Published online: 31 August 2024 DOI: 10.54254/2977-3903/10/2024105

What role does object detection play in autonomous driving?

Yijia Han

Xi'an Jiaotong University Affiliated Middle School, No. 99 Yanxiang Road, Yanta District, Xi'an

hyj20070218@163.com

Abstract. Deep learning-based object detection algorithm is becoming more and more important in autonomous driving area with an increasing amount and trending these days. This article first provides definitions and introduces an autonomous driving and object detection. Subsequently, a detailed discussion is conducted on object detection, comparing traditional object detection methods with deep learning object detection algorithms. The shortcomings of traditional methods highlighted the advantages of deep learning-based object detection algorithms, laying the groundwork for the use of deep learning object detection algorithms in the following text. Finally, several detection objects and detection scenarios are introduced. The detection objects are divided into different parts, including moving targets, stationary targets, and infrared targets. Moving targets such as pedestrians and vehicles, while stationary targets include traffic signs and lanes. The detection scenarios are classified into ordinary scene detection and complex scene detection. In the discussion section, the commonly used datasets for autonomous driving target detection training are listed first, such as the KITTI dataset, the COCO dataset, and so on. Subsequently, a discussion was conducted on algorithms, mainly focusing on the models and features in one stage and two stages. For different types of algorithms, this article discusses the advantages and disadvantages of the algorithms. When judging the superiority or inferiority of algorithms, there are usually two aspects: detection accuracy and detection speed. FPS is a commonly used indicator for detection speed, and detection accuracy mainly covers five aspects, namely accuracy, precision, recall, AP (average precision), and MAP (mean average precision). Finally, how the improved algorithms are applied and solve the existing problems is discussed.
Keywords: object detection, autonomous driving, deep learning

1. Introduction

Car is becoming more and more popular these days and has become the preferred choice for people to travel. As technology develops, the performance of cars is steadily improving, and cars play an important role in our lives. But due the increasing number of cars, it has also caused some negative consequences, such as road congestion during peak hours, frequent traffic accidents, and so on. As a high-tech, the development level of autonomous driving is directly related to the international competitiveness of the automotive industry in various countries and the global industrial division of labor. Therefore, many countries around the world attach great importance to the development of autonomous driving. Autonomous driving technology can effectively avoid traffic accidents caused by human factors and build a stable and harmonious traffic order. Its technology relies on environmental perception, positioning and navigation, trajectory planning, and collaborative execution control, enabling computers to automatically and safely operate vehicles. Autonomous driving technology can not only change our way of travel and provide convenience for our lives but also promote the development of various industries and improve the country's GDP, which is of great benefit. But if the autonomous vehicle wants to run smoothly and safely on the road, it must make timely and accurate judgments about the surrounding environment and obstacles, which is inseparable from efficient and accurate target detection. In practical scenarios, accurate and accurate object detection poses high difficulties, as complex environmental backgrounds and changes in lighting under different weather conditions have a significant impact on object detection. Therefore, these are also urgent issues that need to be addressed in the process of object detection.

2. Research review

2.1. Autonomous driving

Autonomous driving technology relies on environmental perception, positioning and navigation, trajectory planning, and collaborative execution control [1] to enable computers to operate vehicles automatically and safely. As the "eye" of the auto drive system, environment perception plays a very important role and position, and needs to be more accurate and real-time. The main task of the environmental perception system is to safely and quickly detect the position and trajectory of other vehicles, recognize the action intention and posture of pedestrians, and combine it with environmental models. The perception system mainly utilizes sensor technology, positioning technology, and vehicular communication technology to proceed with environmental perception. Environmental perception provides information about the external environment for decision-making in autonomous driving, which is a prerequisite and safety guarantee for the implementation of autonomous driving technology. Object detection is an extremely important part of environmental perception.

2.2. Object detection

Early object detection often uses vision to obtain information. Vision-based object detection methods first selected candidate regions, then extracted features from the candidate regions and determined the type of target, including traditional vision-based object detection methods and deep learning-based object detection methods. Object detection, which accurately finds the positions of various objects in an image and labels the categories of such objects, may seem relatively simple, but it is not easy for computers to solve it because the placement of object angles, changes in size range, and so on may all affect the recognition of the computer. Early object detection often uses visual acquisition of information. Visual-based object detection methods first select candidate regions, then extract features from the candidate regions and determine the target type, including traditional visual-based object detection methods and deep learning-based object detection methods [2]. Real-time object detection based on vision is easily affected by complex environmental backgrounds and changes in light under different weather conditions. Therefore, accurate detection and recognition of road traffic participants still face very serious challenges [3] in the actual detection process.

