ISSN:2790-1688 Volume-10-(2024)

Optimizing Vehicle Detection Through YOLO-Based Deep Learning Strategies

Jietong Chen

School of Mechanical and Automotive Engineering, South China University of Technology, China

202130130255@mail.scut.edu.cn

Abstract. Vehicle detection plays a crucial role in automotive electronic systems and automated driving systems, involving the recognition of specific vehicle types on roads. Addressing the issues of low detection accuracy and slow recognition speed in existing vehicle detection methods, this paper proposes an enhanced vehicle detection model based on YOLO. To account for the varying scales of vehicles and their impact on the detection model, a normalization method is utilized to improve the calculation of prior anchor frame dimensions. Additionally, a multi-layer feature fusion strategy is implemented to enhance the network's feature extraction capabilities by eliminating redundant high-level convolutional layers. Experimental results on the validation dataset demonstrate that the proposed method achieves a mean average precision (mAP) of 90.69% and a mean frames per second (fps) of 19.1, showcasing its effectiveness.

Keywords: Vehicle Detection, Deep Learning, Autonomous Driving, Feature Fusion.

1. Introduction

The rapid advancement of self-driving car technology has spurred a competitive landscape among automakers and technology firms striving to develop and implement autonomous driving solutions. Although many self-driving vehicles remain in the testing and development phase, some are already operational on roads, capable of executing fundamental driving functions like self-parking, autonomous following, and obstacle avoidance. Notably, commercial services such as Waymo and Uber's self-driving taxi service have also emerged.

However, the widespread adoption of self-driving cars faces numerous technical hurdles, including challenges related to perception and sensor fusion, decision-making algorithms, artificial intelligence and machine learning integration, high-precision mapping and positioning systems, and vehicle control and stability. Among these challenges, the domain of visual recognition for vehicle detection plays a pivotal role and is crucial for realizing the full potential and widespread deployment of autonomous driving technologies.

Vehicle detection plays a crucial role in automotive electronic systems and automated driving technologies by enhancing driving safety, reliability, and enabling timely detection and resolution of related issues. Current vehicle detection methods encompass various approaches such as threshold-based determination, classifier-based recognition, dynamic time regularization (DTW) algorithms, traffic surveillance video analysis, and sensor-based detection. To address the challenges associated with these methods, this paper presents a target detection task specifically tailored for vehicles in and around motorways. This is achieved by integrating the YOLOv3 model with multiple training strategies and model compression techniques to enhance vehicle detection accuracy. This research focuses on applications in autonomous driving scenarios, adding significant practical value to the field. The key innovations of this study include:

Adoption of an enhanced version of the YOLOv3 model featuring a multi-layer feature fusion strategy that eliminates redundant high-level convolutional layers, leading to improved object detection accuracy.

Integration of both abstract specialized semantic information from deep networks and fine-grained pixel structure information, provides advantages in classifying and recognizing four types of vehicles.

Practical validation of the model under various conditions, including different lighting environments, backgrounds, road directions, and absence of road conditions, through rigorous testing of collected videos.

The rest of the paper is organized as follows. Section II provides an overview of previous solutions to this problem and related work. Section III describes the model used this time and briefly explains what kind of problem was solved by applying this model. Section IV focuses on the model training process and the experimental results and analyses. Section V concludes the paper.

2. Related works

Currently, common traffic detection methods can be divided into two categories: traditional methods and deep learning-based methods. Traditional methods refer to conventional machine learning algorithms. Some use the Histogram of Orientation Gradients (HOG) method to extract vehicle type features in an image, and then use Support Vector Machines (SVMs) to classify these features for vehicle detection. Although the accuracy of traditional machine learning-based methods for vehicle localization and type recognition is acceptable, these methods include very complex steps, require high human involvement, and take too much time. Therefore, these methods are not suitable for practical application scenarios.

