

A Review of Methods for Applying Object Detection to Intelligent Driving

Yilin Mei^a

Computer Science and Technology (Big Data Direction), China University of Geosciences (Wuhan), Wuhan, Hubei, China

Keywords: Target Detection, Faster R-CNN, YOLO, Transformer.

Abstract: Nowadays, intelligent driving attracts much attention as an emerging industry due to its convenience and the novel feeling it brings to people. However, there are also newspapers reporting that intelligent driving often results in loss of control and causes traffic accidents, which raises public concerns about the safety of this cutting-edge technology. In that case, this article aims to explain the causes of this phenomenon by deeply exploring the very important object detection technology in intelligent driving. Hence, through analyzing the development history, this article is able to illustrate the different solutions of target detection from two-stage target detection (represented by Faster R-CNN), Single-stage target detection (represented by YOLO), to self-attention mechanisms in Transformer and some other methods. Besides, it can also interpret the technical principles, advantages and difficulties faced by these methods. Through comparison and analysis, this article can help readers understand how target detection influences the environmental perception and safety of intelligent driving vehicles, thereby providing references for subsequent research, algorithm optimization, and industrial applications in this field.


1 INTRODUCTION

In recent years, with the rapid development of artificial intelligence and computer vision technology, intelligent driving has gradually moved towards reality and has become a significant development direction of future transportation. In an automatic driving system, environmental perception capability is the key technique to ensure the safe and efficient operation of vehicles. Consequently, target detection takes the responsibility for recognizing and positioning the different targets, such as pedestrians, vehicles, traffic signs and surrounding obstacles, as the key link of the perception system.

Target detection algorithms have seen a significant improvement in terms of detection accuracy, speed and deployment flexibility from initial convolutional neural network (CNN) to two-stage target detection and single-stage target detection. However, since the application scenarios of automatic driving are being more and more complicated, relying solely on traditional network structures is no longer sufficient to meet the robustness and real-time requirements for target

detection in all-weather and all-scenario environments. In that case, emerging methods such as the Transformer-based self-attention mechanism, BEV (Bird's Eye View) perspective, lightweight network structure, and multi-sensor fusion have emerged continuously, further promoting the rapid evolution of intelligent driving perception technology in recent years. Even so, there are still many challenges in intelligent driving target detection. The challenges include detection robustness under extreme weather conditions, multi-target tracking in dynamic occlusion environments, real-time processing on low-power hardware platforms, and reliance on large-scale and diverse labeled data etc. To solve these problems, academia continuously searches for new detection architectures and optimization strategies. They hope to make a balance between detection precision and calculation efficiency.

Through a systematic review of intelligent driving target detection technology, this article helps completely understand the development trajectory and the latest progress of object detection technology, analyze their advantages and limitations and looks

^a <https://orcid.org/0009-0009-5397-0899>

forward to possible research directions and breakthroughs in the future.

2 TARGET DETECTION

2.1 Intelligent Driving Target Detection Based on Two-Stage Method

The two-stage method of target detection evolved from the original convolutional neural network

(CNN) (Girshick et al., 2014). The original CNN mainly aims to recognize the target objects, determining which category the objects in the picture belong to. However, target detection requires some other demands, such as orientation and refining the bounding boxes rather than classification. They need to predict the location of the target in those pictures(usually represented by bounding boxes), which naturally necessitates a higher level of algorithms and calculations. In that case, the traditional CNN is very hard to support the new requirements.

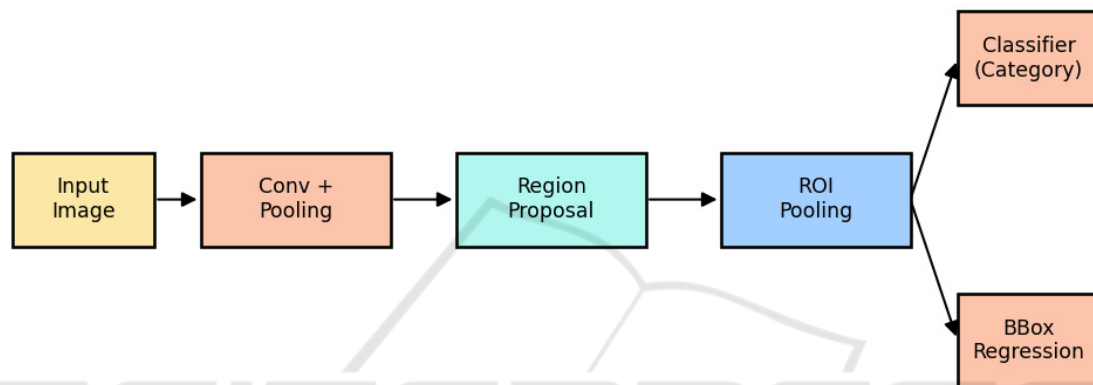


Figure 1: Flowchart of the two-stage object detection method (Picture credit: Original).