2.2.1. Traditional object detection methods

Traditional object detection methods refer to image segmentation based on geometric and statistical characteristics, mainly divided into four stages: image preprocessing, target region selection, feature extraction, and classifier classification. Traditional methods have certain shortcomings, such as lacking specificity and poor robustness in the region selection strategy of sliding windows. Because the traditional target detection algorithm is based on the manual selection of feature areas, there are many problems such as a large amount of calculation, poor detection effect, weak generalization ability, and so on. It is difficult to meet the requirements of high reliability and real-time autonomous vehicles [3]. To address these issues, deep learning methods have appeared.

2.2.2. Deep learning object detection methods

Deep learning object detection methods can be subdivided into two subcategories, namely one stage and two stages. One stage is mainly an algorithm based on regression models [3]. Since 2014, significant breakthroughs have been made in object detection, mainly represented by YOLO, which transforms object detection into a regression problem. It can simply be regarded as designing a network and training on multiple networks and finally obtaining the model The datasets used to test algorithm performance include PASCAL VOC, ImageNet, MS COCO, and other datasets. Deep learning object detection methods have improved many problems existing in traditional methods, so this article mainly discusses deep learning object detection methods.

2.3. Autonomous driving target detection

The core of deep learning-based object detection algorithms is to automatically predict the target categories and location information of interest in input images or videos through detection algorithms, and its functionality matches the demand for autonomous driving. By inputting a large number of autonomous driving scene images, the detection model is trained to obtain the ability to recognize target information such as pedestrians, vehicles, and traffic signs in autonomous driving. However, in autonomous driving scenarios, target occlusion, changes in light intensity, and so on can have a significant impact on detection algorithms. Therefore, designing target algorithms that can shield external interference information and quickly and accurately identify target types and information plays an important role in the development of autonomous driving technology. Autonomous driving target detection can be divided into moving targets, stationary targets, and infrared targets according to different detection objects.

2.3.1. Moving object detection

Motion object detection refers to the detection of moving targets during the driving process of a car, which generally includes vehicles, pedestrians, etc. However, there are several difficulties in detection: 1) There are various degrees of occlusion of the target to be detected; 2) The direction and speed of the target's movement are both uncertain; 3) The detection effect for small targets is not good; 4) The detection speed and accuracy often cannot meet [4] simultaneously. According to the degree of occlusion of the detected object, occlusion is divided into three levels: 0 is unobstructed; 1% to 45% are partially occluded; When it is greater than or equal to 45%, it is considered severe occlusion [5]. Based on the problem of occlusion in detecting targets, Jia Shijie proposed an anchor-free end-to-end object detection algorithm based on improved CenterNet (CB CenterNet) [5]. Firstly, based on the feature extraction network, a differential cascaded feature fusion structure with enhanced composite connection modules is designed, and an attention module with channel and spatial cross dimensional interaction is added to the residual block to guide the network model to focus on occluded targets; Secondly, an improved Gaussian sampling module with controllable Gaussian kernel shape is added to the classification and regression branches to accelerate the training speed of the network [5]. The improved method is very effective which brings a significant improvement in mAP (mean accuracy) and a 30% increase in average speed. Traditional motion object detection methods are mainly based on videos, which detect targets through changes in image information between frames [4]. This method is not effective for relatively static target detection. If the autonomous vehicle is in the parking state, it is easy to cause more missed detection [4]. Li Yunpeng, Hou Lingyan, and Wang Chao (2019) proposed a new method based on such problems - the improved YOLOv3 autonomous driving target detection method [4]. Design a more reasonable loss function based on YOLOv3 to address issues such as low accuracy and slow speed in motion target detection in autonomous driving scenarios. Clustering the target bounding boxes in the dataset using the K- means algorithm, and training an improved motion object detection model through a mixed dataset. Compared with current mainstream object detection models, this model has significantly improved performance and performs well in detecting moving targets in various complex traffic scenes.

Moreover, the average accuracy and detection speed of this model on the test set reached 88.55% and 35FPS9 respectively, meeting the requirements of real-time detection.

