A Real-time Vehicle Detection Method for Unmanned Driving

Lai Wei¹, Bo Zhou², ³, *, Bin Liu¹

¹Shandong University of Technology Shandong, China
²School of Information, Beijing City University, Beijing, China
³School of Mechanical Engineering, University of Science and Technology Beijing, Beijing, China

*Corresponding author: zhoubo@bcu.edu.cn

Abstract. Real-time object detection is the core technology in the perception of unmanned driving. In this paper, a real-time object detection method based on YOLOv4 for unmanned driving is proposed. In the practice of unmanned driving, the illumination and large-scale change effect the visual perception of unmanned vehicles. To overcome the influence of different illumination and scale change in real scenes, we enhance the scale, illumination and other aspects of the images in the training set and the test set, and fuse with the original data. The training set and the test set in this paper use the actual road driving images taken by road traffic monitor. By these ways, more competitive vehicle automatic detection results than the original yolov4 model has been achieved.

Keywords: Unmanned Driving, YOLO, Convolutional neural network, Data enhancement, vehicle detection.

1. Introduction
With the development of artificial intelligence and the increasing innovation of science and technology, intelligent transportation has gradually entered people's vision. From the strategic point of view of realizing national economic prosperity as well as scientific and technological progress, intelligent transportation is also an inevitable development direction in the future. In order to realize driverless system or intelligent traffic monitoring system by artificial intelligence, vehicle detection is an essential technology. Target detection is a relatively simple task in computer vision, which is employed to find some specific objects in a graph. Target detection requires us to identify the types of these objects, and to mark the location of these objects. The application of target detection and target tracking technology on account of deep learning in vehicle detection also makes great object detection on the basis of convolutional neural networks(cnn) become popular in deep learning. A large amount of successful CNN architectures, like R-CNN, SSD, RFCN and YOLO. The research method based on convolution neural network performances well with less target recognition limitation, and can be widely employed in different industries. However, there are still shortcomings and defects in the application of depth learning based target detection in traffic, which requires the introduction of new technology.

Currently, deep learning models need huge data sets to achieve target detection. In order to increase accuracy, it is recommended to use images similar to the actual application scenarios as training sets,
because even for the same objectives, various background environments will have a significant impact on the results. How to obtain such huge data is often a difficult problem for deep learning applications.

A simple and general YOLO model is employed to detect vehicles on normal traffic roads. According to the influence factors that may appear in the actual driving of the vehicle, the enhancement method of data is adopted to improve the performance. The second section mainly introduces the characteristics of the YOLO model and the advantages and innovations of the yolov4. Section 3 details our processing of data sets (including data annotation and data enhancement) methods of use, as well as the characteristics of data sets. By analyzing and comparing the experimental results, the third section shows and explains the advantages and characteristics of yolov4.

2. Related Work

Yolo model. YOLO is a milestone of the target detects the acceleration of speed, although not as accurate as R-CNN series of models, it is still several times faster than other models, which makes it possible to be applied to more scenes. In 2016 the YOLO model was first raised by Joe Redmon. It is an end-to-end learning model with real-time speed. Only a single CNN network is employed in YOLO (including 24 convolutional layers and two fully connected layers). The convolution layer is responsible for feature extraction, and the fully connected layer is employed to generate bounding boxes. YOLO divides the input image into S x S grid units, and each unit has several bounding boxes for prediction. Each bounding box is composed of 5 parameters: x, y, w, h, c, where x and y represent the position of the center point of the bounding box, w and h are used to determine the size of the bounding box, and c is the confidence level. YOLO algorithm regards target detection as a regression problem, employing the mean variance loss function. YOLO network prediction uses NMS algorithms, which mainly solves the problem that a target has been detected many times.

