Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ23816

Comparative Analysis of Object Detection Architectures for Enhanced Performance and Application Suitability

Yaofei Wang

School of Management Engineering, Qingdao University of Technology, Qingdao, China
mochen@uest.edu.gr

Abstract: Object detection is a cornerstone task in computer vision, impacting diverse domains from autonomous systems to medical imaging. This study presents a systematic evaluation of three leading detection architectures—Faster Region-based Convolutional Neural Networks (Faster R-CNN), You Only Look Once version 8 (YOLOv8), and Detection Transformer (DETR), with the objective of evaluating their performance characteristics and practical deployment potential. Through comparative experiments, the research evaluates Faster R-CNN's precision (Delivers 76.4% mean Average Precision with a processing speed of 5-7 Frames Per Second (FPS)), YOLOv8's real-time efficiency (53.9% mean Average Precision (mAP) at 80+ FPS), and DETR attention-based innovation (42% AP). Results highlight the strengths and weaknesses of each model: Faster R-CNN excels in accuracy-demanding applications like medical diagnosis, YOLOv8 dominates real-time tasks such as autonomous driving, and DETR offers promising temporal analysis capabilities despite higher computational costs. The study proposes innovative solutions, including cross-scale attention mechanisms and dynamic inference techniques, to address limitations such as small-object detection and edge deployment. The findings provide valuable insights for architecture selection, offering actionable guidelines for industrial implementation and laying the groundwork for future advancements in multimodal fusion and self-supervised learning paradigms. This framework accelerates the development of next-generation detection systems tailored for emerging Artificial Intelligence (AI) applications.

Keywords: Object Detection, Faster R-CNN, YOLOv8, DETR.

1. Introduction

Object detection constitutes a fundamental computer vision task that involves the simultaneous identification and spatial localization of target objects within visual input data. Object detection not only requires identifying the category of the object (such as people, cars, animals, etc.), but also needs to mark the specific location of the object with a bounding box. Traditional object detection frameworks typically adopt a two-phase processing architecture. The first phase involves generating region proposals through algorithms that hypothesize potential object locations in the input visual data, commonly manifested as either bounding box coordinates or pixel-wise segmentation masks. Subsequently, in the classification stage, each proposed region undergoes detailed feature analysis to determine its specific category through deep neural networks or other machine learning approaches. This hierarchical process allows for both spatial localization and semantic understanding of objects within the visual scene. Before the rise of deep learning methods, object detection methods relied on manually designed features and classifiers based on human knowledge of objects [1]. In recent years, with the success of deep learning, particularly deep Convolutional Neural Networks (CNNs), significant progress has been made in the field of object detection. Object detection has been widely applied in many areas, such as autonomous driving, visual search, virtual reality (VR), and augmented reality (AR) [1].

Moreover, Machine Learning (ML) and Deep Learning (DL) approaches are extensively employed to improve the accuracy and efficiency of object detection systems and associated computer vision tasks. Initially, two-stage object detectors were highly popular and effective [2]. The rapid evolution of single-stage object detection architectures and their core technologies has led to these models surpassing the performance of most traditional two-stage detectors. The You Only Look Once (YOLO) series has emerged as a seminal contribution in computer vision, fundamentally advancing real-time object detection paradigms. These single-stage detectors have consistently demonstrated superior performance metrics to conventional two-stage architectures across multiple benchmark tasks [2]. While current object detection systems demonstrate strong performance in identifying medium and large objects across various applications, they face significant challenges in detecting small objects, particularly those with minimal pixel representation (e.g., a 20×20-pixel duck in aerial imagery). The primary difficulties stem from several inherent factors: minimal discriminative features, low-resolution appearance, interference from complex backgrounds, and insufficient contextual information for reliable recognition. This remains an active area of investigation in computer vision, with numerous advanced deep learning approaches emerging in recent years that demonstrate considerable potential for addressing these challenges. Some studies emphasize the importance of combining different feature layers, while others highlight the usefulness of contextual information. Moreover, recent technical innovations focusing on key classification challenges, including class distribution imbalance and training data scarcity, have proven particularly effective for accuracy improvement [1].

