Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU34649

Machine Vision-Based Object Detection for Autonomous Driving

Ruihong Zhong

College of Automotive Engineering, Jilin University, Changchun, China
zhongrh1523@mails.jlu.edu.cn

Abstract. Vision-based perception is of great significance in autonomous driving, and object detection is one aspect of it. With the development of deep learning, research on machine vision algorithms has increasingly focused on deep learning methods, and these algorithms are widely applied to object detection in autonomous driving. Therefore, this paper provides a general review of vision-based object detection in the field of autonomous driving, especially those based on deep learning. Firstly, vision-based object detection algorithms are introduced. The application of these algorithms in autonomous vehicle object detection, including the integration and optimization of these algorithms for vehicle detection, pedestrian detection, and the recognition of traffic instructions, are discussed in the following section. Finally, the summary of existing challenges and outlook for the future are provided based on the current state of research.

Keywords: Autonomous driving, Object detection, Machine vision

1. Introduction

Recently, with the continuous advancements in computer science and electronic information technology, autonomous driving has emerged as a critical research direction in the automotive industry, and related technologies are also upgrading rapidly. The autonomous driving systems, or intelligent driving systems, utilize the on-board hardware and software to automatically complete driving tasks in specific or all scenarios, thereby replacing human drivers to varying degrees. Generally, the architecture of an autonomous driving system consists of three modules: perception and localization, planning, and motion control [1]. In particular, the environmental perception and localization module is often regarded as a unified functional module that serves as the foundation for subsequent decision-making, path planning, and control. It is a key component of the system and determines the reliability and safety boundary of the whole system. The environment perception of autonomous vehicles uses various sensors to collect data about the surrounding environment and extracts various object features from the data for modeling. It is required to detect and identify static objects, recognize and track dynamic targets, and understand the scene where the vehicle is located through this module. In order to accomplish these tasks, an increasing number of studies are now inclined towards multi-sensor fusion technology. This technology integrates the data from cameras, radars as well as other heterogeneous sensors to overcome the disadvantages of a single type of sensors and enhance the comprehensive environmental perception ability for autonomous vehicles

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU34649

[2]. Among all the sensors, visual images have the high information density and can provide rich texture, grayscale and color [1, 2]. At the same time, the perception of human driving behavior relies primarily on vision, which means that visual information contains the key elements sufficient to complete the driving task. Furthermore, the cost of large-scale production and application of vision sensors is low due to mature hardware technology and well-established supply chain. Therefore, vision sensors have always been an important part of automobile sensor system. This paper mainly discusses the application of visual sensing technology in automobile environment perception.

The vision-based perception of autonomous vehicle involves several interrelated but distinct subtasks. This includes object detection, object tracking, semantic segmentation, and visual localization [1, 3]. Since object detection is a fundamental and major work in vision-based perception, this paper will focus on object detection part. Under road conditions, the objects detected by autonomous vehicles can be roughly classified as vehicles, pedestrians, and traffic instructions including traffic signs and traffic lights. The detection of these objects in road scenarios has its own unique characteristics, so this paper will analyze them separately. Notably, this paper will review research in these three areas with problem-solution logic and the structure is given as follows. Firstly, this paper introduces algorithms for vision-based object detection that are currently in widespread use, including their classification and representative algorithms. Then, it explores several primary object detection tasks in autonomous driving, which involve vehicle detection, pedestrian detection, and the recognition of traffic instructions. In particular, it discusses research on algorithm development and optimization that addresses problems like small-scale detection and detection under adverse conditions, together with the contributions of these studies in object detection tasks. Finally, a summary and outlook are presented.

2. Algorithms for vision-based object detection

Before deep learning technologies are fully developed, machine vision-based vehicle perception technology mainly relied on traditional methods, which required manual feature extraction [4]. These methods are susceptible to interference and have poor generalization ability in complex environments. At the same time, they also have problems with low efficiency and other issues, causing the development of traditional methods to slow down [4]. Currently, due to the advantages of deep learning, the research focus of object detection algorithms has shifted to convolutional neural networks (CNNs). In the field of deep learning methods, there are two categories of object detection algorithms: two-stage algorithms and one-stage algorithms, which are classified by their workflow. An introduction to these two categories and their representative examples is provided in the sections below.

2.1. Two-stage algorithms

The two-stage algorithm, literally, consists of two stages. The first step is to generate candidate region proposals, and the next procedure is to classify and localize objects based on candidate regions. This type of algorithm offers high detection accuracy and is particularly good at handling complex scenarios such as occlusions and small targets. However, its drawbacks include high computational demands and relatively slow processing speed. Faster R-CNN, R-FCN and Mask R-CNN are well-known algorithms in this category.