Yolov2 is an improved version of YOLO, achieving better, faster and stronger target detection. Yolov2 has made many improvements on the YOLO basis. At the same time, a joint training method of detection and classification is proposed, and more than 9000 kinds of objects can be detected. Yolov2 has abandoned the full connection layer and employed convolution as well as anchor boxes to predict bounding boxes. Then, a k-means clustering is added, which better defines bounding boxes. Darknet-19 (including 19 convolutional and 5 maxpooling layers) have been employed as feature extractors.

Yolov3 was proposed in 2018, and in the yolov3, with a deeper network, the feature extractor is a residual model, including 53 layers of 3 x3 and 1x1 filters with jump connections, so it becomes Darknet-53. Besides, the FPN architecture is employed to optimize the bounding box prediction method, which enables yolov3 to generate 3 sets of prior boxes according to the features of different scales. Meanwhile, multi-label classification is realized by employing logical regression instead of softmax in yolov3, of which three basic components are; 1. CBL 2. Res unit 3. Res

The difference between yolov4 and yolov3 is not significant in nature. of the main five basic components of them includes: 1. CBM 2. CBL 3. Res unit 4. CSPX 5. SPP

This is an algorithm for balancing accuracy and speed. The large model has high accuracy, while the speed is too slow; the small model is fast, while the accuracy is not high. Moreover, yolov4 can fulfill the training on a common GPU (1080Ti), and can achieve real-time performance, so that it can be deployed in the production environment, and modify many SOTA methods to make single GPU training more efficient.

Yolov4 has adopted some popular methods in recent years, and finally achieved the improvement of speed and accuracy. The innovation of Yolov4 is mainly carried out in the following four aspects;

1. The Mosaic data enhancement is mainly employed at the input end (on the basis of the CutMix data enhancement mode, the two images are spliced is replaced by the four images are randomly scaled, cut and arranged).

2. Compared with DarkNet53 of YOLOv3, YOLOv4 combines several new ways of BackBone backbone network, including; CSPDarknet53, Mish continuous differentiable activation function (solving the problem of tanh and sigmoid gradient disappearance), Drop block (the enhancement of data on the feature map).
The target detection network inserts SPP modules and FPN+PAN structures between the BackBone and final output layers.

Yolov4 uses CIOU (Complete-IoU) as a Box regression function, thus ciou can achieve better convergence speed and accuracy on the BBox regression problem.

\[
GIoU = IoU - \frac{|A_c - U|}{|A_c|}
\]

Overall, the yolov4 has improved and optimized all parts of the yolov3, which is a little more complex than the v3 network, but with the same input size, the mAP has improved little but the speed grows faster. The structure of yolov4 is shown in Figure 1.

Figure 1. The structure of YOLOv4

3. Our Method

3.1. Generation of data sets

Data set employed in this experiment is DETRAC, this data set consists of four parts, which are images before light intensity change, images after light intensity change, images before stretching and images after stretching. The original images before data enhanced are from the interception of vehicle overhead in the monitoring of 24 regional roads in Beijing and Tianjin. In the data set, more than 1000 original images and the corresponding number of enhanced images are selected, which cover different time periods, different locations and different angles. The labeling program marker data set creation tool is employed to annotate the image to generate xml file, which is finally transformed into a VOC2012 data set format. The diagram contains at least one complete image of a moving vehicle, some even close to ten. These vehicles include various models, colors, sizes and angles. Images of some training and test sets are shown below.

Images of some training sets and test sets are shown in Figure 2.
3.2. Data Annotations

In daily life, there exist many sorts of vehicles, especially in the street, vehicles are different in angle and size. Therefore, there is no uniform marking standard. So, a set of specific labeling rules is made. Labeling program marking data set creation tool is employed to mark the image: the location of the car in each image is manually marked. For vehicles that only appear partially in the image and because of the distance from the lens, the vehicles that show small in the image are not marked, only the vehicles

![Images of training and test sets](image1.png)

(a) Some images from the training set

(b) Some images from the test set

**Figure 2.** Photographs of a training set
that show completely and clearly in the image and are suitable for size are marked. The concrete method is to pull the border along the longest diagonal line of the car body, cut the vehicle completely in the block diagram, name the car as the label name, and generate a corresponding xml file after all the vehicles that conform to the labeling rules are marked. After this repeated operation, all the images of the data set employed in the experiment are annotated and the xml file is generated. The annotation example is shown in Figure 3.

