Title: A Survey of Object Detection: From Region Proposals to End-to-End Transformers

Author: [Your Name/Institution] **Date:** October 2025

Abstract

Object detection, a fundamental task in computer vision, involves identifying and localizing instances of objects within an image or video. It goes beyond simple image classification by not only determining the class of an object but also providing a bounding box that precisely outlines its location. This paper provides a comprehensive survey of the evolution of object detection methodologies, primarily focusing on the deep learning era. We begin by contextualizing the problem with a brief overview of traditional computer vision techniques. The core of the review is dedicated to the two dominant paradigms in deep learning-based detectors: **two-stage detectors**, exemplified by the R-CNN family (R-CNN, Fast R-CNN, Faster R-CNN), which prioritize accuracy through a region proposal mechanism; and **single-stage detectors**, such as YOLO and SSD, which optimize for speed by performing detection in a single pass. We then explore key architectural components like backbone networks, anchor boxes, and non-maximum suppression. The survey culminates with a discussion of modern architectures, including the paradigm-shifting DETR (DEtection TRansformer), which reframes object detection as an end-to-end set prediction problem. Finally, we cover standard evaluation metrics, common datasets, real-world applications, and the ongoing challenges and future directions that are shaping the field.

Keywords: Object Detection, Computer Vision, Deep Learning, R-CNN, YOLO, SSD, Transformer, DETR, Bounding Box, mAP.

Table of Contents

1. Introduction

1.1. Defining Object Detection: Classification and Localization

Object detection is a core computer vision task concerned with answering two fundamental questions about an image: **"What objects are in this image?"** and **"Where are they located?"**. The first question is a **classification** task, assigning a class label (e.g., "cat," "car," "person") to an object. The second is a **localization** task, providing a tight-fitting **bounding box** (typically defined by x/y coordinates and width/height) around each identified object.

1.2. Distinction from Other Vision Tasks

It is crucial to distinguish object detection from related tasks:

- **Image Classification:** Simply assigns one label to an entire image (e.g., "this is a picture of a cat").
- **Semantic Segmentation:** Assigns a class label to every pixel in the image but does not distinguish between different instances of the same object (e.g., all pixels belonging to any person are labeled "person").
- **Instance Segmentation:** Assigns a class label to every pixel *and* differentiates between object instances (e.g., "person 1," "person 2," "person 3"). Object detection can be seen as a precursor to this more complex task.

1.3. The Importance of Object Detection

The ability to detect and locate objects is foundational to how machines perceive and interact with the physical world. It is the technology that enables self-driving cars to see pedestrians and other vehicles, allows doctors to identify tumors in medical scans, and helps robots navigate complex environments. Its broad applicability has made it one of the most actively researched areas in artificial intelligence.

1.4. Paper Structure

This paper will trace the evolution of object detection methods, beginning with a brief look at pre-deep learning techniques. We will then delve into the two primary families of deep learning

detectors: two-stage and single-stage. We will discuss their core components, modern architectures including Transformers, and conclude with evaluation metrics, applications, and future challenges.

---

2. Background and Foundational Concepts

2.1. Traditional Computer Vision Approaches

Before deep learning, object detection relied on hand-crafted features. Methods like the **Viola-Jones framework** (famous for real-time face detection) used simple Haar-like features and a cascade of classifiers. Other approaches used more complex feature descriptors like **HOG (Histogram of Oriented Gradients)**, often paired with a classifier like a Support Vector Machine (SVM), to identify objects. These methods were effective for specific tasks but were brittle and did not generalize well.

2.2. The Sliding Window Method

A common technique was the **sliding window** approach. A window of a fixed size would be slid across all possible locations and scales of an image. For each window, a feature descriptor would be computed and fed to a classifier. This method was computationally exhaustive and prone to errors.

2.3. The Deep Learning Revolution

The success of AlexNet in the 2012 ImageNet classification challenge marked a turning point. Researchers quickly realized that the rich, hierarchical features learned automatically by **Convolutional Neural Networks (CNNs)** were far more powerful than any hand-crafted features. This discovery paved the way for the modern era of object detection.