Faster R-CNN is proposed by Ren et al., and its outstanding achievement is to introduce the region proposal network (RPN) which can improve efficiency [4]. Mask R-CNN extends their framework and can be used for instance segmentation [4]. These algorithms are highly accurate and

DOI: 10.54254/2755-2721/2026.GU34649

are constantly being optimized to reduce inference time. However, their speed still limits their application in real-time detection. Especially when it comes to autonomous driving that inherently has high real-time demand, most implementations use one-stage algorithms, but two-stage algorithms can be utilized as references in many comparative experiments of algorithm optimization research [5, 6].

2.2. One-stage algorithms

One-stage algorithms directly predict object positions and classes from the whole image without an individual region proposal step. This architecture has faster speed and is ideal for real-time detection. However, in the early days, one-stage algorithms often performed less accurately than two-stage algorithms when dealing with extremely small objects and highly overlapping targets. In recent years, with the introduction of techniques such as Focal Loss and Feature Pyramid Network (FPN) , the accuracy of one-stage algorithms has improved significantly [4, 7]. Their representative examples include the YOLO series and SSD.

The YOLO (You Only Look Once) is a classic one-stage algorithm that processes images as a single regression task with the output of several parameters related to the objects [4]. The image input will be divided into a grid, and each cell in the grid will predict the same quantity of bounding boxes with information including the object's bounding box parameters as well as the conditional probabilities of the object belonging to different classes [4]. The YOLO algorithm also has limitations. Its accuracy is inferior to that of two-stage algorithms and its ability to detect small objects is relatively weak. However, with the improvements to the YOLO series, a reasonable balance has now been achieved between real-time performance and accuracy, leading to its increasing application in the field of autonomous vehicle object detection [8].

The SSD (Sing-Shot MultiBox Detector) is another one-stage algorithm. It uses VGG-16 and extra layers to extract multi-layer features, and presets a set of default boxes on each feature map [9]. The network then directly predicts the class probabilities and box adjustment parameters of the objects, and ultimately, the results from different layers are combined and processed together before final output [4, 9]. SSD uses low-level, large-scale feature maps for detection of small objects while the high-level, small-scale ones are used to detect large objects, which performs more effectively when detecting multi-scale targets.

3. Applications of vision-based algorithms in autonomous vehicle object detection

Vehicles, pedestrians, as well as traffic signs and traffic lights are key objects for autonomous vehicle perception. Only by accurately and rapidly identifying, locating and classifying these targets can a strong ability to judge and understand driving scenes be formed, so as to better complete the autonomous driving tasks. At present, there has been a significant amount of research dedicated to improving object detection algorithms for autonomous vehicles and these studies address specific problems in detection and recognition of the objects. Since the two types of traffic instructions share similarities, research on the detection of these objects is classified into three areas. The following is a separate discussion in research on these three aspects.

3.1. Vehicle detection

Vehicles are one of the most important traffic participants on the road, so vehicle detection is essential in autonomous vehicle object detection. Vehicle detection aims to locate and identify

DOI: 10.54254/2755-2721/2026.GU34649

vehicles from images, and its detection methods include traditional manual feature approaches and modern deep learning approaches [7]. Traditional methods rely on prior knowledge or manually designed feature descriptors, while deep learning-based detection is gaining more attention and becoming mainstream thanks to their advantage of automatic feature learning [7]. Currently, many researchers are optimizing existing deep learning algorithms such as the YOLO series and SSD to develop new vehicle detection methods.

In current vehicle detection research, the difficulty of detecting small vehicles has long been one of its major challenges. To address the scale issue in vehicle detection, that is, the problem of low accuracy at small scales, many researchers have optimized existing algorithms. An improved method for small object detection is YOLOv8-QSD [6]. This method employs module replacements in the backbone network of YOLOv8, which can bring richness to the feature space and provide more contextual information; at the same time, it optimizes the neck structure to enhance the ability of multi-scale feature fusion; moreover, it introduces Q-blocks to improve the localization of small objects [6]. This method not only demonstrated improved accuracy compared to YOLOv8 in laboratory experiments but also successfully detected distant vehicles missed by baseline models during real-world testing, proving its excellent small object detection capabilities [6]. The main improvement of this research is the optimization of small objects detection capabilities, and experiments have also proved its effectiveness for detecting small vehicles. However, not all studies solely focus on the detection of small-sized objects, especially considering another issue in vehicle detection about the balance between lightweight model and accuracy. Some studies have carried out research on the development of lightweight models and the improvement of the detection accuracy of small vehicles. One example is the study by Dong et al., which builds on YOLOv5s. Its core approach involves introducing the C3Ghost and Ghost modules to achieve lightweight model, while embedding the CBAM attention mechanism to improve accuracy [7]. The model demonstrated superior detection performance on vehicle datasets, achieving higher detection accuracy than YOLOv5s in challenging scenarios such as dense small vehicles with reduced model parameters [7]. However, their study selected vehicle data from a general object detection dataset without considering driving scenarios, and no experiments were conducted on embedded devices. This means their research merely remaining a general vehicle detection model rather than being specifically designed for the driving scenario.