1. Pedestrian detection

Pedestrian detection is a benchmark problem in object detection and has significant practical value in scenarios such as autonomous driving. It plays an important role in path planning and intelligent obstacle avoidance [6]. Due to limitations in algorithm power consumption and operational efficiency, pedestrian detection in autonomous driving scenarios has problems such as slow detection speed and low accuracy in detecting occluded pedestrians. To address these issues, Ma Yan, Xu Xinkai, and others proposed a method called Scale Attention Parallel Detection Algorithm [6], which effectively solved the problem of pedestrian occlusion and reduced the pedestrian missed detection rate, also improved the robustness of the detector.

2. Vehicle inspection

Accurately and quickly identifying the vehicles ahead is crucial in autonomous driving scenarios. Shi Jianting, Li Xu, and others proposed a YOLOv4 Efficient object detection algorithm that improves detection speed while ensuring detection accuracy [7]. While reducing the number of model parameters, it effectively improves the detection speed of the object detection algorithm. In response to the problem of poor performance in detecting small targets and resulting in low vehicle accuracy, researchers such as Chen Yanfei and Yan Zhangchen proposed a method called improved the YoLov4 algorithm. Firstly, add a small object detection layer to reduce the missed detection rate of small target vehicles. Then, replace the CIoU loss function with the EIoU loss function to reduce the bounding box regression loss of the algorithm and improve its detection accuracy [8]. The improvement method can effectively solve the problem of poor performance in small object detection and improve detection accuracy.

2.3.2. Static object detection

The detection of moving targets is achieved through the changes in image information between frames, so it is difficult to quickly, accurately, and quickly identify moving targets. So here comes the question, would detecting stationary targets that are opposite to moving targets be easier? The answer is no. Static targets such as traffic signs and lane markings. There is a certain difficulty in detecting traffic signs, as small target traffic signs contain very little information, making detection difficult and with low accuracy. For the detection of lane markings, as they are edge lines, deviations are prone to occur during detection.

1. Traffic sign detection

In real-world scenarios, detecting small target traffic signs is of great use for autonomous driving. Due to the fact that small targets are far away from the vehicle camera and occupy a relatively small proportion of the image, there are some problems with small target traffic sign detection. Small targets contain very little information, making detection difficult. In response to the problem of low accuracy of small target traffic signs, researchers such as Hu Junping and Wang Hongshu proposed an improved YOLOv5 small target detection algorithm, which often uses multi-scale feature fusion, incorporating attention mechanisms, improving feature extraction networks, and increasing small target resolution to improve detection accuracy [9]. It is done by applying the idea of cross stage local networks to the spatial pyramid pooling module, adding gradient paths on top of this module enhances the feature aggregation ability of the backbone network; Propose an adaptive feature fusion algorithm that multiplies the fused features at different levels with learnable adaptive weights, suppresses the interference of deep features on small target features, improves the detection accuracy of small target traffic signs [9], and has excellent detection performance. In response to the problem that machine vision cannot simultaneously meet the requirements of accuracy and real-time performance in recognizing road traffic signs from the driver's perspective, researchers proposed the method of using deep learning YOLO v2 model [10]. This method divides the image into grid form, predicts the probability of bounding boxes and traffic sign categories in each grid area, and finally determines the category and position of traffic signs through non maximum suppression. The feature extraction and classification are integrated into the same neural network operation process. Real-time and robustness have been greatly improved, which can well meet the demands of autonomous driving [10].

2. Lane detection

Lane detection is used to extract the edge lines of roads, for the positioning and correction of autonomous vehicles, and is an important prerequisite and guarantee for smooth and safe driving [11]. Researchers Zhang Kaixiang and Zhu Mingji designed a multitasking autonomous driving environment perception algorithm in YOL0v5, which can simultaneously handle three tasks: object detection, drivable area segmentation, and lane detection. The algorithm achieved 76.3 fps on a 1080Ti graphics card and achieved the best performance compared with other multitasking networks on the BDDl00K dataset [11], while achieving multitasking autonomous driving environment perception algorithm.