---

3. Two-Stage Object Detectors: A Focus on Accuracy

Two-stage detectors break the object detection problem into two distinct steps, a paradigm that generally leads to higher accuracy at the cost of speed.

3.1. The "Propose, then Classify" Paradigm

The core idea is to first generate a sparse set of **region proposals**—areas of the image that are likely to contain an object. In the second stage, a classifier is run only on these proposed regions to determine the object's class and refine the bounding box.

3.2. R-CNN: Regions with CNN Features

R-CNN was the first major breakthrough in applying deep learning to this paradigm. However, its process was slow and cumbersome:

1. Generate ~2000 region proposals using an external algorithm like Selective Search.
2. Warp/resize each proposed region to a fixed size.
3. Pass each warped region independently through a pre-trained CNN to extract features.
4. Use a set of SVMs to classify the object in each region.

3.3. Fast R-CNN: Sharing Computation

Fast R-CNN made a significant improvement. Instead of running the CNN 2000 times, it passes the **entire image** through the CNN just once to generate a feature map. The region proposals are then projected onto this feature map. A novel **RoI (Region of Interest) Pooling** layer extracts a fixed-size feature vector from each proposed region, which is then fed into a classifier. This shared computation made the process much faster.

3.4. Faster R-CNN: The Region Proposal Network (RPN)

The bottleneck in Fast R-CNN was the external Selective Search algorithm for proposing regions. Faster R-CNN introduced the **Region Proposal Network (RPN)**, a small neural network that learns to generate high-quality region proposals directly from the CNN features. By integrating the RPN, Faster R-CNN became the first truly end-to-end, unified deep learning object detector, setting a new standard for accuracy.

4. Single-Stage Object Detectors: A Focus on Speed

Single-stage detectors remove the region proposal step and instead perform localization and classification in a single forward pass of the network, making them extremely fast and suitable for real-time applications.

4.1. The "Single Pass" Paradigm

These models treat object detection as a regression problem. They look at the image once and directly predict a set of bounding boxes and their corresponding class probabilities.

4.2. YOLO: You Only Look Once

The YOLO family of models is renowned for its speed. YOLO divides the input image into a grid. For each grid cell, the model simultaneously predicts:

- Several bounding boxes.
- A "confidence" score for each box, indicating how likely it is to contain an object.
- Class probabilities for the object within the box. This unified architecture allows for end-to-end training and blazingly fast inference speeds, making it ideal for video processing.

4.3. SSD: Single Shot MultiBox Detector

SSD aimed to find a middle ground between the speed of YOLO and the accuracy of Faster R-CNN. Its key innovation is using feature maps from multiple layers of the backbone network to make predictions. By making predictions at different scales, SSD is much better at detecting objects of various sizes, particularly small ones, compared to the original YOLO.

5. Key Architectural Components and Innovations

Modern detectors, whether two-stage or single-stage, share several common components.

5.1. Backbone Networks: The Feature Extractors

The **backbone** is a deep CNN (like ResNet, VGG, or MobileNet) pre-trained on a large image classification dataset (e.g., ImageNet). Its role is to act as a powerful feature extractor, converting the raw pixel data of an image into rich, hierarchical feature maps that can be used for detection.

5.2. Anchor Boxes: The Priors for Prediction

Instead of predicting bounding boxes from scratch, most detectors predict offsets relative to a set of pre-defined default boxes called **anchor boxes**. These anchors have various sizes and aspect ratios

and are tiled across the image at different locations. Using anchors reframes the problem from predicting absolute coordinates to refining a well-placed prior, which makes learning easier for the network.

5.3. Non-Maximum Suppression (NMS): Pruning Redundant Detections

A detector will often output multiple, highly overlapping bounding boxes for the same object. **NMS** is a crucial post-processing step that cleans up these redundant detections. It sorts all boxes by their confidence scores, keeps the box with the highest score, and suppresses (discards) any other boxes that have a high overlap with it.

