backbone network and the detection network. In Fig.1, Darknet-53, which is based on the idea of residual network, is used as the backbone network for feature extraction, and the Darknet-53 model contains 53 convolutional layers and 23 jump connections, with deeper convolutional layers compared with the YOLOv2 model. The detection network part uses the FPN feature pyramid structure used in Faster R-CNN to minimize the feature loss and improve the detection accuracy. Among them, three feature layers are extracted: the middle layer with 52*52 output feature resolution, the lower middle layer with 26*26 and the bottom layer with 13*13. The three feature layers pass detection for target objects with small, medium and large resolutions, respectively. After obtaining the three effective feature layers, we fuse the multiple features and predict the effective feature layers, and then use the decoding and prediction module to decode the processed data of the network to obtain the final results.

III. IMPROVED YOLOV3 NEURAL NETWORK ALGORITHM

A. k-means clustering algorithm

The YOLO series algorithm starts from YOLOv3, which uses 9 anchors for prediction, but still uses the same K-means clustering algorithm as YOLOv2 to obtain the size of anchors. k-means clustering algorithm will randomly designate K cluster centers (clusters) as the initial points, and keep averaging the clusters that are close to each other. When the cluster is small, the clusters are saved to determine the initial position of anchor. When the cluster is small, the clusters are saved to determine the initial position of the anchor. The similarity is determined by the IOU value, which is given by the following formula:

$$\mathrm{d}(box, centroid) = 1 - IOU(box, centorid) \qquad (1)$$

where d is the distance from sample points to each cluster center of mass; box is the other edges; centroid is the edge selected as the center in clustering; IOU is the intersection ratio of the target prediction box and the target label box. Due to the randomness of cluster centers and the sensitivity of outliers and isolated points in the K-means clustering algorithm, the clustering effect of the algorithm is easily affected by the improper selection of initial values. In addition, it can also lead to imprecision in the classification of the algorithm and the occurrence of misclassification. In this paper, the K-means++ algorithm is used to replace the original clustering algorithm in order to obtain a more consistent sample first verification frame. k-means++ differs from the K-means algorithm in that the first step is to randomly select a cluster as the initial point, and to avoid noise, the roulette wheel method is used to select a new point that is far away until K clusters are selected; Thereafter, the K-means clustering algorithm is performed. Although the K-means++ algorithm spends more time on the selection of the initial points, it actually alleviates the error caused by the improper selection of the initial clustering center and improves the computational efficiency of the algorithm.

B. Residual Network Module

The deepening of convolutional neural network depth can extract richer features and improve the detection performance. However, as the number of network layers increases, it also increases the training burden of the deeper network and causes a series of problems such as performance degradation of the network. To alleviate the problems caused by deeper networks, the YOLOv3 algorithm adopts a residual network structure similar to the ResNet proposed in the literature [14]. YOLOv2 uses a large number of 3*3 convolutional kernels for convolution, while YOLOv3 first uses a 3*3 convolutional kernel with a step size of 2 for convolution. The structure of YOLOv3 residual network module is shown in Figure 2.

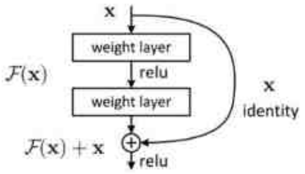

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig.2 Residual network module structure

C. CBAM

Usually, when the human eye is exposed to a scene or an objective thing, the uneven distribution of attention points causes the human attention orientation to shift to the region or information of interest. Through this selective visual attention mechanism, attention resources can be efficiently allocated and ultimately serve the subjective will of the human. Based on this, in order for computer vision to autonomously learn to pay attention to key useful information when recognizing information, researchers have generated the attention mechanism by calculating the relationship between words in the form of a probability distribution to demonstrate the relationship between words. Compared with other mechanisms, the CBAM attention mechanism uses a combination of channel attention and spatial attention, in which feature weights are inferred in the input feature map by 2 dimensions in turn, and then the weights are dot-producted with the input feature map to obtain an optimized output feature map. The overall process is shown in Figure 3.

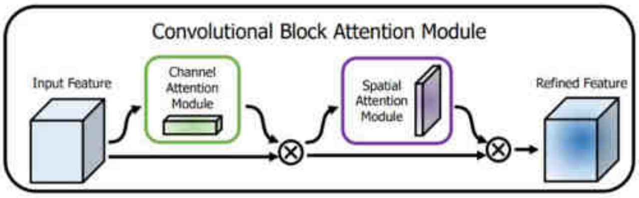


Fig.3 CBAM

The essence of the YOLOv3 detection algorithm is to encode the input image and later decode the target location and category information from it for output . In this process, by adding the CBAM attention mechanism to the network, the YOLOv3 network can be made to apply a larger weight to the pedestrians, thus enhancing the feature extraction capability. As can be seen from Fig. 1, YOLOv3 initially extracts three base feature layers, and each time the base feature layer is stacked and stitched with other upsampled feature layers, five convolutional processes are performed, at which point the attention module is added to the convolutional process.