2.3.3. Infrared target detection

Infrared image target detection is widely used in the field of autonomous driving due to its characteristics of long detection distance, strong anti-interference ability, and easy feature extraction. Lin Jian, Zhang Weiwei, and others proposed an infrared target detection algorithm based on YOLOv5 (YOLOv5 IF), which is structurally divided into three parts: feature extraction (Backbone), feature integration (Neck), and detection head (Head). The input infrared image is processed through a feature extraction network to obtain semantic information about the target at different scales, and the feature map is divided into grids of different sizes [12]. At the same time, after the infrared image is processed through a feature extraction network, the feature information is mapped onto feature maps of different scales. There are a large number of small targets in the infrared image, and excessive sampling rates can make it difficult to reflect the target features on the feature map after multiple sampling [12]. This method successfully optimized some inherent problems such as inaccurate accuracy and parameter quantities.

2.4. Autonomous driving scene detection

For autonomous driving, with the continuous research and innovation of object detection algorithms, universal object detection algorithms can basically meet the problem of object detection in ordinary traffic scenes. However, in complex traffic scenes, there are problems such as a large amount of target occlusion and difficulty in meeting the accuracy requirements of small object detection. Real- time traffic scene object detection is a prerequisite for achieving functions such as electronic monitoring and autonomous driving [14]. In order to improve the detection performance and robustness of the algorithm model. Wen Haiming and Tong Mengjun proposed a semi-supervised learning-based TmsDet (Transfomer Detection) object detection algorithm for autonomous driving scenarios. Firstly, an MSAMark structure is proposed, which utilizes a self- attention mechanism to capture global information and combines convolutional networks to model local features. The algorithm is applied at the end of the feature extraction network and the end of the small object detection head, which enhances the ability of algorithm models to capture remote dependency relationships and enrich contextual information; Secondly, a position attention- weighted feature fusion network LAFFN is proposed for capturing local position and channel information in different feature fusion layers, enhancing the multi-level feature weighted fusion ability and network feature representation ability, improving the attention to small target areas, alleviating the impact of drastic changes in target scale, and improving the detection performance of the model; Finally, this article proposes an efficient object detection algorithm training framework EODS based on semi-supervised learning, which further improves the detection performance and robustness of the TrallsDet algorithm model without increasing the number of parameters and computation [13]. In response to the low detection efficiency of existing object detection algorithms and the problem that most lightweight object detection algorithms have low model accuracy and are prone to false detection and missed targets, Gu Deying et al. improved the YOLOv5 object detection algorithm for model training, which optimized the training process using pseudo label strategy, and then merged the labels into three categories on the KITTI traffic target dataset, Test the trained model [14]. The experimental results show that the improved YOLOv5 final model achieved a mAP of 92.5% on all categories, which is 3% higher than the original YOLOv5 trained model [14]. Finally, the trained model is deployed on the JetsonNano embedded platform for inference testing, and TensorRT is used to accelerate inference. The average inference time per frame is measured to be 77ms, which can achieve real-time detection of targets [14].

3. Discussion/ Development

3.1. Datasets

The dataset is the foundation for analyzing the performance of deep learning neural network models.

The dataset is an indispensable part of the development of autonomous driving technology, and high-quality datasets can often greatly promote algorithm development. When evaluating the superiority or inferiority of an algorithm, it also needs to be done on the same dataset to make sense.

3.1.1. KITTI dataset

The KITTI dataset is a visual algorithm evaluation dataset used in autonomous driving scenarios, co-founded by Karlsruhe Institute of Technology (KIT) in Germany and Toyota University of Technology Chicago (TTIC) [15]. The KITTI dataset includes scenarios such as urban areas, rural areas, and highways. The KITTI dataset [16] was released in 2011, and one of the main obstacles to the application of visual perception systems in the field of autonomous driving is the lack of suitable benchmarks. However, the existing datasets differ significantly from actual requirements in terms of both data volume and collection environment. So they used their own autonomous driving platform to establish a huge dataset based on real scenes, in order to promote the development of computer vision and robot algorithms in the field of autonomous driving. The KITTI dataset consists of 389 pairs of stereo images and optical flow maps, with a visual ranging sequence of 39.2km, image compositions of 3D annotated objects exceeding 200k, a 10Hz sampling frequency, and 3D object detection categories (car, van, truck, pedestrian, pedestrian, cycle, misc) [15].

3.2. COCO dataset

The COCO dataset [17] is a large-scale dataset that can be used for image detection, semantic segmentation, and image title generation. It has over 330K images, including 1.5 million targets, 80 target categories, and 91 material categories, with each image containing five sentence descriptions of images, and 250000 pedestrians with keypoint annotations. The COCO dataset can be used in many areas, including object detection, dense pose estimation, key points detection, Stuff Segmentation, Panoptic Segmentation, and image captioning [18].

