Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 3
CECS 2023

Complex overlapping pedestrian target detection network based on the yolov3 model

Yuchi Zhang*

School of Information Science and Engineering, East China University of Science and Technology, ShangHai200237, China

* Corresponding Author

Abstract. This paper proposes a complex overlapping pedestrian target detection model based on yolov3 model by multi-scale feature fusion and context-aware mechanism. The SONY A7R3a camera shot the model on campus, and the data set was obtained after editing and collating. There were 358 high-definition videos with a resolution of 1920*1080, and the frame rate was 50HZ, about 179,000 frames. Through testing, this paper finds that compared with Single Shot Multibox Detector (SSD), the detection accuracy of the newly proposed model is slightly improved, the detection accuracy is the same as that of Faster R-CNN, and the detection accuracy of the newly proposed model is slightly worse than that of RetinaNet. However, the detection speed of Yolov3 is more than twice that of Single Shot Multibox Detector, RetinaNet and Faster R-CNN. The input size of Yolov3 is 320*320, and the processing of a single image only needs 22ms, so the detection speed of the simplified Yolov3 tiny is faster.

Keywords: yolov3, Computer vision, pedestrian target detection, deep learning.

1. Research background

With the improvement of people's living standards, video surveillance has become an indispensable part of life, and pedestrian detection is a key component of video surveillance, its research purpose is to accurately and quickly detect pedestrian targets in complex scenes. Pedestrian detection is generally used in security monitoring projects, security monitoring is generally used in places where the flow of people is concentrated, such as communities, schools, hospitals, scenic spots, etc., these locations often have higher requirements for security protection. Traditional monitoring work is to rely on manual monitoring of abnormal situations, not only inefficient but also unable to respond in time to the coming risk. However, the use of pedestrian detection technology to automatically monitor pedestrians not only saves time and effort, but also can respond quickly to avoid the misjudgment of the detection results due to people's subjective thinking, and ultimately improve the efficiency of security monitoring. Therefore, pedestrian detection is widely used in pedestrian-intensive, security needs and monitoring needs of high occasions. Such as traffic light intersection violation detection, home elderly fall detection, and campus students missing detection.

In recent years, there have often cases of missing students, campus theft, and campus crimes. The theft cases in the university campus account for 60% ~ 70% of the total number of all kinds of cases in the university, and the trend is rising. Whether it is the safety and track monitoring of campus personnel, or the identification and differentiation detection of off-campus personnel, it is an urgent problem to be solved. An important part of the campus security system is the pedestrian detection system, the campus environment is complex, the flow of people is large, and the pedestrians carry more bags or bicycles, the video image of the pedestrian detection faces many challenges.

This paper mainly studies the real-time pedestrian detection algorithm based on campus pedestrian video, and solves the problems such as slow detection speed, low accuracy and poor real-time detection ability of traditional pedestrian detection algorithm. The structure of the paper is as follows: Section 1 introduces the research background and significance of pedestrian detection, and lists the key scenarios and problems to be solved in this paper; Section 2 reviews the current research status of pedestrian detection, representative algorithms and problems to be solved; Section 3 introduces the

implementation of detection algorithm based on YOLO model; Section 4 introduces the construction of campus pedestrian detection data set and analysis of experimental results. In Section 5, the performance improvement and future research prospects of the proposed algorithm are summarized.

2. Research status of pedestrian detection based on deep learning

2.1. Pedestrian detection method based on deep learning

Typical deep neural networks include convolutional layers, pooled layers, and fully connected layers, etc. These networks do not need to manually design features, but can jointly extract features, capture deformed parts of the human body, and perform classification tasks to effectively optimize model performance.Pedestrian detection technology is a kind of complex and changeable target detection task, which must not only accurately distinguish people or objects, but also accurately locate the target position in the image. Therefore, the accuracy of image classification and the relevance of target location have become important criteria for evaluating pedestrian detection performance.

The CNN architecture has undergone several evolutions over time, such as VGG[1], ResNet[2], Inception[3], and EfficientNet[4]. They effectively capture features at different scales and levels of abstraction through multi-layer convolution and feature pyramid structure, which improves detection accuracy. In terms of target detection frameworks, the widely adopted target detection frameworks in pedestrian detection include Faster R-CNN[5], YOLO (You Only Look Once) [6] and SSD (Single Shot MultiBox Detector)[7]. These frameworks combine area proposal networks (RPN) [8] and convolutional networks to perform both target location and classification, enabling efficient pedestrian detection. In terms of data enhancement and transfer learning, through data enhancement techniques, researchers can expand the training data and improve the robustness of the model. In addition, transfer learning allows models to share knowledge across different tasks or domains, accelerating the training and generalization capabilities of pedestrian detection models. In terms of multi-modal information fusion, the fusion of multiple sensor data, such as RGB images, depth images, and infrared images, helps to improve the robustness of pedestrian detection in complex environments. In terms of attention mechanism and context modeling, the introduction of attention mechanism and context modeling technology can help the model better understand the image content, and improve the performance of pedestrian detection in the case of high congestion or occlusion.

2.2. Research on pedestrian detection algorithm based on deep learning

In recent years, with the rise and continuous development of artificial intelligence technology, artificial intelligence methods with deep learning as the core have been gradually applied in various fields, and pedestrian detection algorithms based on deep learning have improved the speed and accuracy of pedestrian detection algorithms. In 2014, the RCNN] algorithm was proposed by Ross Girshick et al. [9], which can carry out effective classification. After that, a variety of target detection algorithms with excellent performance were successively proposed, such as Fast R-CNN[10], Faster R-CNN [11], SSD, YOLO v1, YOLO v2 and YOLO v3. The stable performance and fast recognition ability of these algorithms make it possible for the models based on deep learning pedestrian recognition algorithms to be applied to the security systems and trajectory tracking systems of major campuses.