D. MHSA(Multi-headed Self-attention)

Convolutional neural networks CNNs usually use smaller convolutional kernels to identify local features of objects when recognizing features, increasing the number of network layers while reducing parameters, but also making the perceptual field of the convolutional layers much smaller relative to the feature map. In particular, in pedestrian detection, it is often necessary to obtain pedestrian features in a larger feature map, allowing the network to gather contextual information from larger adjacent regions of the target and extract better pedestrian features. In order to obtain global information, it is necessary to extend the depth of the network and stack multiple convolutional layers, for which many computational resources are consumed.

Self-attention mechanism (Self-attention) was proposed in 2017 by Vaswani et al.[15] and mainly applied to learn text representation[15]. In text language processing, the self-attention mechanism can better capture contextual information by calculating the attention probability of each word to express the semantic relationship between words. Also, the paper proposes the Multi-headed Self-attention mechanism (Multi-headed Self-attention), in which the results of different attention point weight matrices on each mechanism are stitched together and fused by multiple Self-attention calculations to express a more comprehensive degree of association. For this purpose, it is considered to introduce it into the pedestrian detection network to extract the global features of the image. The structure of the multi-headed self-attention mechanism is shown in Figure 4.

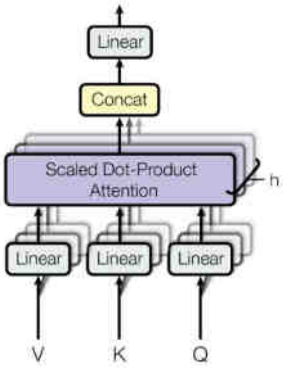


Fig.4 Multi-headed Self-attention

The mechanism maps the Q, K, and V matrices in different dimensions, stores the output parameters, and then fuses the results by stitching and fusing them several times, and then performs another matrix mapping to obtain the output results. In addition, since the use of too many multi-headed self-attentive mechanisms leads to an increase in computer load, which reduces the detection accuracy, it is added after the fourth downsampling.

E. Improved YOLOv3 model structure

In this paper, YOLOv3 is improved by changing K-means to K-means++ clustering algorithm, adding improved residual network module, CBAM attention mechanism and MHSA multi-headed self-attentive mechanism to the network structure.

IV. Experiment

A. Experimental Environment

The experiments in this paper use the Pytorch framework for network structure modification and quantization operations, training on four GPUs on a docker-deployed cluster, and inference and testing operations on an embedded device, the Nvidia Jetson NX.

The embedded device uses the Jetson Xavier NX, the latest GPU edge computing device from Nvidia in 2020, which delivers up to 14 terra operations per second (TOPS) at 10W or 21 terra operations per second at 15W, running multiple neural networks in parallel. Linux operating system.

We installed the Ubuntu 18.04 operating system image on the embedded device and configured the Python 3.6 and PyTorch environments for subsequent testing experiments on the device.

B. Datasets

The datasets selected for the experiments is the open source pedestrian detection datasets from the Chinese University of Hong Kong (CUHK), which includes 1,063 pedestrian images in total. Before starting the experiment, 800 images are selected for debugging and training, and another 100 images are taken later for validation and testing. The initial learning rate is 0.001, and the learning rate decay strategy is 0.05 times for each epoch, with 1,000 iterations and a Batch size of 45. i.e., 45 samples are taken from each epoch in the training set until all samples are traversed to complete the training once.

C. Results And Analysis

In the same experimental scenario, a total of five sets of experiments are conducted with the improved M-YOLOv3 algorithm based on YOLOv3 and the YOLOv3 detection algorithm to verify the performance improvement in pedestrian detection with the addition of each module. The details of the experiments are summarized below.

(1) Experiment A: original YOLOv3 algorithm;

(2) Experiment B: improved YOLOv3 algorithm using K-means++;

(3) Experiment C: The improved YOLOv3 algorithm using K-means++ and residual network module;

(4) Experiment D: Improved YOLOv3 algorithm using K-means++, residual network module, and CBAM attention mechanism;

(5) Experiment E: YOLOv3 algorithm improved by K-means++, residual network module, CBAM attention mechanism and MHSA multi-headed self-attentive mechanism; the average accuracy mAP, recall, accuracy Precision and summed mean F1 Score results of the five experiments are shown in Table 1.

TAB.1 Comparison of detection results of experiments

Method	mAP	Precision	Recall	F_1
Experiment A	54.7	67.6	39.6	53.9
Experiment B	63.3	70.6	50.3	61.4
Experiment C	57.2	54.5	77.2	72.9
Experiment D	60.0	85.5	78.2	74.6
Experiment E	**82.41**	**88.56**	**85.92**	**85.63**

From Table 1, we can see that the mAP value of the proposed M-YOLOv3 algorithm reaches 88.20%, while the mAP value of the original algorithm is 70.75%, which is an improvement of 17.45%; the recall rate and detection accuracy are also improved. The mAP of Experiment E is 8.85% higher than that of Experiment D and E, indicating that the multi-headed attention mechanism is more efficient than the traditional convolutional network.