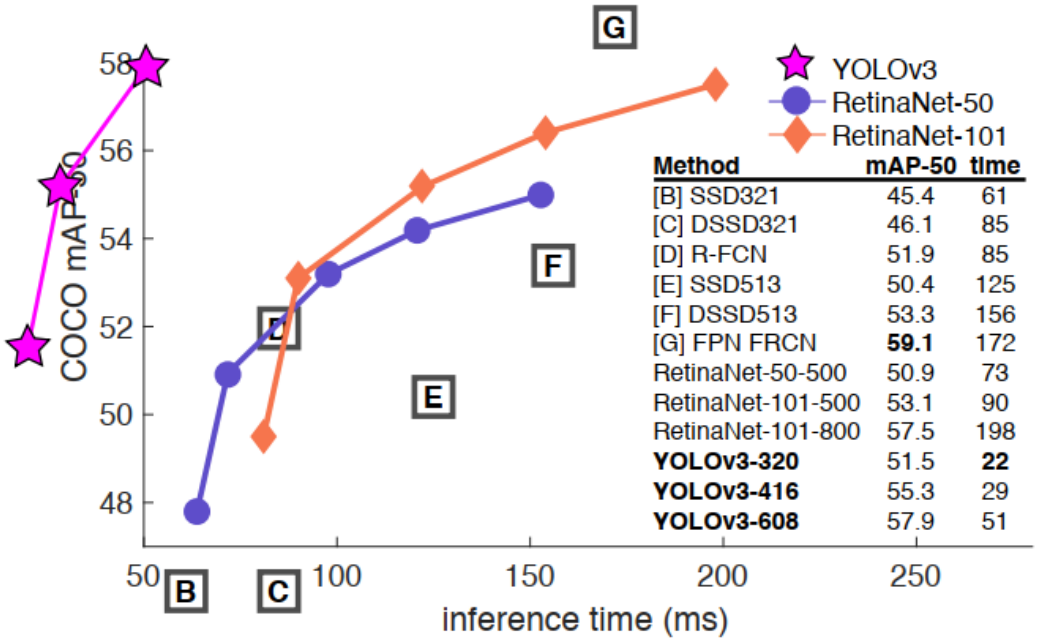
In recent years, deep learning has emerged as a promising approach for target detection and recognition, surpassing traditional methods in performance. Researchers have explored Convolutional Neural Networks (CNNs) for vehicle detection, utilizing extensively labeled vehicle images for network training without manual feature design. Additionally, unsupervised sparse coding techniques have been employed for pre-training the network, followed by softmax-based classification for vehicle categorization.

The evolution of deep learning object detection methods began with R-CNN, incorporating region-based suggestions through selective search. Subsequent advancements like SPP-net, Fast R-CNN, Faster R-CNN, and R-FCN refined region-based approaches. Despite their effectiveness, these methods often suffer from slow processing speeds and suboptimal detection accuracy, necessitating further improvements.

To address these limitations, Redmond et al. introduced YOLO as an end-to-end object detection method, converting direct object detection into regression tasks. YOLOv2, an improved version, significantly enhanced detection speed while maintaining accuracy, making it suitable for detecting diverse categories with distinct differences like people, horses, and bicycles. However, for vehicle detection, which often relies on localized features like tires and headlights, YOLOv2's general approach may not be optimal. Therefore, specialized vehicle detection models such as YOLOv2 Vehicle were proposed, incorporating k-means clustering for anchor box selection. Despite these advancements, challenges persist due to variations in road conditions and video viewpoints, leading to issues such as low detection accuracy and slow processing speeds.

3. Method

The YOLOv3 model employs a Convolutional Neural Network (CNN) to process images, dividing them into N x N grids, with each grid responsible for detecting targets whose centroids lie within it. Each grid cell in YOLOv3 predicts three boxes, each defined by parameters such as (x, y, w, h, confidence), and depending on the dataset, it also predicts probabilities for 20 or 80 categories. This model demonstrates advantages in mean average precision and single-frame detection time at iou=0.5 compared to other models, as illustrated in Figure 1.



