# Highlights in Science, Engineering and Technology EMIS 2024 Volume 119 (2024)

## 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. In recent years, the rapid development of deep learning has significantly advanced the speed and accuracy of general object detection methods. This paper reports the latest research progress in object detection based on deep learning, introducing the research progress from four aspects: dual-stage, single-stage, Transformer-based, and key point methods, including the design ideas and basic processes of representative algorithms. Additionally, this paper quantitatively compares the performance of different methods on common datasets and discusses the challenges that still exist in the field, looking forward to future development directions.

Keywords: computer vision, natural language process, Object detection, anchor-free.

## 1. Introduction

With the fast development and the wide application of artificial intelligence technology, autonomous vehicles are becoming an important part of future transportation. Autonomous driving technology aims to achieve navigation and control of vehicles through computer vision, sensor technology, and deep learning algorithms. This research has brought new opportunities for the development of modern cities and intelligent transportation. Environmental perception, crucial for the safety and reliability of autonomous vehicles, involves accurately detecting and identifying road elements like pedestrians, vehicles, traffic signs, and obstacles, forming the core task of the environmental awareness system.

This paper discusses the development process, addresses public concerns about object detection in autonomous driving, and proposes current challenges and future directions for research.

## 2. Method

### 2.1 Two-stage Detection Approaches

Two-stage detection methods break down the object detection task into region proposal generation and target classification plus bounding box regression. As a highly representative method, Faster R-CNN achieves high-precision detection on multiple datasets using a two-stage network structure. However, it has drawbacks, such as limited performance on small and multi-scale objects and a relatively slow detection speed. Mask R-CNN extends Faster R-CNN to predict a segmentation mask for each detected object, combining efficient computation with accurate instance segmentation.

### 2.2 Single-stage Detection Approaches

Single-stage object detection methods, like YOLO and SSD, provide rapid detection speeds suitable for real-time applications. YOLO divides the image into grid cells and predicts the target category and bounding box directly. SSD, introduced by Liu et al., improves small object detection by using multi-level feature maps and introducing techniques like difficult sample mining to focus on challenging samples during training.

$$L\_{BCE} = -[y \log(p) + (1-y)\log(1-p)] \tag{1}$$

### 2.3 Key point-based Detection Approaches

Key point-based methods such as CornerNet and CenterNet operate by detecting key points of a target for localization. These methods are anchor-free and aim to reduce the dependency on numerous anchor boxes, which can lead to resource-heavy computation and complex hyperparameter tuning.

### 2.4 Transformer-based Detection Approaches

Transformers, utilized in DETR (Detection Transformer), model object detection as an ensemble prediction problem. Despite its innovative approach, DETR requires improvements in small object detection and training efficiency. Deformable DETR addresses these issues by integrating deformable convolution with Transformer architecture to enhance the performance on small object detection.

[Image missing: page2\_img1.jpg] Caption: "DETR Model Framework" | Explanation: "Illustrates how DETR employs transformers to perform object detection. The method simplifies the detection process by eliminating the need for NMS, showcasing the effectiveness of transformers in providing a streamlined approach." | Ref: PDF p.492

## 3. Experiment

### 3.1 Datasets and Metrics

This study utilizes datasets like Pascal VOC and COCO to evaluate the performance of target detection algorithms. Common metrics used include average precision (AP), mean average precision (mAP), precision, recall, and detection speed (FPS).

### 3.2 Performance Comparison

This section presents a quantitative comparison of the different detection methods, highlighting the advantages of Transformer-based methods in achieving higher accuracy on the COCO dataset.

\*\*Table 1. Accuracy (mAP) comparison between different representative methods\*\*

| Method | Category | VOC2007 | COCO |  
|:--------------|:------------------------|--------:|------:|  
| Fast RCNN | Two-stage method | 70.0 | 19.7 |  
| Faster RCNN | | 73.2 | 21.9 |  
| SSD | Single-stage method | 81.6 | 26.8 |  
| YOLOv4 | | 81.8 | 43.5 |  
| Retina-Net | | 79.5 | 36.2 |  
| RefineDet | | 79.2 | 39.1 |  
| CornerNet | Key point-based method | 79.5 | 42.5 |  
| CenterNet | | 78.0 | 41.8 |  
| FCOS | | 78.8 | 44.7 |  
| Deformable DETR | Transformer-based method | 79.5 | 52.3 |  
| Swin Transformer | | 81.0 | 57.7 |

## 4. Discussion

This section discusses the critical role of target detection algorithm invention and optimization in advancing autonomous driving technologies. It highlights the necessity for multi-domain target detection, improvement in the detection of small objects, development of high-precision lightweight network architectures, video detection optimization, and multimodal detection to enhance the robustness and accuracy under various conditions.

## 5. Conclusion

The rapid development of deep learning technology has significantly improved the accuracy and speed of object detection. This paper reviews various object detection frameworks and introduces promising future directions for research. The collective effort of the authors has been crucial in compiling this review and analysis.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Girshick R, et al. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448.  
2. Ren SQ, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv:1506.01497, 2015.  
3. NEUBECK A, VAN GOOL L. Efficient non-maximum suppression. Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong: IEEE, 2006: 850-855.  
4. QIN Y, et al. RoI pooling based fast multi domain convolutional neural networks for visual tracking. Proceedings of the 2016 International Conference on Artificial Intelligence and Industrial Engineering. Beijing: IEEE, 2016: 198-202.  
5. He KM, et al. Mask R-CNN. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988.