{\n "SummaryDoc": "# Highlights in Science, Engineering and Technology EMIS 2024 Volume 119 (2024)\n\n## Abstract\n\nObject detection has always been a fundamental research topic in the computer vision community, which focuses on predicting the category and location of all objects in the scene. In last several years, progressing from the rapid development of deep learning, the speed and accuracy of general object detection methods have also achieved significant breakthroughs. This paper aims to report the latest research progress in the field of object detection based on deep learning to inspire and promote subsequent research. Specifically, this paper systematically introduces the research progress of predecessors from four aspects: dual-stage, single-stage, Transformer-based, and key point, including the design ideas and basic processes of representative algorithms. In addition, this paper also quantitatively compares the performance of different methods on common data sets to further distinguish the benefits and disadvantages of various categories of methods. Finally, this paper summarizes the challenges that still exist in the field of object detection and looks forward to future development directions.\n\nKeywords: computer vision, natural language process, Object detection, anchor-free.\n\n## Introduction\n\nWith the fast development and the wide application of artificial intelligence technology, autonomous vehicles are becoming an important part of future transportation in our daily lives. Autonomous driving technology tries to achieve autonomous navigation and control of vehicles by computer vision, sensor technology, and deep learning algorithms. In recent years, the research of autonomous driving has made major breakthroughs, which not only promote the transformation and upgrading of the automotive industry but also bring new opportunities for the development of modern cities and intelligent transportation.\n\nIn the process of the realization of autonomous driving technology, environmental perception is a vital link, which is directly related to the safety and reliability of autonomous vehicles. The core task of an environmental awareness system is to accurately detect and identify all kinds of aims on the road, including pedestrians, vehicles, traffic signs, road obstacles, and so on. Among them, Object Detection, as the basic link of environment perception, plays a key role. The accuracy and real-time performance of object detection directly affect the decision-making and control ability of automatic driving systems. There is still lots of development space and research value in the future technology. To popularize the public's doubts about object detection technology in autonomous driving, this article will show the development process of the technology, analyze the shortcomings, discuss the social significance, and propose current challenges and future directions.\n\nBased on a review of the general and specialized approaches to object detection task modeling in automated driving, we classify the main approaches into four categories according to the evolution route and development time of the model: two-stage, single-stage, key point-based and transformer-based approaches. Through analyzing the representative models of various methods and the pros and cons of existing technologies, we divided the main features of this generation of methods and the improvement of the next generation of methods over the previous generation. At the end of this paper, the current problems and challenges of object detection technology are discussed. With the continuous progress of autonomous driving technology, object detection algorithms still need to be continuously innovative in the face of complex environments and diverse scenarios. Multi-domain object detection, multi-mode fusion, small object detection, lightweight network architecture, video detection, weak supervision, and small sample detection will become the focus of future research. Breakthroughs in these directions will further promote the intelligence and safety of autonomous driving systems, and promote their faster application to reality.\n\n## Method\n\n### 2.1. Two-stage Detection Approaches\n\nTwo-stage detection methods decompose the object detection task into two stages: region proposal generation, target classification, and bounding box regression. With the refinement of the two phases, the two-phase detection algorithm can identify and localize target objects more accurately, especially when dealing with small and multi-scale targets. Also, two-stage detection possesses greater flexibility. Dual-stage detection allows the algorithm to focus on generating high-quality candidate regions in the first stage, and then focus on classification and precise adjustment of the bounding box in the second stage, a separation that makes the algorithm more flexible. Moreover, two-stage detection also possesses greater scalability and can be applied to different datasets and tasks by simply making appropriate adjustments to the second-stage classifiers.