2.2.1. Deep learning pedestrian detection algorithm based on convolutional neural network

With the rise of deep learning technology, a series of object detection algorithms based on deep neural networks have been proposed. The most representative algorithm is R-CNN series, including R-CNN[9], Fast R-CNN[10], Faster R-CNN[11], etc. These algorithms avoid feature engineering and classifier design in traditional methods by splitting the object detection task into two subtasks: region extraction and classification/regression. In addition, such as YOLO[6], SSD[7], RetinaNet[12], etc. These algorithms usually adopt a single-stage detection method to complete the target detection task by directly regression target box and category score. The object detection algorithm based on

regression does not need the candidate region to generate branches, and directly regress the candidate boxes and categories of the target in multiple positions of the given input image. Therefore, these algorithms usually have faster detection speed and higher accuracy, so they have been widely used in practical applications.

1) Development of Faster R-CNN algorithm

In 2014, Girshick[9] et al. successfully applied Convolutional Neural Networks (CNN) in the field of object detection and proposed the R-CNN algorithm. It combines AlexNet[13] and selective search algorithm [14] to decompose the target detection task into several independent steps (as shown in Figure 1). First, the selective search algorithm is used to extract 2000 candidate regions, and then each candidate region is normalized. The features are extracted from the input CNN one by one. Finally, SVM classification and regional regression are performed for the features.

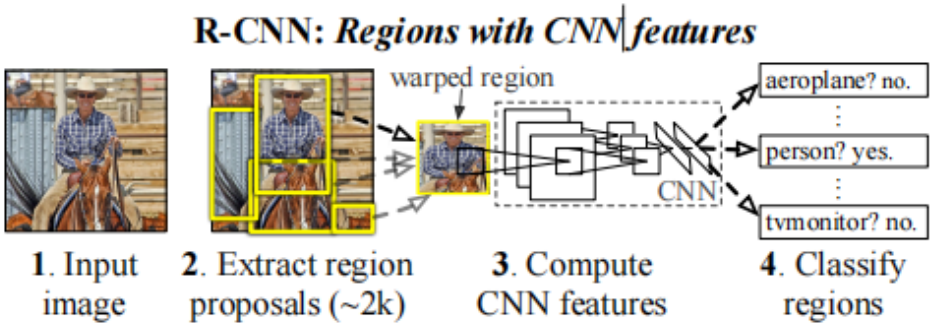

Figure 1. R-CNN network example

In 2015, Girshick et al. [10] proposed the Fast R-CNN algorithm (as shown in Figure 2). Inspired by the SPP-NET algorithm, Simplifying the SPP layer into a single-scale ROI Pooling layer to unify the size of candidate region and proposes the idea of multi-task loss function. The classification loss and bounding box regression loss are trained and learned in a unified manner so that classification and localization tasks can not only share convolution features but also promote each other to improve the detection effect.

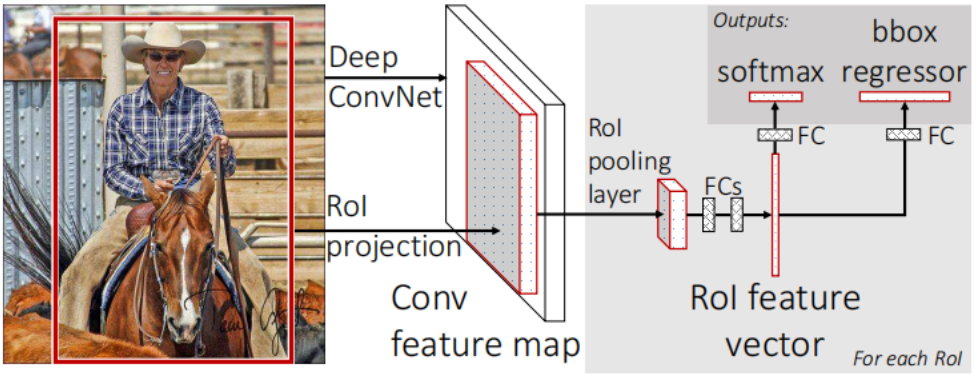

Figure 2. Fast-R-CNN network structure

Although Fast R-CNN effectively speeds up the detection rate, it still relies on selective search algorithms to generate candidate regions. Some studies have shown that the convolutional layer of convolutional neural networks has a good ability to locate the target, but this ability is weakened in the fully connected layer. Therefore, Ren et al. [11] proposed the Faster R-CNN algorithm framework in 2015 (as shown in Figure 3), and designed the RPN to replace the selective search algorithm that assisted in sample generation. RPN is a Fully Convolutional neural Network (FCN) structure, which takes the feature graph of any size as input and generates a series of candidate regions that may contain targets after the convolution operation, which enables the algorithm to achieve end-to-end training and greatly improves the detection speed.