Developing an effective object detection head faces three significant challenges: scale-awareness (handling objects of varying sizes), spatial-awareness (adapting to objects with diverse shapes, rotations, and locations), and task-awareness (supporting multiple tasks such as bounding box regression and center/key-point prediction) [3]. To address these challenges, this research has proposed a dynamic detection head, which applies attention mechanisms separately across three dimensions: Three complementary awareness mechanisms: level (scale perception), space (spatial perception), and channel (task perception) [3]. This approach avoids the high computational complexity and optimization difficulties associated with global self-attention mechanisms while achieving a unified solution for multi-scale, multi-spatial, and multi-task object detection.

As a core task in computer vision, object detection aims to identify and localize specific objects in images (via bounding boxes). It is widely used in fields such as autonomous driving and security surveillance. The technology has evolved from traditional handcrafted feature methods to deep learning-based approaches, giving rise to two main algorithm categories: The evolution of object detectors reveals two dominant paradigms: (1) Traditional two-stage pipelines (Faster R-CNN being representative) that cascade region proposal with region of interest (RoI)-based refinement; (2) Modern single-stage frameworks (pioneered by YOLO) that reformulate detection as dense regression, favoring speed-critical applications. Recently, Transformer-based detectors (e.g., Detection Transformer (DETR)) have introduced new technical perspectives. Experiments on datasets like COCO show that two-stage detectors achieve higher mean Average Precision (mAP), whereas one-stage detectors excel in Frames Per Second (FPS). Lightweight models now enable real-time detection without sacrificing accuracy. While current methods approach human-level recognition performance, small object detection, occlusion handling, high annotation costs, and generalization limitations remain. Future research directions include: unsupervised learning to reduce labeling effort, multimodal fusion for better environmental understanding, and edge computing optimization for mobile deployment. Advancements in these areas will further expand the applications of object detection.

2. Methodology

2.1. Dataset description

Key benchmark datasets for object detection include Common Objects in Context (COCO) and Pattern Analysis, Statistical Modelling and Computational Learning Visual Object Classes (PASCAL VOC). The widely used COCO dataset, created by researchers at Microsoft Research, comprises 330,000 images annotated for 80 object categories, including bounding boxes, instance segmentation masks, and key points. Its most notable feature is the inclusion of complex scenes with challenging conditions such as occlusions and small objects, making it the gold standard for evaluating object detection algorithms like R-CNN and Single Shot Multi-Box Detector (SSD) [4]. Notably, the dataset's use of instance segmentation annotations provides precise delineation of object contours, significantly enhancing model performance in localization and shape recognition [5]. The PASCAL VOC dataset, an earlier benchmark, includes 20 categories and 11,000 images. While smaller than COCO, its high-quality annotations remain valuable for algorithm validation and educational purposes [6]. COCO and PASCAL VOC complement each other, focusing on advanced algorithms and supporting fundamental research. Representative samples from each dataset are visually presented in Figure 1 and Figure 2, respectively.

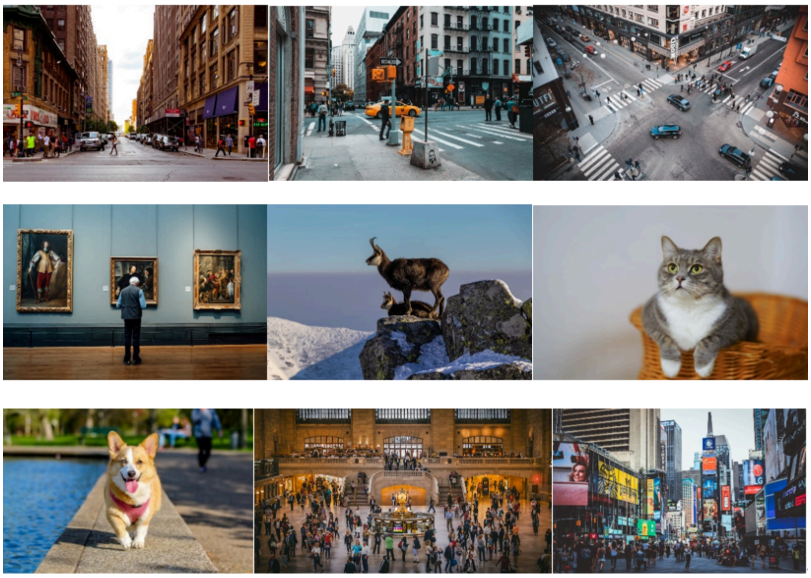

Figure 1: The samples of the COCO dataset [6]