Similarly, focusing on improving the lightweight design and accuracy of vehicle detection models, Chen et al. validated the performance of their method in autonomous vehicle detection using autonomous driving datasets and embedded devices [9]. This algorithm replaces the SSD backbone network from VGG16 with the lightweight MobileNet v2, while introduces the SE attention mechanism and a deconvolution feature fusion module [9]. It achieved excellent detection accuracy on two public datasets in experiments, and has better real-time processing capabilities compared to other models, making it suitable for embedded deployment. Unlike previous studies, this research achieves lighter model and faster processing speeds while maintaining high accuracy.

Nighttime, rain, snow and other adverse conditions pose another major challenge in vehicle detection research. Improving vehicle detection under adverse conditions can be achieved through model optimization and data augmentation based on vision sensors, or by employing multi-sensor fusion. For instance, the 3FL-Net aims to enhance overall detection performance in adverse weather conditions by constructing four subnetworks using pure visual methods [3]. This method consists of three auxiliary subnetworks for training and a lightweight detection subnetwork for detection [3]. On the RID and the iRain two rainstorm datasets, 3FL-Net can increase the average accuracy of the same lightweight detection subnetwork after training compared to its pre-training performance [3]. It

DOI: 10.54254/2755-2721/2026.GU34649

is also proved that the detection speed is equal to the baseline of the lightweight detectors, and it outperforms multiple combination methods that first remove rain and then perform detection [3]. In fact, the method described in that paper does not strictly belong to vehicle detection; rather, it is more like "detection for all" since their detection is not limited to vehicles. However, vehicle data constitutes the majority of their experimental dataset, which is one of the reasons for including their research in this category. Furthermore, it also indicates that if pure vision method is applied while solely detection for vehicles is optimized, the research has limitations especially when current research places greater emphasis on the generality of detection. Therefore, research on multi-class object detection is highly meaningful, and this issue will also be mentioned in the next section. In addition, there have been many studies on multi-sensor fusion to improve vehicle detection in adverse weather; however, due to the limited space and considering the research focus of this paper, a detailed discussion will not be provided here.

3.2. Pedestrian detection

Pedestrians are also important participants in traffic, so another key component in autonomous vehicle object detection is pedestrian detection. Since pedestrians are dynamic objects like other moving vehicles on the road, and functionally speaking, the perception of pedestrians and vehicles by the autonomous driving system is to prevent dangerous incidents such as collisions, so there are similarities in the perception of pedestrians and vehicles, which both require detection, 3D positioning and tracking. When it comes to vision-based object detection, the detection of pedestrians and vehicles also share similarities, such as the use of the same methodologies and the presence of similar challenges. For this reason, many detection algorithms now treat pedestrian and vehicle detection as an integrated task. For example, VP-YOLO is a reconstructed model designed for vehicle-pedestrian detection in complex conditions, which mainly solves three types of complex detection problems: long-distance, truncation and occlusion [5]. This model inserts an enhancement stage inspired by human visual perception between the backbone and neck of YOLOv7 [5]. The results on two public datasets and a self-built dataset show that this model has achieved leading detection performance for both small pedestrians and vehicles, especially in long-distance situation [5]. This method uniformly addresses the problems of small scale and occlusion that exist in both pedestrian and vehicle detection. Meanwhile, both pedestrian and vehicle detection face the same challenge of balancing detection accuracy and speed. Consequently, some researchers have developed lightweight, real-time detection models. One example is an improved architecture based on Tiny YOLO v3 that can deploy on embedded devices [8]. This architecture compresses Tiny YOLO v3 Darknet through Tucker tensor decomposition and realizes real-time vehicle and pedestrian detection on two embedded platforms such as Raspberry Pi [8]. It achieved a significant speed improvement and reduced memory usage at the cost of only about a 2% loss in accuracy compared to the baseline on two different datasets, providing an efficient lightweight solution for resource-constrained devices [8].