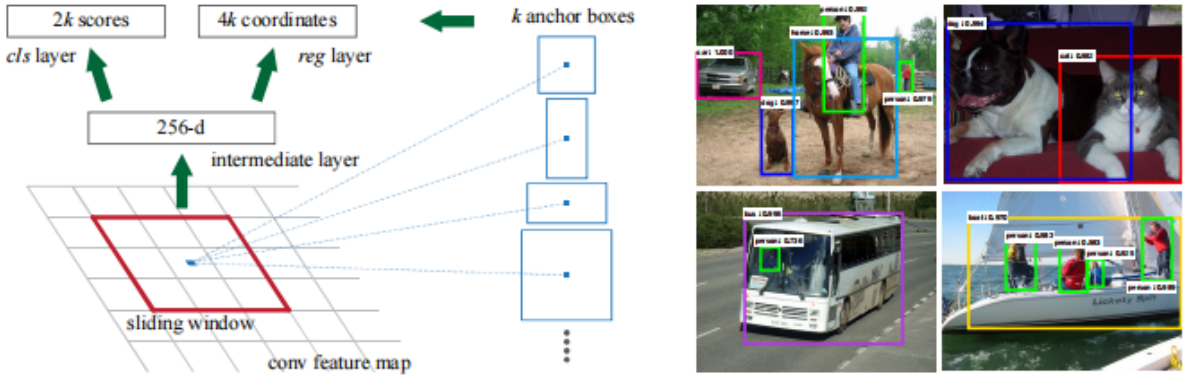

Figure 3. Faster R-CNN network structure

2) SSD algorithm

In 2016, Liu et al. [7] proposed the SSD algorithm. Based on the idea of regression, it effectively combined the idea of multi-scale detection to extract multiple feature maps of different scales for detection, and followed the strategy of using larger feature maps to detect relatively small targets and smaller feature maps to detect larger targets, which significantly improved the detection effect of large targets. The detection of small targets has also been improved. At the same time, by referring to the Anchor mechanism of the Faster R-CNN algorithm, a fixed number of default boxes with different scales and aspect ratios are preset at each position of the extracted feature map. The network can directly conduct intensive sampling and extract candidate boxes on the feature map for prediction. While maintaining the real-time detection speed, The positioning accuracy of the model is improved. As shown in Figure 6, the SSD network is based on the full convolutional network structure, which replaces the full connection layer of the basic network VGG16[15] with the convolutional layer, and adds several auxiliary convolutional layers at the end of the VGG16 network to gradually reduce the size of the feature map for extracting feature maps of different scales. Moreover, the convolution operation is used directly to detect the feature graphs of different scales.

2.2.2. Development and utilization of YOLO algorithm

In 2015, Redmon[10] et al. proposed the YOLO algorithm, which integrates classification, positioning and detection functions into a network. The input image only needs to undergo one network calculation, and the boundary box and category probability of the target in the image can be directly obtained. As shown in Figure 4, the YOLO algorithm divides the entire input image into S×S grid graphs. Each grid is only responsible for the target object whose center falls on the grid and only predicts B bounding box information, and then selects the appropriate confidence threshold to remove those bounding boxes with low probability of having a target. Although YOLO algorithm completely abandoned the step of candidate region generation, greatly improved the detection rate, and can meet the speed requirements of real-time target detection, due to its rough network design, it is far from meeting the accuracy requirements of real-time target detection, and there are problems such as inaccurate target positioning, easy to miss detection, and poor detection effect of small targets and multi-targets.

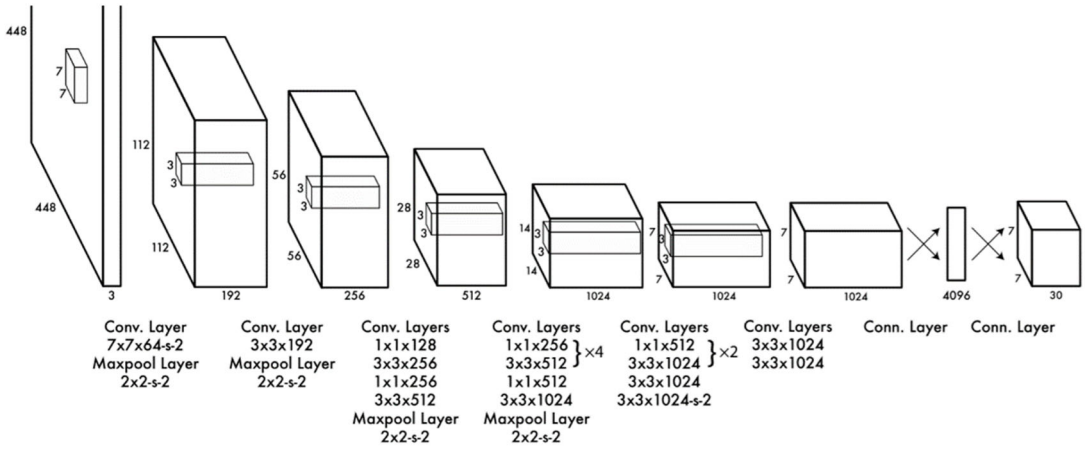

Figure 4. Schematic diagram of yolov1 detection

In 2017, Redmon et al. [16] proposed the YOLOv2 algorithm and made a series of improvements to the YOLO algorithm, focusing on solving the problems of low recall rate and poor positioning accuracy. It uses the Anchor mechanism of Faster R-CNN algorithm for reference, removes the full connection layer in the network, and uses the convolutional layer to predict the position offset and category information of the detection box. And different from the manual design of the original Anchor mechanism, it uses K-Means clustering [17] to learn the best initial Anchor template in the training set. Moreover, YOLOv2 adds a pass-through layer to connect shallow feature maps to deep feature maps, making the network have fine-grained features. In addition, YOLOv2 can adopt the method of joint optimization training of multiple data sets, and use WordTree method to synchronize training on ImageNet classification data set [18] and MS COCO detection data set [19] to achieve real-time detection tasks of more than 9000 target categories.

In 2018, Redmon et al. [20] proposed the YOLOv3 algorithm, which learned from the idea of jump connection in residual networks to build a 53-layer baseline network named DarNet-53, which only uses 3×3 and 1×1 convolutional layers, and has a classification accuracy similar to ResNet-152[19]. But it greatly reduces the amount of computation. To deal with multi-scale targets, three different scale feature maps are used for target detection, and each feature map is a fusion of high and shallow feature maps. When predicting categories, the Logistic regression method is used instead of the Softmax method, so that each candidate box can predict multiple categories, supporting the detection of objects with multiple labels. YOLOv3 algorithm can meet the accuracy and speed requirements of real-time detection tasks and has become one of the preferred target detection algorithms in the current engineering field.