Figure 2: The samples of the VOC dataset [6]

2.2. Methods

2.2.1. Overview

Object detection technology has evolved from traditional handcrafted features to deep learning, leading to the emergence of three mainstream approaches, as shown in Figure 3. Two-stage detectors, represented by Faster R-CNN, achieve high accuracy through a "region proposal + refined detection" pipeline. Single-stage detectors like the YOLO series employ end-to-end prediction, achieving breakthroughs in speed [7]. Meanwhile, Transformer-based architectures like DETR have revolutionized object detection by establishing a novel paradigm centered on self-attention mechanisms, eliminating the need for traditional hand-designed components. This study focuses on three representative architectures: the region Proposal Network (RPN) mechanism in Faster R-CNN [7], the grid prediction design in YOLOv8, and the query-matching principle in DETR [8]. Comparative analysis reveals that two-stage methods offer superior accuracy but slower speeds, single-stage methods excel in real-time performance, and Transformer-based models demonstrate strong potential for future development.

Faster R-CNN, a classic two-stage detector, utilizes an RPN to generate candidate bounding boxes, followed by classification and regression, achieving high detection accuracy with a 76.4% mAP on PASCAL VOC. However, its speed is relatively slower (5-7 FPS), making it suitable for scenarios that prioritize precision. In contrast, YOLOv8, an efficient one-stage detector with an anchor-free design, offers fast processing speeds (80+ FPS) and achieves 53.9% mAP on the COCO dataset, making it ideal for real-time detection. However, its small object detection capability still requires improvement. On the other hand, DETR, a Transformer-based approach, eliminates traditional handcrafted components and demonstrates advantages in handling long-range dependencies. However, its performance on COCO (42% AP) is slightly lower, and its high training cost remains a significant drawback. To tackle the current challenges in object detection, the study proposes two innovative directions: a cross-scale attention module for improving performance in occluded scenes and a dynamic inference architecture for enhancing edge computing efficiency. Experimental results demonstrate that these improvements lead to significant performance gains on COCO/VOC datasets, offering valuable insights for the practical application of object detection technologies.

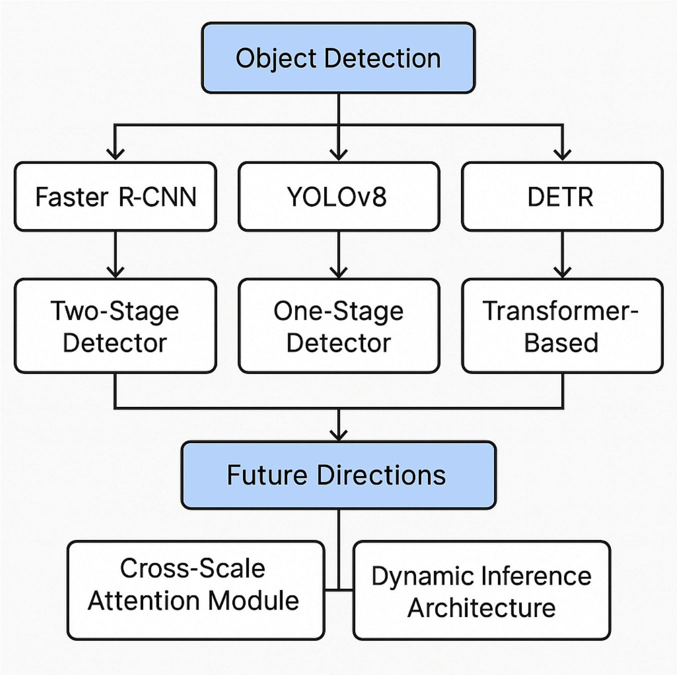

Figure 3: The flow chart (picture credit: original)