Method	mAP-50	time
[B] SSD321	45.4	61
[C] DSSD321	46.1	85
[D] R-FCN	51.9	85
[E] SSD513	50.4	125
[F] DSSD513	53.3	156
[G] FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

Fig. 1 Comparison of YOLOv3 with other models [7]

Here, we employ the YOLOv3-SPP-ultralytics model to address the aforementioned issue of slow detection accuracy. To enhance the algorithm's accuracy in detecting small targets, we incorporate FPN-like upsample and fusion techniques in YOLOv3. This involves performing detection on feature maps of multiple scales, with the final fusion incorporating three scales (the other two scales being 26×26 and 52×52, respectively).

The YOLOv3-SPP model utilizes DarkNet-53 as its backbone network for feature extraction. DarkNet-53 employs continuous 3×3 convolution and 1×1 convolution, comprising a total of 53 layers with associated weights, as depicted in Fig. 2.

	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. 2 Schematic diagram of DarkNet-53

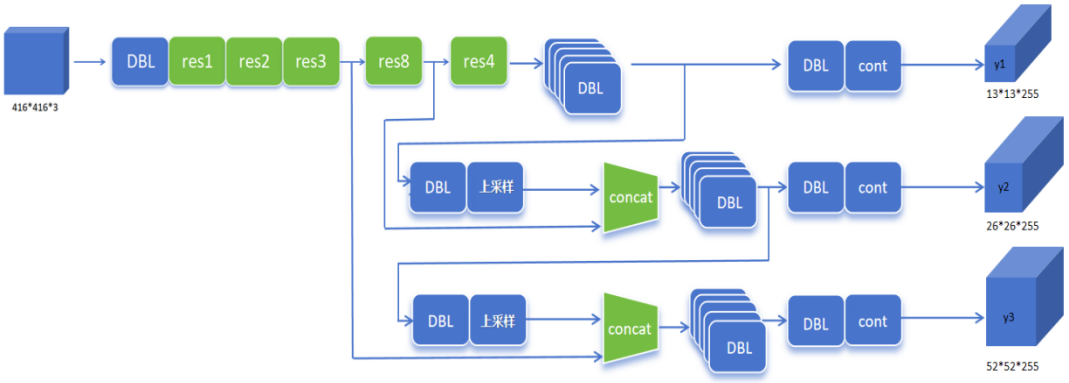


Fig. 3 Structure of yolov3 network.

The YOLOv3 network structure, depicted in Figure 3, underwent modifications in our study. We replaced the YOLOv3 skeleton network with ResNet50-VD to enhance speed and accuracy compared to the native DarkNet53 network. Moreover, we explored different backbone network structures such as ResNet18, 34, and 101 to suit various scene scenarios. To achieve a balance between speed and accuracy, Deformable Convolution was employed instead of continuous 3x3 convolution in the Stage5 segment of the backbone network within our augmented YOLOv3 model. Additionally, the DropBlock module was integrated into the FPN section to enhance model generalization.

The loss function used in this model is as follows:

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} II_{ij}^{obj}\left[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2\right]$$
$$+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} II_{ij}^{obj}\left[\left(\frac{\omega_i-\hat{\omega}_i}{\hat{\omega}_i}\right)^2+\left(\frac{h_i-\hat{h}_i}{\hat{h}_i}\right)^2\right]$$
$$+\sum_{i=0}^{S^2}\sum_{j=0}^{B} II_{ij}^{obj}\left(C_i-\hat{C}_i\right)^2$$
$$+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathrm{II}_{ij}^{noobj}\left(C_i-\hat{C}_i\right)^2$$
$$+\sum_{i=0}^{S^2}\mathrm{II}_{ij}^{obj}\sum_{c\in classes}\left(p_i(c)-\hat{p}_i(c)\right)^2 \quad (1)$$