3.2.1. Mapping traffic sign dataset

The Mapillary Traffic Sign Dataset [19] is the world's largest and most diverse public traffic sign dataset used for detecting and classifying traffic signs from all over the world. Diversity: Covering a global geographic range of various weather, seasons, and times of the day, including urban and rural roads, images, and traffic sign categories, covering six continents.

3.2.2. Tsinghua Tencent 100K tutorial [20]

Tsinghua University has created a large-scale traffic sign benchmark based on 100000 Tencent Street View Panorama images which provided 100000 images containing 30000 instances of traffic signs. These images cover significant changes in lighting and weather conditions. Diversity: covering data under different weather and lighting conditions.

3.2.3. German traffic sign dataset [21]

This dataset was provided in a multi-category classification competition held at the 2011 IJCNN, providing traffic signs in diverse backgrounds such as distance, lighting, weather conditions, and partial occlusion. The dataset includes 43 categories, and the frequency of occurrence of different categories is not balanced. Participants must classify two test sets, each with over 12500 images.

3.3. Algorithm discussion

Neural networks simulate the nervous system of the human brain, forming a network with functions such as learning, association, memory, and pattern recognition, also known as artificial neural networks. However, due to limitations in computing power at the time and the lack of large-scale training data, the research boom in neural network algorithms gradually cooled down. In recent years, with the emergence of large-scale open-source datasets and the rapid development of the computer industry, research on deep learning-based object detection algorithms has made rapid progress. It can be divided into candidate box based two-stage algorithms represented by the R- CNN series and regression-based one-stage algorithms represented by the YOLO series [25].

3.3.1. One stage

The One stage algorithm is mainly represented by the YOLO series. The One stage algorithm does not need to generate candidate regions, but directly generates the class probability and position coordinate values of objects. Only a single detection is needed to directly achieve the final detection result. The One stage algorithm has a fast detection speed.

1. YOLO model

Before YOLOv1 was proposed, the R-CNN series algorithms dominated the field of object detection. Although the R-CNN series has high detection accuracy, its detection speed could not meet real-time requirements due to its two-stage network structure [22]. YOLO, like the basic neural network structure, consists of convolutional layers, pooling layers, and fully connected layers [23].

In 2015, REDMON et al. [24] proposed a regression based one stage object detection algorithm YOLOv1 (You Only Look Once), which directly uses a CNN to complete classification and regression tasks [25]. YOLOv1 divides the input image into A × A grid, if the center of a target falls within a grid, then this grid is responsible for predicting the target [25]. For each grid, YOLOv1 predicts B bounding boxes, each containing 5 predicted values: the position information of 4 bounding boxes and the confidence information of 1 bounding box. YOLOv1 eliminates the operation of generating candidate boxes, thereby reducing the computational complexity of the model, improving detection speed, and achieving real-time detection [25]. However, the output layer of YOLOv1 is a fully connected layer, so the algorithm only supports the same input scale; Secondly, although each grid can predict B bounding boxes, only the bounding box with the highest IoU is ultimately selected as the prediction result, meaning that each grid is only responsible for predicting one object. When the target distribution in the detection image is relatively dense, it can easily cause missed detections and low detection accuracy [25]. Based on the problem of YOLOv1, scholars have studied and proposed YOLOv2, continuously updating and iterating to fill in gaps. To this day, the latest YOLO model is YOLOv5.

2. SSD model

The SSD series can be considered as a combination of FasterRCNN and YOLO, using a regression based model to directly regress the class and position of objects in a network, resulting in fast detection speed. At the same time, the region based concept is utilized in the candidate process, using many candidate regions as ROIs [26]. The SSD (Single Shot Multibox Detector) proposed by LIU et al. [27] in 2016 is a single-stage objective algorithm that eliminates the fully connected layer and all dropout layers on the basis of VGG16, and adds 5 convolutional layers to obtain feature maps of different sizes [25]. The main idea is to uniformly perform dense sampling with different scales and aspect ratios at different positions of the feature map, while performing object classification and regression of prediction boxes. The entire process only requires one step, but uniform dense sampling can cause an imbalance in the proportion of positive and negative samples, making the training process more difficult and reducing the accuracy of model detection. At the same time, adding convolutional layers indirectly deepens the depth of the model, resulting in the loss of location information for small targets in deep feature maps, which reduces the detection accuracy of the SSD algorithm for small targets [25].