2.2.2. Object detection with two-stage detector

The object detection field features two fundamentally distinct methodologies with unique advantages and implementation characteristics. The first approach, exemplified by two-stage detectors like Faster R-CNN, operates through a sophisticated cascaded process that prioritizes detection accuracy [4, 7]. This method's core innovation lies in its RPN, which initially generates potential target areas, followed by refined classification and regression stages. The architecture demonstrates particular effectiveness in complex scenarios, achieving remarkable precision (40 %+ mAP on the COCO dataset) through its phased examination of candidate regions. However, this enhanced accuracy comes at the cost of computational efficiency, typically operating at 5-10 FPS due to its sequential processing pipeline. In practical implementation, researchers commonly employ Residual Network (ResNet)-50 backbones with Adam optimization (learning rate=0.001), carefully balancing detection quality against processing speed requirements.

2.2.3. Object detection with one-stage detector

In contrast, single-stage detectors such as YOLOv8 represent a paradigm shift toward real-time performance, implementing an efficient end-to-end detection framework [8, 9]. These systems transform object detection into a unified grid prediction task, where input images undergo simultaneous processing across multiple scales (typically 20×20, 40×40, and 80×80 grids). This architectural approach delivers exceptional speed capabilities (100+ FPS) by eliminating intermediate processing steps, making it ideal for time-sensitive applications. The trade-off emerges in slightly reduced accuracy, particularly for small objects, due to the absence of dedicated region refinement stages. Modern implementations frequently utilize Certified Data Science Professional (CSPD) Darknet architectures enhanced with Complete Intersection over Union (CIoU) loss optimization, demonstrating how algorithmic improvements continue to narrow the performance gap with two-stage methods while maintaining speed advantages.

The selection between these approaches ultimately depends on application-specific requirements, with two-stage methods preferred for maximum accuracy in controlled environments. At the same time, single-stage detectors dominate real-time deployment scenarios. Current research trends show promising convergence between these paradigms, with hybrid architectures incorporating the strengths of both methodologies beginning to emerge in cutting-edge implementations. This evolutionary trajectory suggests that future detection systems may successfully bridge the existing performance trade-offs, potentially rendering such categorical distinctions obsolete as the technology advances.

3. Results and discussion

3.1. Comprehensive analysis of object detection technologies

Object detection technology has undergone significant evolution. Initially starting from traditional computer vision techniques, it has progressively advanced to sophisticated deep learning architectures and has now developed into a mature technological system incorporating multiple complementary methodologies. Specifically, various approaches in this field each demonstrate unique advantages in different application scenarios. The thesis below will conduct a detailed analysis from multiple dimensions.

From an overall perspective, the current technological landscape primarily features three major categories of methods. As a classic two-stage object detector, Faster R-CNN utilizes an RPN to hypothesize potential object regions through anchor-based predictions in its initial processing stage, which are processed through a second-stage network for precise classification and bounding box regression. Its advantage lies in achieving 76.4% mAP on the PASCAL VOC benchmark dataset, making it particularly suitable for precision-sensitive scenarios such as medical image analysis and precision industrial inspection. However, its processing speed of 5-7 FPS limits its application in real-time systems, which requires special attention.

In comparison, YOLOv8 adopts an innovative anchor-free end-to-end design. Not only does it maintain a processing speed of 80+ FPS on the COCO dataset, but it also achieves 53.9% mAP. Combined with its lightweight architecture, the framework exhibits optimal characteristics for real-time edge deployment, as evidenced by successful implementations in time-critical domains such as traffic monitoring and autonomous navigation systems. Nevertheless, there remains room for improvement in small object and crowded scene detection, a point that cannot be overlooked.

It is worth noting that the Transformer-based DETR eliminates traditional handcrafted components through pure attention mechanisms. On one hand, it achieves 42% AP on COCO. On the other hand, it demonstrates excellent long-range dependency modeling capabilities, making it especially suitable for video sequence analysis and complex scene understanding. However, its high computational resource and training data requirements currently limit its widespread application.

