International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Object Detection Revisited: A Systematic Survey from Traditional Vision Pipelines to Transformer-Based Models

Mushtaq Ahmad Dar

Central University of Kashmir, Srinagar, India

Abstract

Object detection aims to localize and recognize objects within images, a challenging task due to variations in scale, occlusion, and scene complexity. Traditional feature-based methods were limited in adaptability, while CNN-based detectors improved performance but often rely on complex multi-stage pipelines. Transformer-based architectures reformulate detection as an end-to-end set prediction problem, leveraging attention mechanisms for global context. This paper systematically surveys object detection techniques from traditional to CNN- and transformer-based models. We analyze architectural choices, performance trade-offs, and real-world applications, using a fuzzy Multi-Criteria Decision Making (MCDM) framework to rank detectors based on speed (FPS) and accuracy (mAP). Open challenges related to efficiency, robustness, and data dependency are discussed, and future research directions including hybrid CNN–transformer models and lightweight architectures are highlighted.

Keywords: Object detection, deep learning, CNN, transformers, DETR, YOLO, fuzzy MCDM, computer vision

Introduction

Object detection is a fundamental problem in computer vision that involves identifying and localizing objects within an image. It is a core component of many real-world applications, including autonomous driving, video surveillance, healthcare imaging, and robotics. The task remains challenging due to variations in object scale, occlusion, viewpoint, and environmental conditions.

Early object detection methods relied on handcrafted features and classical machine learning models. Techniques such as Histogram of Oriented Gradients (HOG) combined with Support Vector Machines (SVM) and Deformable Parts Models (DPM) achieved reasonable performance but were limited in adaptability and robustness. The emergence of convolutional neural networks (CNNs) revolutionized the field by enabling end-to-end feature learning, leading to two-stage detectors like R-CNN, Fast R-CNN, and Faster R-CNN, as well as single-stage detectors such as SSD and YOLO. These models significantly improved accuracy and, in some cases, inference speed, but often rely on anchors, non-maximum suppression, and heuristic post-processing.

Recently, transformer-based architectures, such as DETR and Deformable DETR, have introduced end-to-end detection paradigms that leverage attention mechanisms to capture global contextual information, simplifying the detection pipeline and reducing dependency on handcrafted components. Despite these

advances, challenges remain in achieving high efficiency, scalability, and robustness, particularly for real-time applications and deployment on edge devices.

In this paper, we present a systematic survey of object detection techniques spanning traditional, CNN-based, and transformer-based models. We provide a comparative analysis using a fuzzy Multi-Criteria Decision Making (MCDM) framework based on speed (FPS) and accuracy (mAP), discuss key applications, and identify open challenges and future research directions. The proposed survey aims to guide researchers and practitioners in selecting appropriate detection models for diverse computer vision tasks.

Background and Related Work

Object detection combines two key tasks: *classification*, which determines what objects are present, and *localization*, which identifies where they are in an image. The evolution of object detection models can be broadly categorized into three stages: traditional feature-based approaches, CNN-based detectors, and transformer-based architectures. Each stage represents significant methodological advancements in addressing the challenges of scale variation, occlusion, and complex backgrounds.

Traditional Feature-Based Models

Early methods relied on handcrafted feature descriptors and classical machine learning models. Histogram of Oriented Gradients (HOG) captures local gradient patterns and, when combined with Support Vector Machines (SVM), achieved high pedestrian detection accuracy . Deformable Parts Models (DPM) represent objects as collections of parts with spatial deformation constraints, improving flexibility and robustness compared to rigid templates . While effective in controlled scenarios, these methods struggled with generalization to large-scale, diverse datasets.

CNN-Based Detectors

The advent of convolutional neural networks (CNNs) enabled end-to-end learning of hierarchical features. Two-stage detectors, such as R-CNN , Fast R-CNN, and Faster R-CNN , first generate region proposals and then classify and regress bounding boxes. Single-stage detectors like SSD and the YOLO family predict object locations and classes in a single forward pass, achieving a favorable trade-off between speed and accuracy. Despite these improvements, most CNN-based detectors rely on anchors, non-maximum suppression, and heuristic post-processing, which increase pipeline complexity and limit adaptability.