![Figure 3. Illustration Example](image)

3.3. Data Enhancement

For deep learning-based target detection algorithms, it is always important for a large amount of data as training phenomenon data parameters are usually employed for rich data sets. There are data parameters of several methods, such as mirroring, rotation, stretching, illumination intensity, and occlusion (clipping), etc. Different methods have different purposes. In practical applications, the characteristics of real road conditions are taken into full consideration, where lighting changes over time, vehicles are often blocked by pedestrians or buildings, and the distance between vehicles and lenses is not fixed. Thus, the main influencing factors are illumination and dimensional stretching and contrast as well as scale preprocessing the training set are made to meet the requirements of the application.

For comparison, the image with contrast of 3 by employing the written Python program to set the contrast value in advance is preprocessed. Batch operation of all images selected from the data set to do contrast preprocessing in a folder, and then the processed images are automatically stored in the folder set in the program. The contrast images as shown in Figure 4.

![Figure 4. Comparison before and after pretreatment](image)

In the same way, the image in size stretch is also preprocessed. In order to better simulate the actual situation, the stretch in a certain proportion is far enough, so it was decided to randomly change five groups of 960*540 pixels each with about 100 images of different time periods, locations and angles into five types of pixels. The specific method is to use the compiled Python program to batch process the
images in the five folders according to the five different pixel sizes set in each program, and the five groups of images are saved to the program set the save path (that is, the same folder). As shown in the figure below, the original picture with 960*540 pixels is stretched into a picture with 1400*1400 pixels according to the above method. The pre-processing contrast diagram see Figure 5.

![Pre-treatment](image1.png) ![Post-treatment](image2.png)

**Figure 5.** Comparison before and after pretreatment

Through the above process, the automatic detection of vehicles in complex road conditions is transformed into objective detection issues, which is different from the common target detection: there are three difficulties in this vehicle detection: (1) the shape of the vehicle is different, some vehicle models and styles are more difficult to detect. (2) the sample distribution is uneven. Some rare vehicle samples are obviously smaller than those of normal models. (3) because the images are mostly selected from the images of the vehicles captured by the lens, the image is often blurred by the speed of the vehicles. Therefore, it is difficult to study the automatic detection of vehicles in this paper.

4. Experiment

4.1. Experiment Setup

Pre-training employing voc2012 data sets, you can get ideal initialization grid parameters, can effectively prevent data overfitting, expedite the convergence of the network during training. Voc2012 the ratio of the training set to the validation set is 4:1. the workbench size is set to 32, epoch set to 50, the learning rate was set to 0.001, if three consecutive periods of loss do not decrease, the drop rate is 90%, if the loss of 10 consecutive epoch doesn't decrease, then stop training and save the final network model parameters. Authors and Affiliations. Laboratory environment configuration; GPU; NVIDIA RTX3070, cuda11.1, Windows 10 pro For Workstations. 16GB Memory. Experimental code uses a darknet framework.

In the experiment, the validity of the method is measured by accuracy and recall. The formulas for accuracy and recall are as follows.

\[
pre = \frac{TP}{TP + FP}
\]

\[
recall = \frac{TP}{TP + FN}
\]

The TP shows that the model prediction is accurate and the true positive number; FP represents the positive sample of the model prediction error; FN means the negative sample of the model error prediction. In other words, the accuracy rate is the ratio of the true positive number to the number of all positive data, and the recall rate is the ratio of the true positive number to the number of all target objects in the test set.
4.2. Experiment Results
Different YOLO models under the same simple training set and complex test set are compared in this paper. As it can be seen from Table 1. and Figure 6, Yolo has difficulty to cope with this special dataset. As for yolov2 and yolov3, if the recall value is low, they will show relatively good accuracy. However, as recall values rise, its accuracy began to drop rapidly, and its AP value is also low. For yolov4, the result is simple and good, that is, the accuracy value of YoloV4 remained stable in a wide range, and the AP value reached 0.8768. Figure 7 indicates results of different models.