YOLO series models have the characteristics of real-time, simplicity and high accuracy. The ability to perform object detection at real-time speeds is suitable for many applications, such as autonomous driving and real-time video analytics. YOL model transforms the object detection problem into a single regression problem, which is easy to understand and implement. YOLO models can handle complex scenes and small-size targets. YOLO models can perform multi-class detection. yolov3 model adopts direct regression target box and category score to complete the target detection task. It detects the pedestrian whose shape and behavior are relatively complex, and provides the target detection method for the pedestrian target in the overlapping state and the action state. Due to the complexity of the campus background, not only pedestrians will frequently appear on the identification screen, but other moving objects such as bicycles, suitcases, and cars will also overlap with pedestrians. The objective of the data set mainly adopted in this paper is analyzed from the time sequence, which has the characteristics of partial time concentration and overall time dispersion. During the peak dining period, the number and complexity of pedestrians are relatively high, while during theOf f peak dining hours, the number of pedestrians is relatively loose, but the number is still large. The whole presents the characteristics of complex and changeable, regular and concentrated. Spatially, this paper mainly adopts that more than 95% of the targets in the data set are in motion,

including the crowd of backpacks, the cyclists, and the overlapping crowd. In the background, there are a large number of stationary targets, such as motorcycles, bicycles, etc., which show the characteristics of moving and overlapping in space. In addition, the pedestrian targets in the self-built data set used in this paper are cohesive and complex, and the background is similar to the targets, which is difficult to detect and analyze. As a relatively mature target detection algorithm, yolov3 model meets the requirements of speed and accuracy required by the task. Considering the stability problem, the yolov3 model is selected in this paper.

3. Methodology

3.1. Analysis of yolov3 detection methods

YOLOv3 is an efficient object detection model that adopts a real-time detection method called "You Only Look Once". The method achieves object detection by dividing the entire image into smaller grids, each of which predicts a set of bounding boxes and corresponding class probabilities.

3.1.1. Input image preprocessing

First, the input images of the YOLOv3 model were preprocessed. The original image is scaled to a size of 416x416 and then normalized to scale the pixel values to between 0 and 1. Next, the image is converted into three feature maps of different sizes, which are used to detect large, medium, and small-sized targets respectively.

3.1.2. Feature extraction network

YOLOv3 uses Darknet-53 as its feature extraction network. Darknet-53 can efficiently learn high-level features of images. The network structure of Darknet-53 adopts the residual network structure, which can effectively avoid the gradient disappearance problem and has a high accuracy.

3.1.3. Detection layer

After the feature extraction network, three detection layers are added to the YOLOv3 model, which is used to detect large, medium and small-size targets respectively. Each detection layer consists of three convolution layers and a final output layer. Each output layer predicts a set of bounding boxes and corresponding class probabilities.

3.1.4. Adjusting Prediction Enclosures

When the YOLOv3 model outputs the test results, the anchor frame is adjusted to improve the test accuracy. Anchor boxes are a set of predefined boxes used to represent targets of different sizes and aspect ratios. For each detection layer, the YOLOv3 model uses anchor frames of different sizes and proportions in order to be able to detect targets of various sizes and proportions.

3.1.5. Non-maximum suppression

Finally, the YOLOv3 model uses Non-maximum suppression (NMS) [21] to filter the overlapping detection results. The NMS algorithm can identify overlapping bounding boxes and select the box with the highest category probability as the final output result.

To sum up, YOLOv3 model adopts an efficient target detection method, which can quickly and accurately detect targets in images. Through pre-processing, feature extraction network, detection layer, prediction frame adjustment and non-maximum suppression, the YOLOv3 model can quickly process large-scale image data and has a high detection accuracy. In addition, the YOLOv3 model has a lower computational cost and memory footprint, and is suitable for resource-constrained environments such as embedded systems and mobile devices. However, the YOLOv3 model still has some limitations, for example, the detection of small targets is not accurate enough, and the processing of target occlusion and deformation is not robust enough. In addition, YOLOv3 model also has a high false detection rate, and further optimization algorithm is needed to improve the detection accuracy.

3.2. Yolov3 Network Architecture

3.2.1. yolov3 Architecture

yolov3 uses the Darknet-53 network as the backbone architecture, as shown in Figure 5.

	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			

Figure 5. Darknet-53 architecture

Backbone has evolved from Darknet-19 in the Yolov2 period to Darknet-53, deepening the number of network layers and introducing cross-layer addition operations in Resnet. The Darknet-53 processes 78 graphs per second, slower than the Darknet-19, but faster than the ResNet with the same precision. (As shown in Table 1)

Table 1. Precision performance comparison of Darknet

Backbone	Top-1	Top-5	Bn Ops	BFLOP/s	FPS
Darknet-19	74.1	91.8	7.29	1246	171
ResNet-101	77.1	93.7	19.7	1039	53
ResNet-152	77.6	93.8	29.4	1090	37
Darknet-53	77.2	93.8	18.7	1457	78

3.2.2. yolov3 Network Structure

Yolov3 uses Darknet-53 as the classification backbone of the entire network. According to the code, tidy up the data flow as shown in Figure 6.