3.3.2. Two stages

Two stages is a type of basic deep learning object detection algorithm. The target detection process is mainly carried out through a complete convolutional neural network, so CNN features will be used to extract the description of the features of candidate area targets through the convolutional neural network. The main representative is: R-CNN to fast RCNN. Compared with traditional algorithms, the two-stage algorithm has a significant improvement in accuracy. Although its speed is slightly inferior to one stage, its detection accuracy is high.

1. R-CNN

In 2014, Girshick R et al. (2013) proposed the R-CNN model. R-CNN first uses a selective search algorithm to extract approximately 2000 candidate boxes, then scales the extracted candidate regions to the same size and inputs them into a convolutional neural network for feature extraction. Finally, the extracted feature maps are input into an SVM (Support Vector Machine) classifier for classification Input to the fully connected network to obtain the target location information [25]. The disadvantage of R-CNN is that it causes image distortion when scaling candidate regions, with 2000 candidate regions input into the neural network, resulting in high computational complexity. The training steps are all run separately, also the intermediate data requires a large amount of hard disk space for storage [25]. Therefore, He Kaiming et al. analyzed and improved the shortcomings of R-CNN by proposing SPP Net [29], which adds a Spatial Pyramid Pooling (SPP) layer between the last convolutional layer and the fully connected layer.

This can pool feature maps of different scales and generate fixed size feature maps, avoiding image distortion. However, the various training processes of SPP Net are still running separately, and the intermediate data requires a large amount of hard disk space for storage [25].

2. Fast R-CNN

In 2015, Ross et al. drew inspiration from SPP and proposed Fast R-CNN [30], replacing SPP with RoI Pooling (Region of Interest Pooling) and SVM classifier with softmax classifier [25] by inputting the entire image to be detected into the backbone network VGG16 to extract convolutional feature maps, and using selective search algorithms to extract candidate regions from the input image; Then perform RoI Pooling on each candidate region on the convolutional feature layer to obtain fixed scale features; Finally, classification and regression are performed through a fully connected layer [25]. Fast R-CNN proposes a Multi-task Loss function, which simultaneously solves the problems of classification and position regression.

Compared to R-CNN and SPP Net, The training process of Fast R_ CNN is no longer distributed, reducing the occupation of hard disk space, but external algorithms are still needed to generate candidate boxes [25].

3. Faster R-CNN

In 2015, Ren et al. proposed Fast R-CNN [31], which replaced the selective search algorithm [25] in Fast R-CNN with Region Proposal Network (RPN). RPN maps anchor boxes with multiple sizes and aspect ratios to the original image by performing sliding window operations on each point on the feature map to obtain candidate regions [25]. By setting a predetermined IoU threshold, the generated candidate regions are distinguished into positive and negative samples, completing the training of the RPN network for coarse classification and localization of targets in the image [25]. Faster R-CNN performs the process from extracting original image feature maps, obtaining candidate regions to final classification and bounding box regression in the neural network, and the entire training process can share the feature information extracted by the convolutional neural network, improving the computational speed of the model [25]. However, due to the fact that Faster R-CNN only uses the last feature layer for prediction, a lot of low-level position information is lost, resulting in poor detection performance of the network for small targets. Moreover, the network is still divided into two stages: extracting candidate regions and classifying, which makes the detection speed of the network unable to meet the requirements of real-time detection [25].

3.4. Advantages and disadvantages of algorithms and areas for improvement

3.4.1. Algorithm superiority and inferiority

When judging the superiority or inferiority of algorithms, we often use detection speed and detection accuracy to measure them. The detection speed is usually represented by FPS (frames per second). FPS is a performance evaluation index for object detection, where detection speed represents the computational performance of the object detection algorithm and model. FPS refers to the number of frames processed per second of the image, with higher values indicating faster detection speed [32]. The detection accuracy is mainly judged from several aspects which are accuracy, precision, recall, AP (average precision), and MAP (mean average precision).

1. Precision

Accuracy is the probability of a value being detected correctly among all detected targets [32].
Precision=TP/TP+FP
Accuracy is defined from the perspective of prediction results. Accuracy applies to all samples, but Precision only applies to the portion of samples detected (including false positives) [32].