Of these, the x_i and y_i are the centre coordinates of the box of the ith grid cell, and ω_i and h_i are the width and height of the box of the ith grid cell, and C_i is the confidence level of the box of the ith grid cell, and $p_i(c)$ is the class probability of the box of the ith grid cell. The λ_{coord} denotes the weight of coordinates lost, and λ_{noobj} denotes the weight of the bounding box without object loss. Finally, the S^2 denotes the S×S grid cell, and B denotes the box, and II_i^{obj} denotes whether the object is located in cell i, and II_{ij}^{obj} denotes the jth box predictor in cell i that is "responsible" for the prediction. In Equation (1), the first line computes the coordinate loss, the second line computes the bounding box size loss, the third line computes the bounding box confidence loss with objects, the fourth line computes the bounding box confidence loss without objects, and the last line computes the class loss.

We transform the target detection problem into a regression problem and train it using the regression loss function. We Train our prediction frames to approximate the correctly labeled frames using the regression equation loss function.

$$L(o,c,O,C,g)=\lambda_1 L_{conf}(o,c)+\lambda_2 L_{cla}(O,C)+\lambda_3 L_{loss}(l,g) \quad (2)$$

$\lambda_1 L_{conf}(o,c)$ is the confidence loss, the $\lambda_2 L_{cla}(O,C)$ is the classification loss, and $\lambda_3 L_{loss}(l,g)$ is the localisation loss, and λ_1、λ_2、λ_3 is the balance coefficient.

4. Experimental procedure and analysis

4.1 Dataset

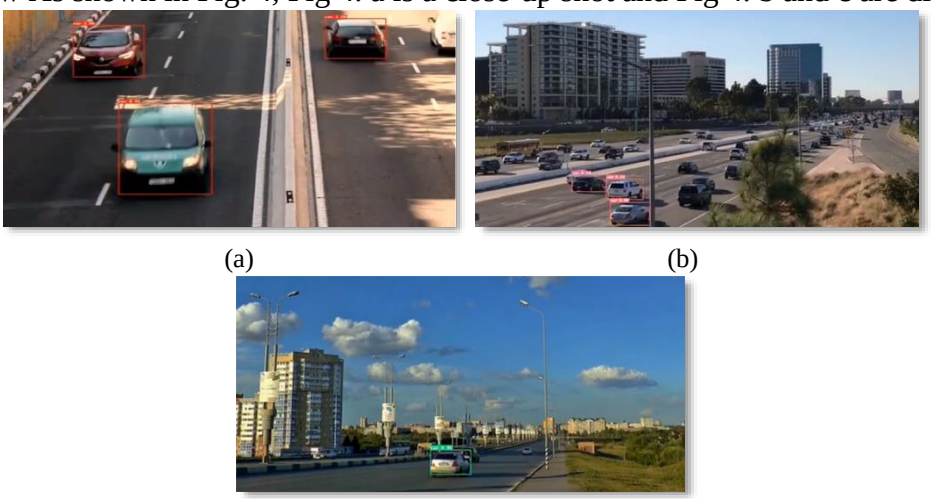
This paper utilizes the UA-DETRAC dataset [8] for vehicle detection and tracking experiments. This large-scale dataset is primarily collected from road overpasses in Beijing and Tianjin (Beijing-Tianjin-Hebei scenarios) and meticulously annotated with 8,250 vehicles and 1.21 million target object out-frames. The vehicles are categorized into four types: cars, buses, vans, and other vehicles. Weather conditions are classified into four categories: cloudy, nighttime, sunny, and rainy. Consequently, this dataset serves as a rigorous benchmark for real-world multi-target detection and tracking tasks.

The dataset comprises 10 hours of video footage captured using a Canon EOS 550D camera at 24 distinct locations across Beijing and Tianjin, China. The videos were recorded at 25 frames per second (fps) with a resolution of 960×540 pixels. The dataset encompasses a training set of 55,817 images, a validation set of 9,851 images, and a test set of 16,417 images, totaling 82,085 images for comprehensive evaluation and analysis.