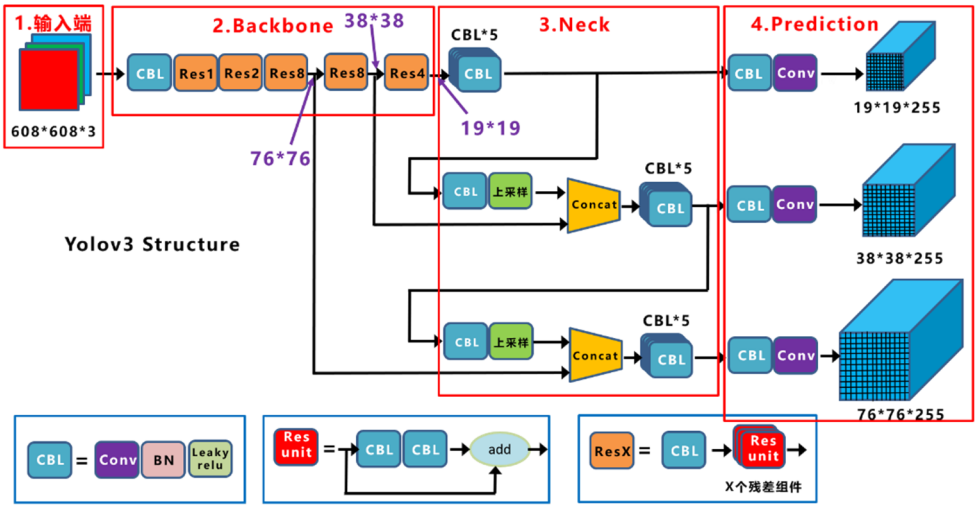

Figure 6. Schematic diagram of yolov3 model

1) In Yolov3, only the convolution layer can control the size of the output feature map by adjusting the convolution step. So, there is no special limit on the size of the input image. In the flowchart, the input picture is 256 x 256.

2) Yolov3 draws on the pyramid feature map idea. sampling 32 times under the first feature map, 16 times under the second feature map, and 8 times under the third feature map. The output dimension of the feature graph is N×N×[3×(4+1+80)], and N×N is the number of grid points of the output feature graph. There are three Anchor boxes in total, and each box has 4-dimensional prediction box values tx,ty,tw,th, 1-dimensional prediction box confidence, and 80-dimensional object category number. Therefore, the output dimension of the first layer feature graph is 8×8×255.

3) Yolov3 outputs a total of 3 feature maps. The first feature map is subsampled 32 times, the second feature map is subsampled 16 times, and the third one is subsampled 8 times. The input image passes through Darknet-53 (without a full connection layer) and the feature map generated by Yoloblock is used as a dual-purpose. The first use is to generate feature figure 1 after 3*3 convolutional layer and the 1*1 convolutional layer. The second use is to combine the output results of the middle layer of the Darnet-53 network with a 1*1 convolutional layer and sampling layer. This generates feature figure 2. The same loop is followed by the feature figure 3.

4) Difference between concat operation and add and sum operation. The add and sum operation comes from the idea of ResNet, which adds the input feature graph to the corresponding dimension of the output feature graph, that is, y=f (x) +x, while the concat operation comes from the design idea of DenseNet network, which directly concatenates the feature graph according to the channel dimension. For example, the 8*8*16 feature map is spliced with the 8*8*16 feature map to generate the 8*8*32 feature map.

5) upsample layer: The role is to generate large-size images from small-size feature maps through interpolation and other methods. For example, using the nearest neighbor interpolation algorithm, the 8*8 image is transformed into 16*16. The upper sampling layer does not change the number of channels in the feature map.

Yolo's network drawing on the essence of Resnet, Densenet, FPN, can be said to be a combination of all the most effective target detection techniques in the industry today.

3.2.3. Yolo Output Feature Map decoding (forward process)

According to different input sizes, the output feature maps of different sizes will be obtained. Taking the input image 256 × 256 ×3 in Figure 2 as an example, the output feature maps are 8×8×255, 16×16×255, 32×32×255. In the design of Yolov3, three different prior frames are configured in each lattice of each feature graph, so the last three feature graphs are 8×8×3×85, 16×16×3×85, 32×32×3×85, which is easier to understand.

The three feature maps are the detection results of the entire Yolo output, including the detection box position (4d), detection confidence (1d), and category (80d), which add up to 85 d. The last dimension 85 of the feature graph represents this information, While the other dimensions of the feature map, N × N × 3, N × N represents the reference position information of the detection frames, and 3 is the 3 different scales of anchor. How to decode the detection information is described in detail below:

1) anchor

Yolov3 followed the technique of anchor in Yolov2, and used k-means to cluster label boxes in the data set, and obtained 9 boxes of the category center point as anchor. In the COCO data set (the original picture resized 416×416), the nine boxes are (10×13), (16×30), (33×23), (30×61), (62×45), (59×119), (116×90), (156 ×198), (373×326) in the order w×h.

Note: The anchor is only related to the w and h of the detection box, not to x and y.

2) Detection box decoding

With the anchor and the output feature map, the detection box x, y, w, and h can be decoded.

$$b_x = \sigma(t_x) + c_x \tag{1}$$

$$b_y = \sigma(t_y) + c_y \tag{2}$$

$$b_w = p_w e^{t_w} \tag{3}$$

$$b_h = p_h e^{t_h} \tag{4}$$

As shown in Figure 7, the offset of the grid point coordinates based on the upper left corner of the center point of the rectangular box is the activation function. The width and height of the actual prediction box can be calculated by using the width and height of the sigmoid prior box through the above formula.

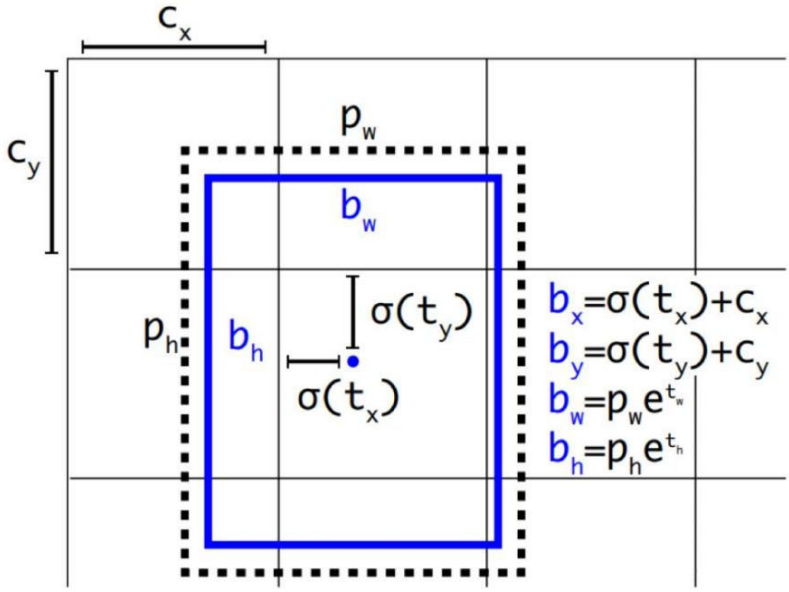

Figure 7. Decoding of the detection box

3) Detect confidence decoding

The detection confidence of objects is very important in Yolo design, which is related to the detection accuracy and recall rate of the algorithm. The confidence occupies a fixed bit in the output 85 dimensions, which can be decoded by the sigmoid function, and the numerical interval after decoding is in [0,1].