Under that circumstance, two-stage object detection emerged (Girshick et al., 2014). As shown in Figure 1, this method first takes an input image and passes it through several layers of convolution and pooling to extract feature maps. Then, it generates a sequence of region proposals, which probably contain those potential targets. These candidate regions are further processed through ROI pooling to align their features. Finally, the network branches out to classify each region into specific categories and regress the precise location of the bounding boxes. Compared to traditional algorithms, two-stage object detection achieves significant improvements in accuracy and loss rate, and it is widely applied in scenarios that demand high precision, such as intelligent driving (Ren et al., 2016).

Take an example of the Faster R-CNN algorithm, which was first applied in the intelligent driving field (Ren et al., 2016), it needs to detect if there are several different types of targets in one transportation photo, such as pedestrians, traffic signs, signal lights, road structures and so on. These functions allow the method are able to assist intelligent driving systems in choosing the correct roads or directions. To be specific, Faster R-CNN is very similar to traditional

CNN in its first stage, using some networks such as ResNet to extract features from the input images; then using Regional Proposal Network (RPN) to slide the window on the feature images, predicting if there is a target in the predictive recognition frames. Finally, output a series of candidate locations.

When it comes to the next stage, the network needs to orient the target location accurately and extract the features in candidate regions. To begin with, Faster R-CNN would make use of RoI Align technique to extract the regional features of fixed dimensions and then classify them into different categories, such as background, pedestrians and cars with fully connected layers (Ren et al., 2016). Meanwhile, the regression edge frames lead the location more accurately and make the frame background less redundant. Ultimately, the algorithm uses Non-maximum Suppression (NMS) to remove redundant predictions and output the detection outcomes.

By using this method, the two-stage target detection achieves great improvement compared to the past (Ren et al., 2016). Nevertheless, there also remain some problems. For instance, the detection speed of Faster R-CNN is very slow, requiring much

time to do analytical reasoning, which does not satisfy the demand of high immediacy for today's automatic driving. Besides, the model is too complex and needs high-level computer hardware configuration. Hence, because of its property of high precision and low immediacy, the two-stage target detection is more likely to be used in scenarios for offline high-precision map production and algorithm verification rather than modern real-time object detection, which often prefers single-stage detectors such as SSD (Liu et al., 2016).

2.2 Intelligent Driving Target Detection Based on YOLO

With the development of time, YOLO was born as a representative of single-stage object detection (Redmon et al., 2016). Different from the two-stage method, which first generates the candidate regions and then conducts classification and regression, YOLO divides the graph into a fixed number of grids. Each grid can directly predict the location of edge frames, their size and class probability. As a consequence, YOLO is able to complete the whole detection mission in a single stage (Redmon et al., 2016). The true end-to-end detection process has been realized. This type of YOLO structure can output the targets' location and classification at the same time through a single neural network. It allows the model structure more simpler and effective. Therefore, it is significantly suitable for intelligent driving scenarios with extremely high requirements of detection velocity.

The key principle of YOLO is to partition the input image into an $S \times S$ grid, and each grid is responsible for detecting and predicting any possible target in its certain regions (Redmon et al., 2016). It includes multiple bounding frames (Usually, B bounding boxes are predicted), each bounding frame's confidence and class probability. The confidence contains not only the probability of the existence of the target, but also the degree of overlap (IoU) between the bounding frame and the real target. Finally, the model will output all of the prediction outcomes and remove the redundant frames through the non-maximum suppression (NMS) method. It only retains some detection results with high confidence. This type of "regressive" thinking avoids many generations' steps in the lengthy candidate region in the two-stage method. Consequently, the inference speed can be greatly improved (Redmon et al., 2016).

The initial YOLO version (YOLOv1) was approximately 300 times faster than the two-stage

method in terms of speed (Redmon et al., 2016). However, its accuracy rate is 6.6 times lower than Faster R-CNN due to its limited faculties for detecting small targets. To solve these problems, the YOLO series gradually adds some new methods to make up for these deficiencies in its subsequent versions (Bochkovskiy et al., 2020).

Firstly, taking advantage of some good ideas of Faster R-CNN, the YOLO algorithm also introduces multi-scale anchor boxes to improve the detection ability of small targets and dense targets;

Second, YOLO uses Feature Pyramid Network (FPN) to achieve Multi-scale feature extraction and fusion, which supports detecting big targets and small targets from different resolution layers (Bochkovskiy et al., 2020).