\n\nAs the most representative method, Faster R-CNN achieves high-precision detection performance on multiple datasets through a two-stage network structure with a region proposal network (RPN) and a classifier, which is easy to migrate to other scenes by changing the target class in the dataset [1-2]. Moreover, by decreasing the number of shared convolutional layers, the amount of computation is reduced, which is widely used in the fields of automatic driving, security monitoring, face detection, object recognition and so on. Despite the above advantages of Faster R-CNN, there are some drawbacks, such as (1) the small resolution of the feature maps of the convolutional extraction network, which may not be conducive to the detection of small and multi-scale objects; (2) NMS may not be friendly to obscured objects, leading to missed detections [3]; (3) The loss of accuracy in RoI Pooling is limited [4]; (4) A large number of parameters and unshared computations in the fully connected layer; (5) the fact that positive and negative sample equalization methods may not be optimal; (6) the relatively slow detection speed.\n\nMask R-CNN [5] has added a new branch to Faster R-CNN for the purpose of predicting the segmentation mask for each object to achieve accurate instance segmentation. In addition, it inherits the RPN part of Faster R-CNN and integrates a multi-task learning framework, demonstrating the trend of integrating complex models in the field of deep learning. This algorithm is capable of generating a segmentation mask for each detected object to achieve accurate instance segmentation, which gives it a wide potential for practical applications such as human pose estimation, object segmentation, etc., to achieve efficient computation while maintaining accuracy. As a result, Mask R-CNN has achieved remarkable results in instance segmentation tasks through its innovative network structure and training method and has become a significant milestone in the field of computer vision.\n\n### 2.2. Single-stage Detection Approaches\n\nSingle-stage object detection is a class of object detection methods that regresses the target bounding box and categories directly from the input image without going through the Region Proposal step. Compared to two-stage methods, single-stage methods typically have higher detection speeds and are suitable for real-time applications.\n\nIn order to cope with the problems of the complexity of the two-stage object detection model, multiple parameters, long training time, and poor real-time detection, Redmon et al. [6] proposed the YOLO (You Only Look Once) algorithm in 2015. This is the first single-stage detector in the deep learning space. the big advantage of YOLO is that its processing speed is very fast, with the enhanced version reaching up to 45 FPS on the GPU, and the fast version even reaching 155 FPS. The YOLO algorithm directly divides the whole image into S × S grid cells through a single neural network, and the task of each grid is to detect targets within it and predict the target class, confidence level, and bounding box location [7]. The whole process passes through several layers of convolutional and pooling layers and then two fully connected layers to output the classification result and bounding box of the target. Finally, the algorithm uses threshold filtering to remove low-confidence targets and in order to detect objects, non-maximum suppression (NMS) will remove redundant bounding boxes.\n\nAlthough YOLO significantly outperforms the two-stage Faster R-CNN in terms of detection speed, there are some limitations. First, since YOLO uses an S × S grid for prediction, if there are multiple targets in the same grid, it may lead to missed detection, especially poor detection of small targets. Second, instead of using the anchor frame mechanism of Faster R-CNN, YOLO directly predicts the absolute position, which increases the difficulty of model training. In addition, the aspect ratio of the prediction frame is preset based on the training set, which is a weak generalization to new datasets. Finally, since YOLO can only output the bounding box that has the highest IoU when dividing the grid, this means that at most only one target can be detected per grid, which is less effective for detecting scenes containing multiple small targets (e.g., flocks of birds).\n\nTo address the shortcomings of YOLO in small object detection, Liu et al. [8] provided the SSD (Single Shot MultiBox Detector) algorithm in 2015. SSD combines the advantages of YOLO's rapid detection speed and Faster R-CNN's accurate localization by introducing multi-level feature maps for detecting targets at various scales. Specifically, SSD uses shallow feature maps to detect small targets and deep feature maps to detect large targets, thus dramatically improving the detection of small targets [9-10]. SSD utilizes multi-reference and multi-resolution detection techniques to detect objects of different sizes through different levels of networks. Meanwhile, SSD borrows the Anchor trick from Faster R-CNN to set up a priori frames with different aspect ratios in advance, and predicts the object detection frames based on them, thus reducing the training complexity. In order to further improve the detection effect, Liu et al. also introduced the difficult sample mining technique to address the issue of focusing on difficult samples during the model training process. However, SSD still has some limitations: first, feature maps at different scales are detected independently, which may lead to repeated detection of the same target and increase the computational effort of the model. Second, although the shallow feature map is used for detecting small targets, there is still opportunitjson\_output = { " \n }