2. Recall rate

The recall rate refers to the probability of correct identification among all positive samples [32].
Recall=TP/TP+FN

3. Accuracy

Accuracy refers to the proportion of correct predictions among all predictions [32].
Accuracy=TP+TN/TP+FP+TN+FN

4. AP

AP (Average Precision): AP is composed of the area enclosed by the P-R curve and coordinates, used to represent the average accuracy of detection at different recall rates [32].

5. MAP

MAP is the average value of each category of AP, used to take the average effect of detecting all target categories [32]. For example, if there are classes A and B, then MAP is their sum divided by two. The larger the ratio, the better the performance.

3.4.2. Areas that need improvement

Nowadays, with the deep learning and improvement of autonomous vehicles and object detection technology, many significant research achievements have been made in this field. However, there are still some problems that exist. (1) Regarding the issue of low detection accuracy and missed or erroneous detection of small targets, although YOLO's series of algorithms meet the real-time requirements of object detection, there are still significant issues with detection accuracy. For example, YOLOv1 can easily cause missed detections and low detection accuracy when the target distribution in the detection image is relatively dense [25]. Scholars have updated and upgraded the YOLO algorithm in response to this. After generations of evolution, the YOLO algorithm is able to achieve high accuracy, greatly avoiding the problem of missed and false detections of small targets. (2) To resolve poor robustness, the main issue is the optimization of the loss function. A reasonable loss function can improve the robustness of the detection algorithm [25]. (3) Visual-based object detection methods rely on high-performance computers and powerful graphics card capabilities, but the limited memory and computing resources in cars cannot meet this requirement. Therefore, the development direction of the visual-based object detection method aims to develop lightweight vehicle applications with low computational complexity and fast recognition speed, suitable for installation on mobile devices.

4. Conclusion

Object detection algorithms are the foundation of environmental perception, whether it is single visual object detection environmental perception or multi-sensor fusion environmental perception, object detection is essential. It detects obstacles, pedestrians, and other things on the road ahead through the camera. From the perspective of object detection, it can be divided into moving targets, stationary targets, and infrared targets. For movable animals, object detection can detect pedestrians, cars, etc. in real-time. For stationary objects, object detection can detect traffic signs, lane markings, and so on in real-time. Moving object detection and stationary object detection are aimed at ordinary images, while the all-weather, long-distance, and anti-interference characteristics of infrared imaging also occupy an important position in the field of object detection. From the application of object detection algorithms in different scenarios, they can be divided into two categories: ordinary scenes and complex scenes. A universal object detection algorithm that basically meets the problem of object detection in ordinary scenes. However, in complex scenes, there are some problems, such as the inability to detect small targets and poor detection of occluded targets. The improved object detection algorithm solves the problem of low detection efficiency of existing object detection algorithms, as well as the low accuracy of most lightweight object detection algorithm models, which are prone to false detection and missed detection of targets. It improves the accuracy and robustness of object detection algorithms.

References

[1] Mao, Z., Zhu, J., Wu, X., & Li, J. (2022). Review of YOLO based target detection for autonomous driving. *Computer Engineering and Applications, 58*(15), 68-77. https://doi.org/10.3778/j.issn.1002-8331.2203-0310

[2] Li, A., Gong, H., Huang, X., & Cao, J. (2022). Overview of target detection methods for autonomous vehicles. *Journal of Shandong Jiaotong University, 30*(3), 20-29.

[3] Song, S., Xia, H., & Li, G. (2023). Research on improved YOLOv5 algorithm and its application in multi-object detection for automatic driving. *Computer Engineering and Applications, 59*(15), 68-75.

[4] Li, Y., Hou, L., & Wang, C. (2019). Moving objects detection in automatic driving based on YOLOv3. *Computer Engineering and Design, 40*(4), 1139-1144.

[5] Shi, Y., & Jis, S. (2022). CB-CenterNet based autonomous driving occlusion target. *Journal of Dalian Jiaotong University, 43*(3), 115-120.

[6] Xu, X., Ma, Y., Qian, X., & Zhang, Y. (2021). Scale-aware Efficient Det: Real-time pedestrian detection algorithm for automated driving. *Journal of Image and Graphics, 26*(1), 93-100.

[7] Shi, J., Li, X., Liu, W., & An, X. (2022). Target detection and recognition based on YOLOv4-Efficient. *Intelligent Computer and Applications, 12*(3), 123-127.