Additionally, by using the advanced data augmentation methods such as Depth-separable convolution, Pruning, quantification and Mosaic, YOLO gets more lightweight and regularization. It obviously improves the detection effect while reducing the computational overhead (Bochkovskiy et al., 2020). Besides, YOLO improves its backbone networks, such as Darknet-53 and CSPDarknet and manages to enhance the feature extraction capability (Bochkovskiy et al., 2020; Wang et al., 2023).

So when it comes to the beginning of 2025, YOLO has developed the newest version to YOLOv11. It not only gets similar or better precision in the public datasets such as COCO and KITTI compared to the two-stage method, but also supports real-time operation at extremely high frame rates in embedded or vehicle-mounted environments (Wang et al., 2023). It allows YOLO to be one of the most representative detection algorithms in real-time intelligent driving scenarios.

Take a concrete example of intelligent driving, YOLO is capable of immediately coping with the video stream collected by the vehicle-mounted cameras and recognizing the key targets such as pedestrians, cars and transportation signs with high efficiency (Redmon et al., 2016; Wang et al., 2023). For example, YOLOv4 is improved with a lightweight backbone and feature pyramid to detect vehicles and pedestrians more robustly in urban scenes, achieving a 3.5% mAP gain on KITTI while keeping real-time speed (Ma et al., 2021).

These positive reinforcements provide reliable visual input for subsequent path planning or collision warning. In this case, the single-stage end-to-end structure reduces the latency effectively and enhances the safety and stability in a dynamic road environment for cars.

Meanwhile, nowadays academia is still trying to optimize the YOLO series to further enhance its robustness and generalization ability in extreme environments such as complex weather, night, rain and snow. With the introduction of new technologies such as Transformer, attention mechanisms, and unsupervised learning, YOLO is expected to play a greater pivotal role in intelligent driving target detection and support the construction of higher-level autonomous driving perception systems (Wang et al., 2023).

2.3 Intelligent Driving Target Detection Based on Transformer

The application of Transformer originated from the field of natural language processing, first proposed by Google (Vaswani et al., 2017). Its main self-attention mechanism was used for machine translation at the beginning. However, academia then discovered that it is also very excellent in many other fields with high accuracy and speed. And its performance is also far ahead of ordinary CNN and RNN in various rankings. Computer vision is also one of its applications (Dosovitskiy et al., 2020).

Different from traditional neural networks, the particular self-attention mechanism of Transformer allows it to handle input sequences of any length (Vaswani et al., 2017). This competence helps Transformer completely solve the problem of long-range dependence in RNN and has a powerful parallel computing performance. In target detection, using self-attention helps capture global information in the images and analyze the relation between different location features, so that the systems can understand the spatial interaction relationship between objects. And this is very important, especially in complicated situations. For example, an intelligent driving system usually needs to analyze the dynamic relationships among different vehicles and pedestrians at the same time on city roads (Dosovitskiy et al., 2020). For instance, TransTrack is proposed, which integrates a Transformer with a tracking-by-detection framework to jointly detect and track multiple objects in driving scenes, achieving higher accuracy and robust trajectory prediction compared to CNN-based trackers (Sun et al., 2020).

The Transformer model is capable of transforming feature map information into higher-level semantic information. Through capturing the relative location relationships between important targets, self-attention can analyze their behavioral trends. In that case, the Transformer can recognize the relationship between the distance and speed of

vehicles inside and outside the lane, which is helpful for predicting other cars' movement tracks (Dosovitskiy et al., 2020).

In addition, combining with other advanced techniques would also grant the Transformer greater efficacy. For instance, when combined with BEV (Birds-eye View), it brings about a global environmental perception capability, which can effectively reduce blind spots around the vehicles (Li et al., 2024). In high-speed roads or complex intersections, the cars need to recognize the lanes precisely in order to keep driving within the lane. Under those circumstances, the traditional camera recognition is easy to be influenced by light and viewing angle. On the contrary, the BEV + Transformer utilizes an all-around birds-eye view to stably identify lane lines even under harsh conditions, making it safer for vehicles to change lanes or make sharp turns (Li et al., 2024).

2.4 Other Methods in Intelligent Driving

In addition to the mainstream two-stage methods, the single-stage YOLO series, and the emerging detection methods based on Transformer, there are also some highly adaptable detection approaches. Among these, the lightweight detection network is one of the research hot spots in recent years. Since the computing power in vehicle-mounting scenarios is limited, some researchers introduced techniques such as network pruning, parameter sharing and efficient operators in the model structure design (Howard et al., 2017). All of these are aiming to minimize the model volume as much as possible while ensuring the detection ability, making it more convenient to run in real time in embedded devices. Some representative methods, such as MobileNet-SSD and Tiny-YOLO, implement these methods very well (Howard et al., 2017).