| Model   | AP    |
|---------|-------|
| Yolov2  | 0.5401|
| Yolov3  | 0.7937|
| Yolov4  | 0.8768|

Figure 6. PR curves on car test set for different models

Figure 7. YOLOV2
The results of various pretreatment methods are displayed in Figure 10 and Table 2. It can be seen from the table that the yolo AP value is increased by 2 after stretching pretreatment, and the yolo AP value is increased by 1 after brightness enhancement pretreatment. The yolo AP value is raised to 0.9011. As can be seen from the curve in Figure 10, the non-pretreated model drops much faster than the treated model.

Table 2. AP values under different preprocessing results under YOLOV4

| Pretreatment     | AP       |
|------------------|----------|
| Null             | 0.8768   |
| Stretch          | 0.8920   |
| Contrast         | 0.8940   |
| Stretch + Contrast | 0.9011 |

The above conclusions show that when employing the yolov4 model, the data results are obviously better than the results of yolov2 and yolov3, and the V4 model can be employed in automobile detection tasks. At the same time, when the model and the original image are trained together with the original image after the brightness adjustment and stretching are added, according to the experimental results, it can be seen that the preprocessing model can improve the accuracy of model detection, on condition that the ensuring of the correct rate. The accuracy of occlusion object recognition is improved.
5. Conclusion
This paper analyzes the detection effect of common deep learning models. The task is to try to make the machine learn the features of the car, then identify the existence of the vehicle under complex conditions. The results show that the yolov4 is suitable for special training and test sets, while the yolov3 model is less effective than the yolov4 model, which shows that technology does realize the improvement of machine learning from simple memory to preliminary understanding of things. According to the actual application scene, different preprocessing methods are employed to preprocess the training set, which improves the performance and achieves good results. It shows that the accuracy of preprocessing can be improved by selecting the preprocessing method correctly according to the actual application.

YOLO is a brilliant end-to-end target detection model, of which real-time performance meets the requirements of many real-world scenarios. The experimental results can be employed as the guiding foundation for the monitoring system.

The system is more than suitable for traffic roads. The results show that the simple training set can also obtain higher accuracy in special scenarios, thus reducing the construction cost of the training set. The proposed yolov4 model can identify vehicles in various harsh environments, including but not limited to insufficient illumination, occlusion, abnormal angle, etc.

Acknowledgments
This work was supported by the teaching and research project of Beijing City University (JYC20201034).

References
[1] Yang Jinsheng, Yang Yannan, Li Tianjiao, "Traffic sign recognition algorithm based on deep separable convolution", Liquid Crystal and Display, 2019 (12).
[2] Xu Xing, Song Xiaopeng, Du Chunhui, "Verification Code Image Recognition based on Deep Learning", Journal of Test Technology, 2019 (02).
[3] Gong Jing, Cao Li, Ma Lin, Li Liangrong, "Research on yolov2 Vehicle Target Detection Method", Electronic Technology, 2018 (06).
[4] Li Chuandong, Tian Yuan, Chen Ling, Ge Junhui, "Memristive neural network learning with multiple connection models", Journal of Chongqing University, 2014 (06).
[5] Ji Liner, "Talking about the development and trend of vehicle video detection technology", Automobile and Safety, 2014 (06).
[6] Aharon Bar Hillel, Ronen Lerner, Dan Levi, Guy Raz, "Recent progress in road and lane detection; a survey", Machine Vision and Applications, 2014 (3).
[7] Redmon J, Farhadi A, "yolov3; An Incremental Improvement", IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[8] Guan Junlin, "A mask wearing detection method based on YOLOV4 convolutional neural network", Electronic technology and information science, 2020.