6. Modern Architectures and the Rise of Transformers

6.1. Balancing Speed and Accuracy: EfficientDet

The EfficientDet family of models introduced a systematic way to scale detectors for different resource constraints. It uses a highly efficient backbone (EfficientNet) and a novel feature fusion mechanism (BiFPN) to achieve state-of-the-art efficiency, balancing high accuracy with low computational cost.

6.2. DETR: End-to-End Object Detection with Transformers

DETR (DEtection TRansformer) represents a major paradigm shift. It completely eliminates the need for hand-crafted components like anchor boxes and NMS. DETR uses a standard Transformer encoder-decoder architecture, similar to those used in NLP. It treats object detection as a **direct set prediction problem**: the model ingests image features and directly outputs the final set of unique object detections. This simplifies the detection pipeline significantly and has opened up a new and exciting research direction.

---

7. Evaluation Metrics and Datasets

7.1. Intersection over Union (IoU)

IoU is the fundamental metric used to measure the "correctness" of a predicted bounding box. It is calculated as the area of overlap between the predicted box and the ground-truth box, divided by the area of their union. A detection is typically considered a "true positive" if its IoU with a ground-truth box is above a certain threshold (e.g., 0.5).

7.2. Average Precision (AP) and mean Average Precision (mAP)

Average Precision (AP) is the primary metric for evaluating the performance of a detector on a single object class. It is calculated from the precision-recall curve and effectively measures the detector's accuracy across all confidence levels. **Mean Average Precision (mAP)** is the average of the AP values across all object classes and is the standard metric for comparing different object detection models.

7.3. Landmark Datasets

The field has been driven by large-scale, publicly available datasets, most notably **PASCAL VOC** and **Microsoft COCO (Common Objects in Context)**. The COCO dataset, with its large number of object categories and instances per image, is the current benchmark for modern object detectors.

8. Applications and Real-World Use Cases

Object detection is a deployed and impactful technology across numerous industries.

- **Autonomous Vehicles:** Detecting cars, pedestrians, cyclists, and traffic signals is essential for safe navigation.
- **Medical Imaging:** Assisting radiologists by automatically locating tumors, lesions, or other anomalies in X-rays, CT scans, and MRIs.
- **Retail:** Powering cashier-less stores, monitoring shelf inventory, and analyzing customer foot traffic.
- **Security and Surveillance:** Automatically detecting intruders, unattended baggage, or monitoring crowd density.

9. Challenges and Future Directions

Despite immense progress, several challenges remain.

- **Detecting Small and Occluded Objects:** Models still struggle to reliably detect objects that are very small, far away, or partially hidden.
- **The Speed vs. Accuracy Trade-off:** While models are becoming more efficient, the fundamental trade-off between real-time speed and maximum accuracy remains a key design consideration.
- **Domain Adaptation and Generalization:** A model trained on daytime, sunny weather data may fail when deployed at night or in the rain. Improving robustness to new environments is a major challenge.
- **Few-Shot and Zero-Shot Detection:** Training models to detect new object categories with very few (or zero) labeled examples is an active and important area of research.

10. Conclusion

Object detection has undergone a remarkable transformation, moving from slow, brittle systems based on hand-crafted features to highly accurate and efficient end-to-end deep learning models. The evolution from the methodical two-stage R-CNN family to the rapid single-stage YOLO and SSD detectors, and now to the elegant, anchor-free Transformer-based models like DETR, showcases the field's rapid pace of innovation. As a core enabling technology for machine perception, object detection continues to solve critical real-world problems and will undoubtedly remain a central focus of AI research for years to come.

---

11. References

[Viola & Jones, 2001] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[Girshick et al., 2014] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*.

[Redmon et al., 2016] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[Liu et al., 2016] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *European conference on computer vision*.

[Carion et al., 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *European conference on computer vision*.