4) Category decoding

The COCO dataset has 80 categories, so the number of categories accounts for 80 of the 85 dimensions of output, with each dimension independently representing the confidence level of a category. The sigmoid activation function replaces softmax in Yolov2, canceling the mutual exclusion between classes, and making the network more flexible. A total of 8×8×3 + 16×16×3 + 32×32×3 = 4032 boxes and corresponding categories and confidence can be decoded from the three feature maps. The 4032 boxes are used differently in training and reasoning. During training, all 4032 boxes are sent into the labeling function for the calculation of the label and loss function in the latter step. When reasoning, select a confidence threshold, filter out the low threshold box, and then through NMS (non-maximum suppression). It can output the prediction results of the whole network.

3.2.4. yolov3 training Strategy and loss Function (reverse process)

1) The prediction box is divided into three cases: positive, negative, and ignore.

2) Positive example: Take any ground truth and calculate the IOU with 4032 boxes. The largest prediction box of the IOU is a positive example.A prediction box can only be assigned to one ground truth. For example, if the first ground truth already matches a positive example detection box, then the next ground truth is found to be the largest IOU detection box among the remaining 4031 detection boxes as a positive example. The order of the ground truth is negligible. Positive examples generate confidence loss, detection frame loss, and category loss. The prediction box is the corresponding ground truth box label (reverse coding is required, and the real x, y, w, h is used to calculate (tx,ty,tw,th); The category label corresponds to the category 1, and the rest is 0. The confidence label is 1.

3) Ignore example: except for the positive example, if the IOU with any ground truth is greater than the threshold value (0.5 is used in this paper), the example is ignored. Ignoring the sample does not result in any loss.

4) Negative example: Except the positive example (the IOU is the largest detection box after calculation with ground truth, but the IOU is less than the threshold, it is still a positive example), and the IOU of all ground truth is less than the threshold (0.5), it is a negative example. Negative cases produce loss only with confidence, and the confidence label is 0.

Loss function:

$$Loss = \text{loss}_{N_1} + loss_{N_2} + loss_{N_3} \tag{5}$$

a) λ is a weight constant, which controls the ratio between detection box Loss, obj confidence Loss and the proportion of no obj confidence Loss. Usually, the number of negative cases is dozens of times more than the number of positive cases, and the detection effect can be controlled by weight over parameter.

b) 1ijobj output 1 if positive example, otherwise 0; 1ijnoobj outputs 1 if negative, 0 otherwise; Ignore the sample output 0.

c) x, y, w, h use MSE as a loss function, and can also use smooth L1 loss (from Faster R-CNN) as a loss function. smooth L1 makes training smoother. Since the confidence and class labels are 0,1 and 2 classifications, cross-entropy is used as a loss function.

3.2.5. Accuracy and performance of yolov3 model

Table 2. Precision comparison diagram

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++	ResNet-10**1-C4**	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI	Inception-ResNet-v2	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	52.1
One-stage methods							
YOLOv2	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
SSD513	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
1RetinaNet	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	**51.2**
YOLOv3 608 x 608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

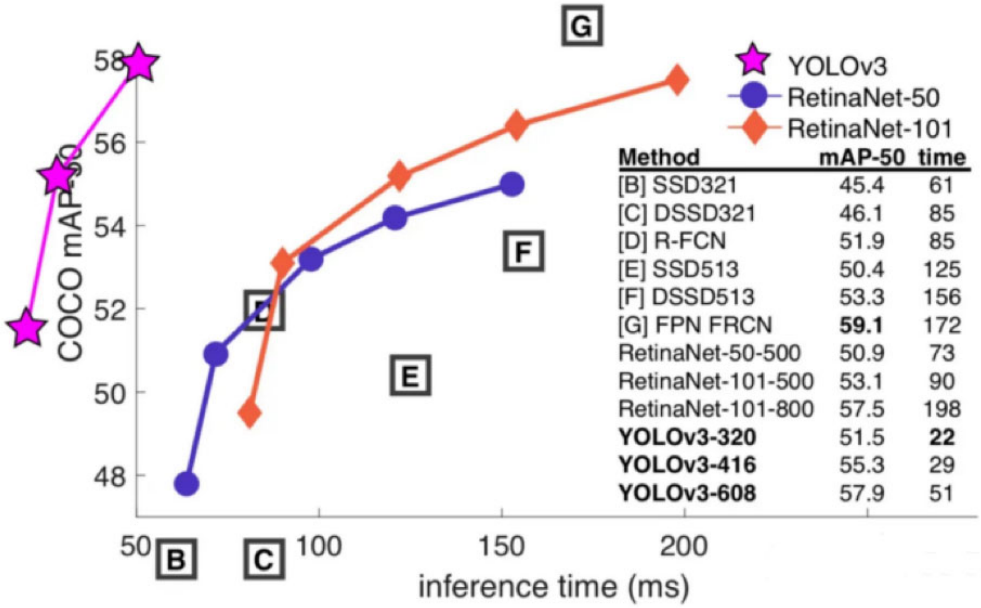

Figure 8. Performance comparison diagram

As shown in Table 2 and Figure 8, the accuracy of Yolov3 is higher than SSD, equal to Faster R-CNN, but lower than RetinaNet. However, the processing speed is more than twice that of SSD, RetinaNet, and Faster R-CNN. Input size of 320*320 Yolov3, single image processing only needs 22ms, simplified Yolov3 tiny can be faster.