Transformer-Based Detectors

Transformer-based models, such as DETR and Deformable DETR , reformulate detection as a direct set prediction problem. By leveraging global attention mechanisms, these models capture long-range dependencies and remove the need for anchors or NMS. While they simplify the detection pipeline, transformers often require larger datasets and exhibit slower convergence compared to CNN-based models, motivating hybrid approaches and efficiency-focused variants.

Fuzzy Multi-Criteria Decision Making in Object Detection

Evaluating detectors solely based on accuracy or speed can be limiting. Fuzzy Multi-Criteria Decision Making (MCDM) provides a structured approach to rank detectors by combining multiple performance

metrics. In this survey, we define fuzzy sets for speed (FPS) and accuracy (mAP) to compute a defuzzified ranking score for each detector. This methodology offers a quantitative comparison across heterogeneous models, guiding the selection of appropriate detectors for different application scenarios.

Summary of Related Surveys

Several surveys have focused on specific object detection models or deep learning approaches, but few provide a comprehensive comparison across traditional, CNN, and transformer-based detectors using a quantitative ranking framework. Our work addresses this gap by integrating a systematic review, fuzzy MCDM analysis, and discussion of real-world applications and open challenges.

Comparison of Object Detectors and Fuzzy MCDM Analysis

To assist in selecting appropriate object detection models, we provide a comparative evaluation of representative detectors spanning CNN-based and transformer-based architectures. The comparison considers both accuracy (mAP on COCO dataset) and inference speed (FPS), reflecting the trade-offs between performance and efficiency in practical applications.

Representative Detector Performance

Representative Object Detectors: FPS and COCO mAP

Model	Architecture	FPS	mAP (COCO)
Faster R-CNN	Two-stage CNN	7	42
SSD	Single-stage CNN	22	31
YOLOv3	Single-stage CNN	45	33
YOLOv7	Single-stage CNN	60	51.4
DETR	Transformer	28	42
Deformable DETR	Transformer	40	50

Strengths and Limitations

Strengths and Weaknesses of Major Detectors

Model	Strengths	Weaknesses
R-CNN	High accuracy	Slow; high computation
YOLO	Real-time inference	Struggles with small objects
SSD	Balanced speed and accuracy	Less accurate than YOLOv7
DETR	End-to-end, no NMS	Slow convergence
Deformable DETR	Handles multi-scale features	Computationally intensive

Fuzzy MCDM Analysis

Fuzzy sets for speed (FPS) and accuracy (mAP) are defined to combine multiple criteria into a single defuzzified ranking score. The results are presented in Table 3.

Defuzzified Fuzzy Ranking of Detectors

Model	Score (Combined)
YOLOv7	0.556

YOLOv3	0.495
Deformable DETR	0.434
DETR	0.330
SSD	0.195
Faster R-CNN	0.060

Fuzzy Logic Based Model Evaluation

Input Variable

Combined Score

Range:0 -> **1**

Fuzzy Sets

We define 5 linguistic variables:

Fuzzy Set	Range
Very Low	0.0 – 0.15
Low	0.1 – 0.35
Medium	0.3 – 0.55
High	0.5 – 0.75
Very High	0.7 – 1.0

Fuzzy Rule Base

Rule No.	IF Combined Score is	THEN Performance is
R1	Very Low	Poor
R2	Low	Below Average
R3	Medium	Average
R4	High	Good
R5	**Very High**	**Excellent**

Fuzzy Classification of Models

Model	Score	Fuzzy Category	Performance
YOLOv7	0.556	High	Good
YOLOv3	0.495	Medium	Average
Deformable DETR	0.434	Medium	Average
DETR	0.330	Low/Medium	Below Average
SSD	0.195	Low	Below Average
Faster R-CNN	0.060	Very Low	Poor

Membership Functions

Very Low

$$\mu_{VL}(x) = \begin{cases} 1 & x \le 0.05 \\ \dfrac{0.15 - x}{0.10} & 0.05 < x < 0.15 \\ 0 & x \ge 0.15 \end{cases}$$

Low

$$\mu_L(x) = \begin{cases} 0 & x \le 0.10 \\ \dfrac{x - 0.10}{0.15} & 0.10 < x < 0.25 \\ \dfrac{0.35 - x}{0.10} & 0.25 \le x < 0.35 \\ 0 & x \ge 0.35 \end{cases}$$

