Proceedings of CONF-MLA Workshop: Mastering the Art of GANs: Unleashing Creativity with Generative Adversarial Networks
DOI: 10.54254/2755-2721/80/2024CH0058

A Review of Research on Object Detection Algorithms

Xiangjin Kong

Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China

xk2003@hw.ac.uk

Abstract. Object detection is a fundamental task in computer vision, widely used in fields such as autonomous driving, security surveillance, medical imaging, and drone image analysis. With the continuous advancement of technology, object detection algorithms have evolved from traditional methods to deep learning approaches. This paper categorizes object detection algorithms into four types based on their technical characteristics and implementation methods: two-stage algorithms, one-stage algorithms, keypoint-based algorithms, and emerging Transformer-based methods. Through a performance comparison on existing datasets, it was found that two-stage algorithms excel in accuracy but consume significant computational resources, leading to slower speeds; one-stage algorithms have a clear advantage in speed but show lower accuracy in detecting small objects; keypoint-based methods effectively balance speed and accuracy; additionally, the emerging Transformer-based methods perform well in capturing global information but require large amounts of training data and computational resources. This paper summarizes the advantages and disadvantages of each type of algorithm and discusses future research directions.

Keywords: Object Detection, Deep Learning, Two-Stage, One-Stage, Transformer.

1. Introduction

As one of the core tasks in computer vision, object detection plays a crucial role in various fields. The goal of object detection is to identify specific object classes within images or video sequences and to locate their positions. It is widely used in areas such as autonomous driving, security surveillance, medical imaging, and drone image analysis. With the continuous advancement of technology, object detection algorithms have evolved from traditional methods to deep learning approaches and are now moving towards lightweight and multimodal development.

Before the rise of deep learning, object detection primarily relied on manually designed feature extraction methods and classifiers. In the 1990s, the sliding window approach was widely used, which involved sliding a window of different sizes and steps across an image and classifying each window. This method had high computational complexity and low detection accuracy. Later, in 2005, Dalal and Triggs proposed a method based on Histogram of Oriented Gradients (HOG) features and a Support Vector Machine (SVM) classifier for pedestrian detection. HOG features describe the distribution of local gradient orientations in an image, while the SVM is used to classify these features. Although the HOG+SVM method performed well in pedestrian detection, it had limited capability in handling complex scenes and multi-class object detection [1].

DOI: 10.54254/2755-2721/80/2024CH0058

In recent years, thanks to advancements in deep learning technology, particularly the development of Convolutional Neural Networks (CNNs), object detection has gradually integrated with deep learning, sparking a new wave of development. Girshick et al. (2014) were the first to apply deep learning to object detection using a region-based convolutional neural network (R-CNN), which significantly improved the accuracy of object detection [2]. Subsequently, numerous deep learning-based object detection algorithms have emerged, showing substantial improvements in both accuracy and computational efficiency. This paper posits that current deep learning-based object detection algorithms can be broadly categorized into the following types: two-stage methods, one-stage methods, anchor-free methods, and the emerging Transformer-based methods. While the various algorithms available today exhibit strong performance, each has its own shortcomings. Additionally, object detection currently faces many new challenges, such as detecting small objects and detecting objects in three-dimensional space. Moreover, improving model generalization, enhancing the understanding of complex scenes, and minimizing the resource consumption of model training have gradually become key research topics in the present and future.

This paper will review the development of object detection by focusing on the aforementioned categories, organized both by classification and chronological order. It will summarize the advantages and limitations of each type of existing algorithm and discuss the challenges and future trends in object detection. The aim is to provide researchers in related fields with a comprehensive understanding framework and potential future research directions.

2. Key Technologies Overview

2.1. Two-Stage Methods

Two-stage methods achieve object detection by generating candidate regions and then performing classification and regression on each region. In 2014, Ross Girshick introduced R-CNN, which generates candidate regions using selective search, then uses a CNN to extract features from each region, and employs an SVM for classification. R-CNN improved detection accuracy but had slow processing speed, as each candidate region needed to be processed individually [2]. Subsequently, in 2015, Girshick proposed Fast R-CNN, which introduced the Region of Interest (RoI) pooling layer and shared the computation results of convolutional layers, thereby improving processing speed. However, it still relied on external methods for generating candidate regions [3]. In 2016, Shaoqing Ren et al. further introduced the Region Proposal Network (RPN) and proposed Faster R-CNN, achieving end-to-end training and significantly enhancing detection speed and accuracy. RPN could directly generate candidate regions from the image, unifying region proposal and object detection, though it still consumed considerable computational resources [4].

Following this, Mask R-CNN was introduced, adding a segmentation branch to Faster R-CNN for instance segmentation [5]. Additionally, Cascade R-CNN, proposed in 2019, performed multi-stage object detection and regression in a cascade manner, using multiple cascaded detectors with progressively increasing thresholds. This approach refined the detection results step by step, thereby improving both accuracy and recall rate [6].