Besides, multi-task learning and end-to-end integrated perception have also become one of the important development directions of intelligent driving target detection. These improvements not only output classification and location, but also accomplish other missions such as semantic segmentation, instance segmentation, and object tracking, demonstrating higher insights into complicated road scenarios. For example, Mask R-CNN can not only detect the target, but also output the accurate segmentation results of the target area (He et al., 2017); CenterTrack adds a real-time multi-target tracking function on the basis of detection,

which greatly improves the environmental perception ability of vehicles in dynamic scenes (He et al., 2017).

These diverse methods continuously enrich the technical routes of intelligent driving target detection and complement the mainstream methods, jointly promoting the practical and intelligent development of autonomous driving perception systems.

3 DATASETS

In order to support the research and verification of target detection in intelligent driving, abundant high-quality labeled datasets have been made public by academia. These datasets often contain different types of transportation scenarios, weather, luminosity and information collected by sensors. And the datasets are widely used in detecting multiple types of targets such as vehicles, pedestrians, non-motorized vehicles, and traffic signs.

Table 1 summarizes the most representative and widely used public datasets in the field of intelligent driving at present.

4 CURRENT LIMITATIONS AND FUTURE PROSPECTS

Although the target detection technique has made great progress in the intelligent driving field, it still

has some limitations. So the academia is going to do some deep research in several aspects as follows.

At first, the robustness of models in extreme weather environments is still not sufficient. For instance, the visual detection is more likely to get missed detection and false detection when in harsh weather such as rain, snow and smog, which negatively influences the safety of the intelligent driving system. In the future, it might integrate more high-quality sensors to achieve multi-modal fusion perception to enhance the robustness and stability of the detection system in a particular environment.

In addition, the capabilities of multi-target detection and occlusion processing in complex dynamic scenes still need to be improved, especially in dense crowds and mixed traffic of non-motorized vehicles situations. Hence, it is significant to improve the adaptive ability to recognize unknown targets and support incremental learning of vehicles during operation to continuously enhance detection capabilities.

Moreover, the construction and update of labeled datasets are very expensive. Existing datasets still have disadvantages in scene diversity and long-tail distribution, making it difficult to cover all actual road conditions. Therefore, unsupervised learning, which is suitable for unlabeled or weakly supervised learning with few labels, can be introduced to enhance the generalization ability of the model and reduce its reliance on large-scale manually labeled data.

Table 1: Summary of available datasets in the field of intelligent driving.

Dataset Name	Year Released	Main Sensors	Number of Samples	Main Scenario	Key Features
KITTI	2012	Monocular & Stereo Cameras, LiDAR	15K+ frames	Urban Roads	Classic benchmark; includes detection, tracking, and 3D tasks
Cityscapes	2016	Monocular Camera	5K+ images	Urban Streets	High-resolution pixel-level annotation, detection and segmentation
BDD100K	2018	Monocular Camera	100K+ frames	Urban & Highways	Diverse weather, day/night; detection and tracking labels
nuScenes	2019	Cameras, LiDAR, IMU, GPS	1.4M frames	Urban Roads	Multi-sensor synchronized labels; supports 3D detection and tracking
Waymo Open Dataset	2019	Cameras, LiDAR	12M+ frames	Urban Roads	Large-scale dataset for autonomous driving; rich sensor annotations
ApolloScape	2018	Monocular Camera, LiDAR	140K+ frames	Urban & Suburban	Opened by Baidu; complex scenarios with lane markings
Argoverse	2019	Cameras, LiDAR, HD Maps	290K+ frames	Urban Roads	Includes high-definition map context; suitable for trajectory prediction

5 CONCLUSIONS

As a crucial part of environmental perception, target detection directly determines the safety and reliability of a vehicle's understanding of its surrounding environment and decision-making in intelligent driving systems.

This article reviews the development history and principles of intelligent driving target detection algorithms from two-stage to single-stage target detection and then to self-attention mechanism detection method based on Transformer, with some other methods for lightweight emerging in recent years. It also compares and analyzes the core ideas, performance and applicable scenarios of different techniques.

In summary, with the continuous improvement of algorithm theory and hardware computing power, object detection technology has made remarkable progress in intelligent driving. But it still needs deeper research for greater precision, robustness and immediacy to satisfy the greater requirement in complex scenarios in the future. It is evident that some excellent methods, such as Multimodal perception, lightweight network structure, and few-shot learning, are changing this field. Maybe one day they can make a great surprise for all.

REFERENCES

- Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
- Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., ... & Dai, J. (2024). Bevformer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *Computer Vision–ECCV 2016* (pp. 21–37). Springer International Publishing.
- Ma, L., Chen, Y., & Zhang, J. (2021, May). Vehicle and pedestrian detection based on improved YOLOv4-tiny model. In *Journal of Physics: Conference Series* (Vol. 1920, No. 1, p. 012034). IOP Publishing.
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).
- Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149.
- Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., ... & Luo, P. (2020). Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7464–7475).