Medium

$$\mu_M(x) = \begin{cases} 0 & x \le 0.30 \\ \dfrac{x - 0.30}{0.125} & 0.30 < x < 0.425 \\ \dfrac{0.55 - x}{0.125} & 0.425 \le x < 0.55 \\ 0 & x \ge 0.55 \end{cases}$$

High

$$\mu_H(x) = \begin{cases} 0 & x \le 0.50 \\ \dfrac{x - 0.50}{0.125} & 0.50 < x < 0.625 \\ \dfrac{0.75 - x}{0.125} & 0.625 \le x < 0.75 \\ 0 & x \ge 0.75 \end{cases}$$

Fuzzy Classification of Models

Fuzzy classification results

Model	Score	Fuzzy Category	Performance
YOLOv7	0.556	High	Good
YOLOv3	0.495	Medium	Average
Deformable DETR	0.434	Medium	Average
DETR	0.330	Low/Medium	Below Average
SSD	0.195	Low	Below Average
Faster R-CNN	0.060	Very Low	Poor

Defuzzified Ranking

Rank	Model	Final Fuzzy Decision
1	YOLOv7	Good
2	YOLOv3	Average
3	Deformable DETR	Average
4	DETR	Below Average
5	SSD	Below Average
6	Faster R-CNN	Poor

The fuzzy inference system indicates that:

- YOLOv7 achieved the highest fuzzy performance classification (“Good”).
- YOLOv3 and DETR variants achieved “Average” performance.
- SSD and Faster R-CNN showed lower suitability under the combined evaluation metric.

Applications

Autonomous Driving

Real-time detection of pedestrians, vehicles, lanes, and traffic signs is critical for autonomous vehicles. High-speed detectors such as YOLOv7 are preferred for on-road scenarios, while transformer-based models can be used for detailed scene understanding.

Healthcare

Medical imaging applications, including tumor, organ, and lesion detection, require high accuracy. Transformer-based architectures, particularly Deformable DETR, provide enhanced feature representation for complex structures.

Retail

Automated checkout, inventory monitoring, and shelf management benefit from real-time detection. Single-stage CNN detectors like YOLO and SSD offer a good balance between speed and accuracy.

Security and Surveillance

Monitoring of restricted areas, anomaly detection, and crowd analysis demand efficient and reliable detectors. Fuzzy MCDM analysis can guide the selection of models that satisfy both speed and precision requirements.

Challenges and Future Directions

Despite significant progress, several challenges remain:

- **Real-Time Inference on Edge Devices:** Lightweight architectures are required for deployment on mobile and embedded platforms.
- **Data Efficiency:** Most deep learning detectors require large annotated datasets; semi-supervised and self-supervised learning approaches can reduce data dependency.
- **Robustness:** Detectors must generalize across domain shifts, lighting variations, and occlusions.
- **Efficient Transformer-Based Models:** Transformers offer high accuracy but are computationally intensive; hybrid CNN-transformer architectures can balance efficiency and performance.

Future research includes developing lightweight transformer models, exploring unsupervised pretraining, and combining fuzzy decision-making frameworks with multi-objective optimization.

Conclusion

This paper presents a comprehensive survey of object detection models, tracing the evolution from handcrafted features to CNN-based and transformer-based architectures. We performed a structured comparison using fuzzy MCDM analysis to evaluate speed and accuracy trade-offs. Our study highlights the strengths and limitations of each detection paradigm and identifies open challenges, particularly in efficiency, robustness, and data dependency. Future research should focus on hybrid CNN–transformer models, lightweight architectures, and adaptive learning approaches to advance practical, high-performance object detection systems.

References

1. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *CVPR*, 2005.
2. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Trans. PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.
3. R. Girshick, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *CVPR*, 2014.

4. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *NIPS*, 2015.
5. W. Liu et al., "SSD: Single shot multibox detector," *ECCV*, 2016.
6. J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, arXiv:1804.02767.
7. N. Carion et al., "End-to-end object detection with transformers," *ECCV*, 2020.
8. X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," *ICLR*, 2021.