This indicates that the multi-headed attention mechanism has stronger feature extraction ability than the traditional convolutional network. A comparison of the actual test results of the original YOLOv3 and the improved M-YOLOv3 algorithm proposed in this paper is shown in Figure 6. For the detection of small targets, after comparing Fig. 6(a) with Fig. 6(b), it can be clearly seen that the improved algorithm has better portrayed the edge contours of pedestrian targets and improved the feature detection effect. For the part of the pedestrian target that is heavily occluded and overlapped, Fig. 6(d) can still be recognized compared with Fig. 6(c). This shows that the improved algorithm improves the feature extraction ability for small targets and occluded parts, and is able to detect pedestrian targets more accurately.

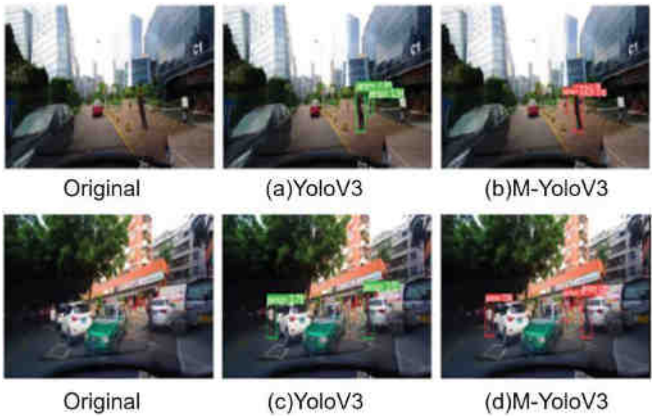


Fig. 6 Comparison of detection effects of different experiments

V. Conclusion

This paper proposes an improved pedestrian detection method based on the YOLOv3 network by selecting the K-means++ clustering algorithm, adding an improved residual network module to the network structure, and the CBAM attention mechanism and the MHSA multi-headed self-attentive mechanism. The algorithm is trained and tested on CUHK datasets, and the experimental results show that the optimized algorithm has stronger feature extraction ability, which greatly improves the pedestrian detection effect of YOLOv3 algorithm.

However, there are still problems with the proposed method, such as the improved algorithm has better improvement on the current training set, but the detection effect decreases on other data sets when the pedestrians in the images are heavily obscured or far away. Secondly, the algorithm has not been tested on actual roads and scenes, and further research will be conducted subsequently to try to improve the anti-interference and real-time detection capability of the algorithm.

REFERENCES

[1] FUJIYOSHI H,LIPTON A J. Real-time human motion analysis by image skeletonization [C]// Proceedings of the 4th IEEE Workshop on Applications of Computer Vision. Princeton,NJ,USA:IEEE, 2002:15-21.

[2] WANG Heng,KLASER A,SCHMID C,et al. Action recognition by dense trajectories [C] // IEEE Computer Society Conference on Computer Vision and Pattern Recognition.Colorado Springs,Co, USA:IEEE,2011:3169-3176.

[3] GAVRILA D M.Pedestrian detection from a moving vehicle [C] // European Conference on Computer Vision. Berlin/Heidelberg: Spring-Verlag,2000:37-4D.

[4] BOBICK A,DAVIS J.An appearance-based representation of action [C]// Proceedings of the 13th International Conference on Pattern Recognition(ICPR).Vienna,Austria:IEEE,1996:307-312.

[5] VIOLA P,JONES M.Rapid object detection using a boosted cascade of simple features [C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition.Kauai, HI,USA:IEEE Computer Society,2001:511.

[6] DALAL N,TRIGGS B.Histograms of Oriented Gradients for Human Detection [C]//IEEE Computer Society Conference on Computer Vision Pattern Recognition. San Diego,CA,USA: IEEE Computer Society,2005:886-893.

[7] GIRSHICK R,DONAHUE J,DARRELL T,et al.Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// Computer Vision and Pattern Recognition. Columbus,OH,USA:IEEE,2014:580-587.

[8] GIRSHICK R.Fast R-CNN[C]//IEEE International Conference on Computer Vision.Santigago,Chile:IEEE,2015:1440-1448.

[9] REN S,GIRSHICK R, et al.Faster R-CNN:Towards real-time object detectionwith region proposal networks J].IEEE Transactions on Pattem Analysis & Machine Intelligence,2017,39(6):1137.

[10] REDMON J,DIVVALA S,GIRSHICK R, et al. You only look once:Unified,real-time object detection [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas,NV,USA:IEEE,2015:779-788.

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

[12] JEONG J,PARK H,KWAK N.Enhancement of SSD by concatenating feature maps for object detection [J].arXiv preprint arXiv:1705.09587,2017.

[13] LI Z,ZHOU F. FSSD: feature fusion single shot multibox detector [J].arXiv preprint arXiv:1712.00960,2017.

[14] HE K,ZHANG X,REN S,er al.Deep residual learning for image recognition[C]//Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition.New York:IEEE,2016:770-778.

[15] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need [C]// Advances in Neural Information Processing Systems.Long Beach,CA,USA:NIPS Foundation,2017:30.