It is worth noting that the YOLO-ACE algorithm, an upgraded version of YOLOv10, can reduce the model size and speed while maintaining high accuracy [10]. It optimizes feature extraction by introducing the Add-CGLU architecture and the FPSC module, improves the fusion of multi-scale features by designing the EMBS pyramid network, and enhances generalization ability and robustness through a double distillation strategy [10]. The comparative experiments and the generalization experiment have verified its ability to balance detection accuracy and computational efficiency, as well as its robustness under snow, fog and other adverse weather scenarios [10].

DOI: 10.54254/2755-2721/2026.GU34649

Although improving the detection of pedestrians and vehicles as a whole is an important research direction, pedestrian detection still has its own unique characteristics. Therefore, there are studies focusing on the optimization or development of algorithms and models specifically for pedestrian detection. For instance, under adverse conditions such as at night, the multi-sensor fusion method is beneficial for enhancing detection robustness, while the characteristics of pedestrians make them suitable for detection using thermal images [2]. As a result, some researchers have proposed to conduct fusion detection by combining standard optical cameras with thermal cameras [2]. Additionally, pedestrians are dynamic objects with specific movement patterns, making the analysis of their movement a significant task. Kilicarslan et al. have pointed out that pedestrian walking exhibits certain characteristics which can be analyzed by examining the chain trajectories formed by their legs in the time and space domain [11]. This enables the real-time acquisition of the walking direction and speed [11]. This type of pedestrian motion analysis can facilitate the detection of walking pedestrians, thereby mitigating risks.

3.3. Recognition of traffic signs and traffic lights

The recognition of traffic signs and traffic lights in autonomous driving is quite different from vehicle and pedestrian detection. Traffic signs and traffic lights provide information regarding traffic regulations, traffic warnings, and route guidance. In order to fully understand that information, the traffic sign recognition mainly focuses on accurately identifying and classifying various types of traffic signs in road scenarios, while traffic light recognition primarily involves identifying and judging the status of traffic signal lights. Their recognition is a critical means to ensure road safety and the legality of autonomous driving. Since these two instructions rely almost entirely on vision to convey information, the recognition of them is generally carried out using visual approaches especially during the classification stages.

Whether it is the recognition of traffic signs or traffic lights, they can generally be divided into traditional methods based on experience and artificial intelligence methods [12, 13]. Nowadays, artificial intelligence methods, which commonly includes machine learning and deep learning methods, are widely applied in the recognition tasks, and the research on deep learning methods is particularly more prominent. It is worth noting that the integration of machine learning-based detectors and CNN models can also achieve the purpose of optimization. For example, Triki et al. developed a method that combines a Haar cascade detector with a new CNN classifier incorporating an attention mechanism [12]. The experiments demonstrated that not only does this novel classifier itself exhibit high accuracy and strong generalization ability, but the entire integrated architecture could achieve high-precision, real-time recognition of various types of traffic signs on embedded platforms [12].

In terms of deep learning methods, recent research has also been directed toward solving the problem in traffic signs and lights recognition. Similar to vehicle and pedestrian detection, the recognition of traffic signs also has the issue of small scales. Therefore, optimizing the architecture for multi-scale detection is a key research direction currently. For instance, Wang et al. improved YOLOv5 by designing the AF-FPN feature pyramid and employing an automatic data augmentation strategy, thereby significantly enhancing the detection accuracy of small traffic signs on a traffic sign dataset [14]. Another challenge in traffic sign and traffic light recognition is detection under adverse conditions, such as at night. However, due to the characteristics of traffic signs and traffic lights, it is difficult to conduct their detection through multi-sensor fusion. As a result, existing solutions still rely on methods such as model optimization and data augmentation and other approaches to enhance visual perception capabilities and robustness. For traffic light recognition,

DOI: 10.54254/2755-2721/2026.GU34649

EG-TLR is a detection framework specifically designed for nighttime traffic light recognition [12]. It integrates two components: a residual denoising module (RDM) for suppressing noise and a mixed attention traffic light detection module (MATLDM) for strengthening recognition ability [12]. Comparative experiments on a self-built traffic signal dataset demonstrated that EG-TLR achieves better detection accuracy in the task of nighttime traffic signal recognition, and its generalization capability has also been validated on another public dataset [12]. In terms of traffic sign recognition, Qu et al. proposed an algorithm based on YOLOv5s for complex weather conditions [15]. This approach not only increases the accuracy of traffic sign detection under various weather conditions but also enhances the detection ability for small traffic signs [15].