4.2 Model training and results

This experiment is divided into a training stage and a testing stage. in the training stage, YOLOv3 clusters the dataset and then uses the center of the clusters as anchor boxes (a priori boxes). Each anchor box predicts four values related to the coordinates. The most suitable anchor size (center of nine clusters) is calculated by k-means clustering. Our training set is used to train our anchor boxes to be closer to the correct box of the training set, thus forming our desired prediction box. In the testing phase, the data frames from the test set are processed and put into a convolutional neural network to generate predictions, which are then parsed, filtered for thresholding using non-maximal suppression (threshold iou=0.5) and visualized and finally spliced into a video.

Effect Show As shown in Fig. 4, Fig 4. a is a close-up shot and Fig 4. b and c are distant shots.



(a) (b)

(c)

Fig. 4 Effect display diagram

4.3 Analysis and Extensible Applications

The experimental results from the validation dataset demonstrate the effectiveness of the method, achieving a mean average precision (mAP) of 90.69% and a mean frames per second (fps) rate of 19.1. Leveraging feature recognition and melting techniques, it enhances recognition accuracy, particularly in detecting and recognizing moving vehicles within the traffic system. This capability

finds application in various domains such as traffic flow detection, statistics on violated vehicles, vehicle tracking, and apprehension.

Traffic flow monitoring is essential for effective traffic management, control, and information dissemination, directly impacting overall road section operation and management. Current traffic flow monitoring technologies include coil detection, geomagnetic detection, microwave detection, and video detection. Video detection, offering non-contact, high precision, and real-time monitoring advantages, holds significant potential in urban traffic flow monitoring applications.

Illegal parking poses a common challenge contributing to road congestion and traffic order disruptions. Many cities are grappling with a significant number of illegal parking incidents, necessitating measures like increased patrols, electronic eye systems installation, and higher fines. The conclusions drawn in this study aid in identifying illegal parking through road vehicle monitoring, facilitating better traffic regulation and order maintenance.

In vehicle pursuit scenarios, real-time monitoring and tracking of suspect vehicles assist law enforcement in quickly locating and apprehending suspects during pursuit incidents. Additionally, implementing video surveillance at community entrances and exits enhances community security and management efficiency by swiftly identifying vehicles. Analyzing mainline traffic parameters and combining them with ramp traffic queue length enables traffic light-controlled regulation of traffic flow onto highways.

5. Conclusion

In this paper, we utilized an enhanced vehicle detection model algorithm that improves upon the YOLOv3 model. This improvement includes optimizing anchor boxes through k-means++ clustering, improving the loss function with normalization to achieve scale consistency, and designing a multi-layer feature fusion network to enhance feature extraction capabilities. Experimental results demonstrate the effectiveness of this model, with a mean average precision (mAP) of 90.69% and an average frames per second (fps) of 19.1. The enhanced recognition accuracy of the model addresses challenges in traffic detection systems, making it applicable in multiple areas such as traffic flow detection, violation vehicle statistics, and vehicle tracking. However, future improvements are needed to optimize convergence speed and training efficiency, achieving better scalability and wider application possibilities.

References

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

[2] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[3] Purkait P, Zhao C, Zach C. SPP-Net: Deep absolute pose regression with synthetic views[J]. arXiv preprint arXiv:1712.03452, 2017.

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

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

[6] Sang J, Wu Z, Guo P, et al. An improved YOLOv2 for vehicle detection[J]. Sensors, 2018, 18(12): 4272.

[7] https://arxiv.org/abs/1804.02767

[8] https://detrac-db.rit.albany.edu/

[9] Wu S, Li X, Wang X. IoU-aware single-stage object detector for accurate localisation[J]. Image and Vision Computing, 2020, 97: 103911.