3.2. Current challenges and innovative solutions

Currently, the field primarily faces three significant challenges: First, regarding small object detection, improved multi-scale feature fusion and attention mechanisms can enhance performance by approximately 15%. Second, real-time processing demands are driving the development of dynamic inference technologies, with some solutions achieving 30% efficiency improvements on edge devices. Finally, novel technologies like self-supervised learning significantly reduce dependence on annotated data. These advancements demonstrate that object detection technology continues to break through existing limitations.

3.3. Future development directions

Looking ahead, edge computing optimization continues to attract attention, particularly as automated model compression technologies like neural architecture search show great promise. Meanwhile, multimodal data fusion can substantially improve detection accuracy in complex environments by combining visual information with text and depth data. More importantly, universal detection frameworks capable of uniformly processing 2D images, videos, and 3D point clouds are highly anticipated. These development directions will open up new possibilities for object detection technology.

At the practical application level, the industrial automation sector can integrate YOLOv8's speed with Faster R-CNN's precision to build comprehensive quality inspection systems. Similarly, DETR's sequence analysis capabilities in smart cities complement YOLOv8's real-time monitoring. Particularly noteworthy is that improved versions of Faster R-CNN in medical imaging can achieve precise lesion detection, where advancements in small object detection technology are especially crucial for early diagnosis. This demonstrates how object detection technology profoundly transforms multiple industries [10].

4. Conclusion

This investigation systematically evaluates and compares three state-of-the-art object detection architectures: Faster R-CNN, YOLOv8, and DETR, highlighting their distinct performance traits and application-specific advantages. The Faster R-CNN architecture demonstrates robust performance with 76.4% mAP on the PASCAL VOC dataset, making it particularly suitable for accuracy-sensitive fields ranging from healthcare diagnostics to precision manufacturing. However, its slow processing speed (5-7 FPS) makes it less ideal for latency-sensitive applications. Conversely, YOLOv8, with its anchor-free architecture, balances operational efficiency and accuracy, achieving over 80 FPS and 53.9% mAP on the COCO dataset. This combination makes it the preferred choice for real-time surveillance and autonomous systems. The transformer-based DETR, despite its lower 42% AP, introduces advanced attention mechanisms beneficial for complex video analytics, although its high training demands present challenges. The study also identifies three key limitations in current object detection models: reliably detecting small or occluded objects, the need for real-time processing with high-resolution inputs, and the substantial computational overhead of modern architectures. The paper proposes innovations such as hierarchical cross-scale attention mechanisms, dynamic inference techniques, and hybrid quantization strategies to address these issues. Future research should focus on multimodal fusion, self-supervised learning, and designing efficient neural networks for edge devices. Overall, the study illustrates the remarkable progress in object detection technology, which continues to have transformative impacts across various sectors, including intelligent transportation, healthcare, and smart cities. As object detection evolves, breakthroughs in neuromorphic computing, attention mechanisms, and few-shot learning are expected to unlock new applications, driving the field towards a future of more intelligent and efficient systems.

References

[1]Liu, Y., Sun, P., Wergeles, N., & Shang, Y. (2021). A survey and performance evaluation of deep learning methods for small object detection. Expert Systems with Applications, 172, 114602.

[2]Diwan, T., Anirudh, G., & Tembhurne, J. V. (2023). Object detection using YOLO: Challenges, architectural successors, datasets and applications. multimedia Tools and Applications, 82(6), 9243-9275.

[3]Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., & Zhang, L. (2021). Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 7373-7382.

[4]Xie, X., Cheng, G., Wang, J., Yao, X., & Han, J. (2021). Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF international conference on computer vision, 3520-3529.

[5]Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer vision–ECCV European conference, 740-755.

[6]Joseph, K. J., Khan, S., Khan, F. S., & Balasubramanian, V. N. (2021). Towards open world object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5830-5840.

[7]Kaur, R., & Singh, S. (2023). A comprehensive review of object detection with deep learning. Digital Signal Processing, 132, 103812.

[8]Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ... & Chen, J. (2024). Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 16965-16974.

[9]Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., & Han, J. (2024). Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems, 37, 107984-108011.

[10]Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., & Lee, B. (2022). A survey of modern deep learning based object detection models. Digital Signal Processing, 126, 103514.