4. Data sets

4.1. Self-constructed dataset collection

The object of this paper is the complex flow of people in the campus, using coco dataset and self-built dataset for model training, through the camera in the campus flow of large intersections and doorways for photographic records, the collection of including overlapping crowds, cycling crowds,

backpacking crowds, including the complex crowd dataset of the campus characteristics, the range of 1920*1080 to 192*108 pixels are selected as the images in the dataset, and a part of the data is selected as the training set and a part of the data is selected as the test set by classifying the collection. Shooting in multi-scene and multi-angle makes the image set have complexity and diversity, which can effectively improve the speed and stability of the model training, and at the same time, improve the detection accuracy and frame rate.

This paper uses labelImg software [22] to annotate the dataset, which is written in python and Qt, and by annotating the rectangular box, we can get the annotation information of the xml format file after saving. In this paper, three folders are firstly created during dataset annotation as Annotations, ImageSets and JPEGImages, and .xml file is generated after annotation. As shown in Figure 9.

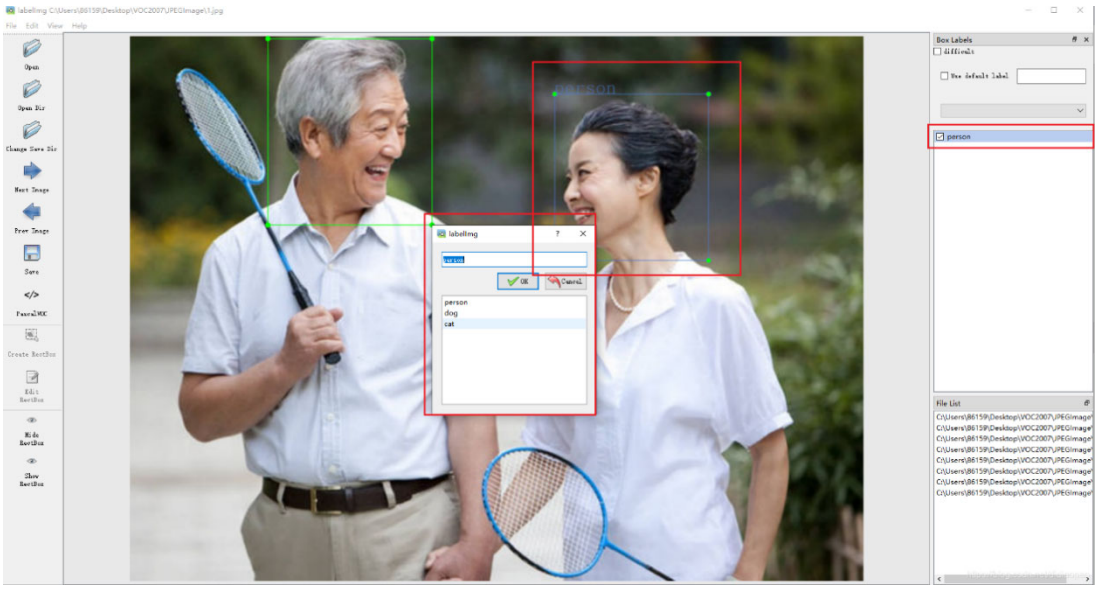

Figure 9. labelimg annotation interface

4.2. Construction of self-constructed dataset

The dataset used in this paper consists of a total of 358 videos with a resolution of 1920*1080 HD video and a frame rate of 50 HZ.The dataset was acquired in the early stage by using a SONY A7R3a camera for field shooting on the campus, after which the dataset videos were edited by using Adobe Premiere, and the final 358 videos were obtained, totaling about 179,000 frames.

4.3. Analysis of the detection results of the dataset

In this paper, 358 videos collected were detected and evaluated using the trained yolo detection model, and more excellent detection data were obtained, the following are the sample detection results of the featured population.

4.3.1. Backpack and suitcase crowd

As shown in Figure 10, this type of crowd tends to appear densely at school exit locations such as school entrances, and suitcases and backpacks affect the evaluation and judgment of the pixel points, in the improved model, the crowd that appears to overlap can achieve 70% accuracy, while the accuracy of the detection of suitcases is at 55%.

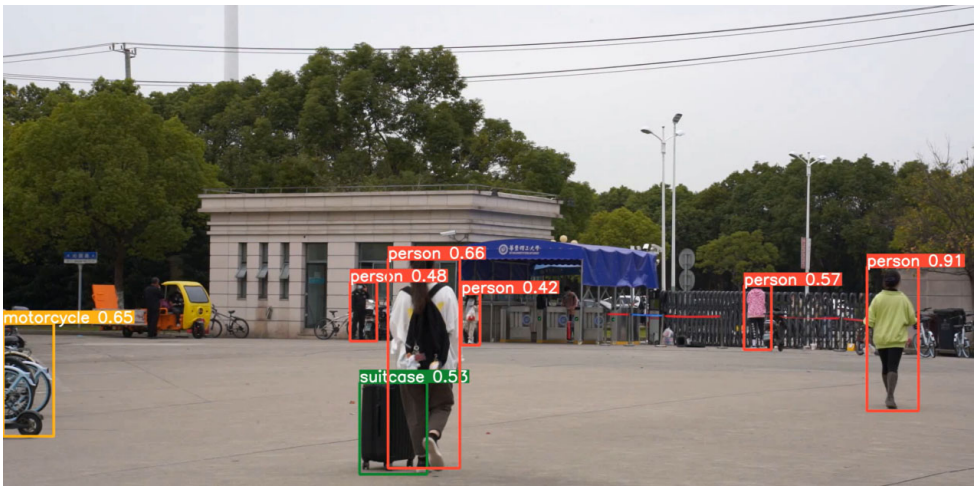

Figure 10. Crowd of backpackers pulling suitcases

4.3.2. Highly overlapping population

This type of crowd tends to appear in the business district of the campus, with a large number of targets overlapping or moving at high speed. In the model, 60 to 80% accuracy can be achieved. As shown in Figure 11, 12.

