# Research Advanced in Object Detection based on Deep Learning

## Abstract

Object detection has always been a fundamental research topic in the computer vision community, focusing on predicting the category and location of all objects in the scene. Progress from the rapid development of deep learning has significantly advanced the speed and accuracy of general object detection methods. This paper reports recent research progress in object detection based on deep learning, introducing the research progress of predecessors from four aspects: dual-stage, single-stage, Transformer-based, and key point methods. It includes the design ideas and basic processes of representative algorithms and quantitatively compares different methods on common datasets to highlight their benefits and disadvantages. Lastly, the paper summarizes the current challenges in the field of object detection and anticipates future development directions.

\*\*Keywords:\*\* computer vision, natural language processing, Object detection, anchor-free.

## 1. Introduction

In recent years, autonomous vehicles have increasingly become significant in future transportation developments through advancements in artificial intelligence technologies, including computer vision, sensor technology, and deep learning algorithms. Object detection forms a critical element of environmental perception essential for autonomous driving, directly impacting the safety and reliability of these vehicles. This opens substantial development space and research value, and clarifies public concerns about object detection technology in autonomous driving, introducing the developmental stage of the technology, analyzing shortcomings, discussing social significance, and proposing current challenges and future directions. The paper categorizes object detection methods into dual-stage, single-stage, key point-based, and transformer-based approaches, examining their characteristics and reviewing the advancements made by these methods in automatized driving.

## 2. Method

### 2.1 Two-stage Detection Approaches

Two-stage methods break down the object detection task into two steps: region proposal and subsequent target classification and bounding box regression. These methods offer increased flexibility and scalability, adapt to various datasets, and are particularly effective in handling small and multi-scale targets. Faster R-CNN, a predominant model, delivers high precision using a two-stage network structure featuring a region proposal network (RPN) and a classifier. Despite its advantages, there are limitations such as small feature map resolutions which may not effectively detect small objects and a relatively slower detection speed.

### 2.2 Single-stage Detection Approaches

In contrast to two-stage methods, single-stage approaches directly predict bounding boxes and categories from the input image without an intermediary step. Highlighted by methods like YOLO and SSD, these algorithms present faster detection times suited for real-time applications. While YOLO speeds up the detection, it faces challenges with small object detection and the generalization to new datasets. SSD improves upon YOLO by employing multi-level feature maps that enhance the detection of small targets but still struggles with independent feature map detection leading to potential repeated detections.

### 2.3 Key point-based Detection Approaches

Anchor-free key point-based methods like CornerNet and CenterNet offer significant advances by reducing reliance on traditional anchor box strategies. These methods streamline the model architecture and reduce parameter counts, enhancing operational speeds and facilitating the application across different datasets.

### 2.4 Transformer-based Detection Approaches

Transformer models such as DETR utilize an end-to-end architecture that simplifies the object detection process by combining CNN and Transformer technologies to process the data. While they offer improved detection accuracies and simplify the detection process by eliminating steps like NMS, their training convergence is considerably slower, and they generally require more computational resources.

## 3. Experiment

### 3.1 Datasets and Metrics

This study utilizes standard datasets like Pascal VOC and COCO to analyze and compare the effectiveness of various object detection methods. These datasets provide a comprehensive platform for benchmarking improvements in detection techniques across different model architectures.

### 3.2 Performance Comparison

\*\*Table 1. Accuracy (mAP) Comparison between Different Representative Methods\*\*  
| Method | Category | VOC2007 | COCO |  
|:--------|:---------|-------------:|----------:|  
| Fast RCNN | two-stage | 70.0 | 19.7 |  
| Faster R-CNN | two-stage | 73.2 | 21.9 |  
| SSD | Single-stage | 81.6 | 26.8 |  
| YOLOv4 | Single-stage | \*\*81.8\*\* | \*\*43.5\*\* |  
| Retina-Net | Single-stage | 79.5 | 36.2 |  
| CenterNet | Key point-based | 78.0 | 41.8 |  
| Deformable DETR | Transformer-based | 79.5 | \*\*52.3\*\* |  
| Swin Transformer | Transformer-based | \*\*81.0\*\* | \*\*57.7\*\* |

## 4. Discussion

This section discusses the ongoing challenges and future research directions, emphasizing the need for high-precision lightweight model architectures, video detection optimization, multisensory data integration for improved detection in adverse conditions, and reduced dependency on extensive labeled data collections through techniques such as transfer learning and weak supervision.

## 5. Conclusion

The maturation of deep learning technologies has propelled significant enhancements in object detection accuracies and processing speeds. This review analyzes and compares various detection frameworks, advocating for innovations and further research to overcome existing limitations and scale up the applicability and efficiency of object detection technologies in real-world scenarios.

## Authors Contribution All authors contributed equally to this work and are listed in alphabetical order based on their first names.

## References [1] Girshick R., et al. 'Fast R-CNN', ICCV 2015. Santiago: IEEE, 2015. pp. 1440–1448. [2] Ren S.Q., et al. 'Faster R-CNN: Towards real-time object detection with region proposal networks', arXiv:1506.01497, 2015. [3] Neubeck A. and Van Gool L., 'Efficient non-maximum suppression', ICPR 2006. IEEE, 2006. [4] Qin Y., et al. 'RoI pooling based fast multi domain convolutional neural networks for visual tracking', Artificial Intelligence and Industrial Engineering Conference, Beijing, 2016. [5] He K.M., et al. 'Mask R-CNN', ICCV 2017. IEEE, 2017. [6] Divvala S., et al. 'You only look once: Unified, real-time object detection', CVPR 2016. IEEE, 2016. [7] Li G., et al. 'A new method of image detection for small datasets under the framework of YOLO network', IEEE 2018. [8] Liu W., et al. 'SSD: single shot multibox detector', European Conference on Computer Vision, 2016. [9-10] Chen H.J., et al. 'SSD object detection algorithm with multi-scale convolution feature fusion', Journal of Frontiers of Computer Science and Technology, 2019. [11] Lin T.Y., et al. 'Focal loss for dense object detection', arXiv:1708.02002, 2017. [12-15] Relevant literature citations continuing as per the provided list in a detailed manner reflecting their respective contributions to the field of object detection.