4. Conclusion

This paper focuses on object detection based on machine vision and reviews the recent research status in three specific aspects using a problem-oriented approach. From the analysis above, it is evident that there are some shared challenges across the detection of vehicles, pedestrians, and recognition of traffic signs and traffic lights. These include: the small-scale problem, which refers to the issue of low detection accuracy or even missed detection for distant objects due to their small size; the occlusion problem, which refers to the decline in detection accuracy when objects are obscured; the adverse condition problem, which refers to the lack of detection accuracy and stability under adverse conditions including but not limited to rainy and snowy days or at night; the balance issue, which refers to the trade-off between accuracy, processing time, and model size in detection algorithms. As current research continues to solve these issues, several trends are worthy of attention, as they are possible reference directions for future research. First of all, the introduction of attention mechanisms has become popular. It can suppress noise or other interference and improve accuracy. Meanwhile, efforts to enhance robustness in complex environments, especially with the application of multi-sensor fusion technology, are becoming a trend. Although this article does not focus on the fusion methods, the perception through integration of multiple sensors, whether including vision or not, remains a central research topic currently. In addition, lightweight model design with good real-time performance have been extensively studied. The example includes the use of knowledge distillation for model compression. Certainly, knowledge distillation is not only helpful to lightweight design, but also an important means to enhance the generalization ability of the model.

References

[1] Zhao, J., Zhao, W., Deng, B., Wang, Z., Zhang, F., Zheng, W., Cao, W., Nan, J., Lian, Y., Burke, A.F., 2024. Autonomous driving system: a comprehensive survey. Expert Systems with Applications 242, 122836.
[2] Dasgupta, K., Das, A., Das, S., Bhattacharya, U., Yogamani, S., 2022. Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. IEEE Trans. Intell. Transport. Syst. 23, 15940–15950.
[3] Huang, S.-C., Jaw, D.-W., Hoang, Q.-V., Le, T.-H., 2023. 3FL-Net: an efficient approach for improving performance of lightweight detectors in rainy weather conditions. IEEE Trans. Intell. Transport. Syst. 24, 4293–4305.
[4] Wu, X., Sahoo, D., Hoi, S.C.H., 2020. Recent advances in deep learning for object detection. Neurocomputing 396, 39–64.
[5] Liu, W., Qiao, X., Zhao, C., Deng, T., Yan, F., 2025. VP-YOLO: a human visual perception-inspired robust vehicle-pedestrian detection model for complex traffic scenarios. Expert Systems with Applications 274, 126837.
[6] Wang, H., Liu, C., Cai, Y., Chen, L., Li, Y., YOLOv8-QSD: an improved small object detection algorithm for autonomous vehicles based on YOLOv8. IEEE Trans. Instrum. Meas. 73, 1–16.

DOI: 10.54254/2755-2721/2026.GU34649

[7] Dong, X., Yan, S., Duan, C., 2022. A lightweight vehicles detection network model based on YOLOv5. Engineering Applications of Artificial Intelligence 113, 104914.
[8] Falaschetti, L., Manoni, L., Palma, L., Pierleoni, P., Turchetti, C., 2024. Embedded real-time vehicle and pedestrian detection using a compressed tiny YOLO v3 architecture. IEEE Trans. Intell. Transport. Syst. 25, 19399–19414.
[9] Chen, Z., Guo, H., Yang, J., Jiao, H., Feng, Z., Chen, L., Gao, T., 2022. Fast vehicle detection algorithm in traffic scene based on improved SSD. Measurement 201, 111655.
[10] Xie, Y., Du, D., Bi, M., 2025. YOLO-ACE: a vehicle and pedestrian detection algorithm for autonomous driving scenarios based on knowledge distillation of YOLOv10. IEEE Internet Things J. 12, 30086–30097.
[11] Kilicarslan, M., Zheng, J.Y., 2023. Deepstep: direct detection of walking pedestrian from motion by a vehicle camera. IEEE Trans. Intell. Veh. 8, 1652–1663.
[12] Triki, N., Karray, M., Ksantini, M., 2023. A real-time traffic sign recognition method using a new attention-based deep convolutional neural network for smart vehicles. Applied Sciences 13, 4793.
[13] Yao, Z., Liu, Q., Zhao, Z., Qin, Y., Zhu, J., Xia, T., Li, B., Wang, L., 2025. Night-time traffic light recognition based on enhancement-guided object detection. IEEE Trans. Syst. Man Cybern, Syst. 55, 4410–4422.
[14] Wang, Junfan, Chen, Y., Dong, Z., Gao, M., 2023. Improved YOLOv5 network for real-time multi-scale traffic sign detection. Neural Comput & Applic 35, 7853–7865.
[15] Qu, S., Yang, X., Zhou, H., Xie, Y., 2023. Improved YOLOv5-based for small traffic sign detection under complex weather. Sci Rep 13, 16219.