Figure 11. Highly overlapping population

Figure12. Highly overlapping population

4.3.3. Crowd of cyclists

This kind of crowd is also a common detection problem in the campus, because cyclists are fast, and their body posture is different from pedestrians, this model can basically reach 60% of the detection level of this kind of target, and complete the detection accuracy of about 80% of the bicycle driven by them. As shown in Figure 13,14.

Figure 13. Crowds of cyclists

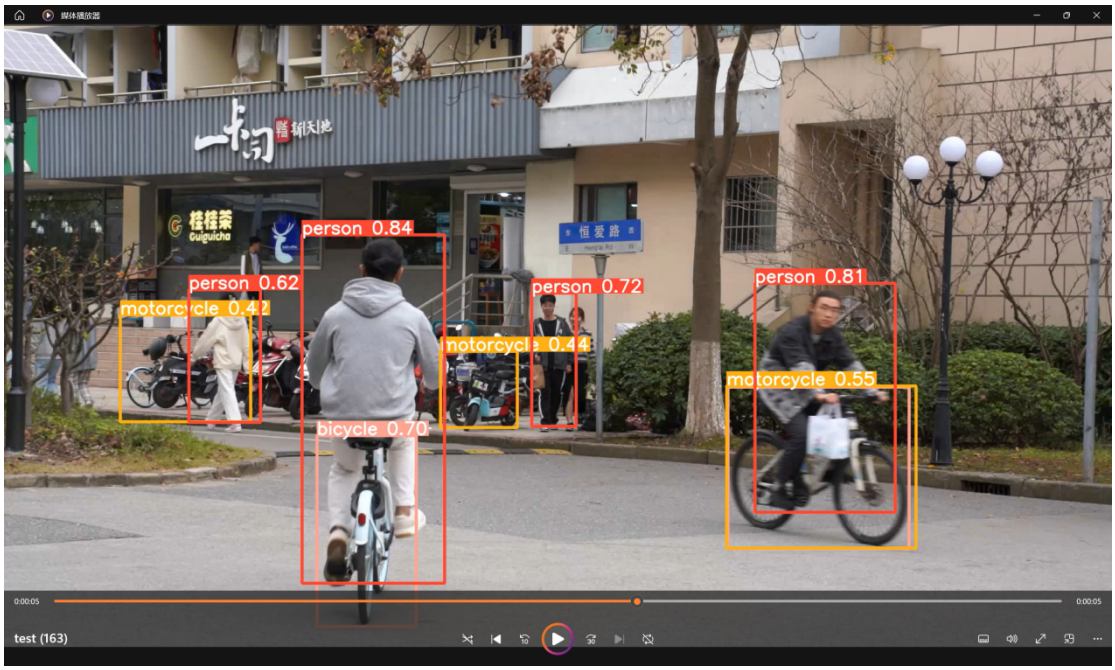

Figure 14. Crowds of cyclists

5. Summary

This paper introduces a complex overlapping pedestrian target detection network based on YOLOv3 model. By introducing multi-scale feature fusion and context-aware mechanism, this network improves the detection accuracy and stability of overlapping pedestrian targets in complex scenes.

Firstly, the network uses multi-scale feature fusion technology to detect targets on different feature maps, thus improving the accuracy and robustness of target detection. Secondly, the network uses a context-aware mechanism to improve the accuracy and stability of target detection by introducing context information. Finally, the network is verified by experiments on the pedestrian target detection dataset, and the results show that the network has better performance in overlapping pedestrian target detection than other target detection methods. In the follow-up work, this method can be combined with engineering application to study the monitoring software with high practical value.

References

[1] Sengupta A, Ye Y, Wang R, et al. Going deeper in spiking neural networks: VGG and residual architectures. Frontiers in neuroscience, 2019, 13: 95.

[2] Targ S, Almeida D, Lyman K. Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029, 2016.

[3] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision, Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[4] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR, 2019: 6105-6114.

[5] Girshick R. Fast r-cnn: Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection, Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[7] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[8] Fan Q, Zhuo W, Tang C K, et al. Few-shot object detection with attention-RPN and multi-relation detector, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4013-4022.

[9] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[10] Girshick R. Fast r-cn, Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[11] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, 28.

[12] Wang Y, Wang C, Zhang H, et al. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery. Remote Sensing, 2019, 11(5): 531.

[13] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016.

[14] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition. International journal of computer vision, 2013, 104: 154-171.

[15] Qassim H, Verma A, Feinzimer D. Compressed residual-VGG16 CNN model for big data places image recognition,2018 IEEE 8th annual computing and communication workshop and conference (CCWC). IEEE, 2018: 169-175.

[16] Redmon J, Farhadi A. YOLO9000: better, faster, stronger,Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[17] Hamerly G, Elkan C. Learning the k in k-means. Advances in neural information processing systems, 2003, 16.

[18] You Y, Zhang Z, Hsieh C J, et al. Imagenet training in minutes,Proceedings of the 47th International Conference on Parallel Processing. 2018: 1-10.

[19] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context,Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[20] Redmon J, Farhadi A. Yolov3: An incremental improvement.arXiv preprint arXiv:1804.02767, 2018.

[21] Bodla N, Singh B, Chellappa R, et al. Soft-NMS--improving object detection with one line of code,Proceedings of the IEEE international conference on computer vision. 2017: 5561-5569.

[22] Tzutalin D. LabelImg. GitHub repository, 2015, 6.