[8] Chen, Y., Yan, Z., Zhou, C., & Huang, Y. (2023). Vehicle detection in autonomous driving scenarios based on improved YOLOv4. *Detecting Technology and Data Processing, 38*(1), 59-63, 85.

[9] Hu, J., Wang, H., & Dai, X. (2023). Real-time detection algorithm for small-target traffic signs based on improved YOLOv5. *Computer Engineering and Applications, 59*(2), 185-193.

[10] Wang, C., & Fu, Z. (2018). Traffic sign detection based on YOLOv2 model. *Coden Jyiidu, 38*(z2), 276-278.

[11] Zhang, K., & Zhu, M. (2022). Environmental perception algorithm for multi-task autonomous driving based on YOLOv5. *Computer Systems Applications, 31*(9), 226–232. https://doi.org/10.15888/j.cnki.csa.008698

[12] Lin, J., Zhang, W., Zhang, K., & Zhang, Y. (2022). Infrared target detection based on YOLOv5. *Flight Control & Detection, 5*(3), 63-71.

[13] Wen, H., & Tong, M. (2023). Object detection in automatic driving scenarios based on semi-supervised learning. *Microelectronic & Computer, 40*(2), 22-36.

[14] Gu, D., Luo, Y., & Li, W. (2022). Traffic target detection in complex scenes based on improved YOLOv5 algorithm. *Journal of Northeastern University (Natural Science), 43*(8), 1073-1079.

[15] Wildcraner. (2021). Detailed explanation KITTI dataset. [Blog post]. https://blog.csdn.net/weixin_46195203/article/details/115870752

[16] Geiger, A., et al. (2013). Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research, 32*(11), 1231-1237.

[17] Doan, A., Okatan, A., & Etinkaya, A. (2021). The code: Vehicle classification and tracking using convolutional neural network based on Darknet YOLO with COCO dataset. [ResearchGate publication]. https://www.researchgate.net/publication/353958426_Vehicle_Classification_and_Tracking_Using_Convolutional_Neural_Network_Based_on_Darknet_Yolo_with_Coco_Data

[18] Waiting for Godot. (2022). Introduce COCO dataset. [Blog post]. https://blog.csdn.net/qq_44554428/article/details/122597358

[19] Ertler, C., et al. (2022). The code: The Mapillary traffic sign dataset for detection and classification on a global scale. [arXiv preprint]. https://arxiv.org/abs/1909.04422

[20] Zhu, Z., et al. (2016). Traffic-sign detection and classification in the wild. [Conference paper]. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Zhu_Traffic-Sign_Detection_and_CVPR_2016_paper.pdf

[21] Murakami, M., et al. (2022). Radiative thermal conductivity of single-crystal bridgmanite at the core-mantle boundary with implications for thermal evolution of the Earth. *Earth and Planetary Science Letters, 578*, 117329. https://doi.org/10.1016/j.epsl.2021.117329

[22] AI Bacteria. (2023). YOLO series of algorithms in detail: The road from YOLOv1 to YOLOv8. [Blog post]. https://blog.csdn.net/wjinjie/article/details/107509243

[23] He, X., Luo, Y., & Jiang, L. (2021). Front vehicle detection based on YOLO. *Ship Electronic Engineering, 41*(1), 137-139.

[24] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You only look once: Unified, real-time object detection. *arXiv preprint*. https://arxiv.org/abs/1506.02640

[25] Zhao, Z. (2022). Analysis of target detection algorithm and its application in autonomous driving scenarios. *Automobile Applied Technology, 47*(4), 29-33.

[26] Plain white water. (2022). Object detection->SSD arithmetic. [Blog post]. https://blog.csdn.net/wanchengkai/article/details/12477589

[27] Liu, W., et al. (2015). SSD: Single shot multibox detector. [arXiv preprint]. https://doi.org/10.48550/arXiv.1512.02325

[28] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint*. https://doi.org/10.48550/arxiv.1311.2524

[29] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1904-1916.

[30] Girshick, R. (2015). Fast R-CNN. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 1440–1448). https://doi.org/10.1109/iccv.2015.169

[31] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint*. https://doi.org/10.48550/arxiv.1506.01497

[32] Mao, Z., Zhu, J., Wu, X., & Li, J. (2022). Review of YOLO based target detection for autonomous driving. *Computer Engineering and Applications, 58*(15), 29-33.