2.2. Single-stage Methods

Single-stage methods approach object detection as a regression problem, directly generating the categories and bounding boxes of objects from the input image. To improve the speed of object detection, Joseph Redmon proposed You Only Look Once (YOLO) in 2016, which divides the image into grids, with each grid directly predicting bounding boxes and class labels [7]. YOLO has a clear advantage in terms of speed, capable of real-time detection, but it performs poorly when detecting small objects and dealing with complex scenes. The workflow of YOLO includes dividing the image into an SxS grid, where each grid predicts B bounding boxes along with their confidence scores and class probabilities, and redundant bounding boxes are removed via non-maximum suppression (NMS). Following this, Wei Liu et al. proposed Single Shot MultiBox Detector (SSD) in 2016, which combines regression with a

DOI: 10.54254/2755-2721/80/2024CH0058

convolutional feature pyramid structure to perform detection from feature maps at different scales, ensuring both speed and accuracy, particularly excelling in handling multi-scale objects [8]. However, SSD still has limitations when dealing with very small objects. Additionally, EfficientDet is a new approach that combines EfficientNet and Bidirectional Feature Pyramid Network (BiFPN), enhancing detection accuracy while maintaining computational efficiency [9].

2.3. Anchor-free Methods

Anchor-free methods, which is also known as keypoint-based methods, achieve object detection by detecting keypoints of objects, thus avoiding the complexity involved in generating candidate regions in traditional approaches. In 2018, Hei Law and Jia Deng introduced CornerNet, the first algorithm that infers bounding boxes by detecting object corner points [10]. The innovation of CornerNet lies in defining bounding boxes through a pair of keypoints (the top-left and bottom-right corners of an object). This method simplifies the process of detection and effectively improves the accuracy of object detection. However, CornerNet's inference speed is relatively slow because it needs to handle multiple keypoints and combine them afterward. To further enhance detection efficiency, in 2019, Xingyi Zhou et al. proposed CenterNet. Unlike CornerNet, CenterNet directly predicts bounding boxes by detecting the center point (centroid) of objects [11]. The core advantage of CenterNet is its simplicity; by reducing the number of keypoints and directly regressing the size and offset of the targets, it achieves higher inference speeds and better real-time performance. Simultaneously, for finer details in object detection, Jingdong Wang et al. introduced High-Resolution Network (HRNet) in 2019 [12]. Although initially designed for pose estimation, its unique high-resolution feature extraction method makes it equally effective in object detection tasks, though its complex model structure results in relatively slower inference speeds.

2.4. Emerging Transformer-Based Methods

In recent years, Transformer methods have been introduced into the field of object detection. In 2020, Nicolas Carion et al. from Facebook AI introduced Detection Transformer (DETR), which leverages the Transformer architecture combined with features extracted by CNNs to perform object detection via self-attention mechanisms [13]. The process of DETR involves using a convolutional network to extract image features, after which positional encodings are added to these features to preserve spatial information. These features are then fed into a Transformer encoder-decoder, which generates the object detection results. DETR can perform detection directly without requiring region proposals, simplifying the detection process. It excels particularly in capturing global information and handling complex scenes due to its self-attention mechanism, which allows DETR to effectively capture global dependencies within images, making it powerful in complex scenarios. More recently, Transformer-based methods have also included Deformable DETR, which addresses issues with small object detection and training efficiency in the original DETR by introducing deformable convolutions [14].

3. Methods

3.1. Common Datasets

In the evaluation of object detection algorithms, commonly used datasets include:

- COCO: Covers a variety of objects found in everyday life and provides rich multi-object detection scenarios. The COCO dataset includes 80 common object categories, with multiple instances per category, making it suitable for evaluating the general performance of detection algorithms [15].
- Pascal VOC: A classic dataset primarily used for evaluating basic object detection algorithms. The Pascal VOC dataset comprises 20 common object categories and provides standardized training and testing sets, serving as the main dataset for early object detection research [16].
- ImageNet: A large-scale dataset mainly used for image classification, but also includes an object detection task. The ImageNet dataset covers over 1000 object categories and provides a wealth of

DOI: 10.54254/2755-2721/80/2024CH0058

images and annotation information, suitable for evaluating large-scale object detection algorithms [17].

- KITTI: Focuses on autonomous driving scenarios and provides high-quality street scene images. The KITTI dataset includes various types of road objects such as vehicles, pedestrians, cyclists, etc., making it suitable for evaluating object detection algorithms in the context of autonomous driving [18].
- DOTA: Used for object detection in aerial imagery, containing objects under various complex backgrounds. The DOTA dataset includes a variety of aerial imagery objects such as planes, ships, vehicles, etc., and is suitable for evaluating object detection algorithms in high-resolution images [19].

3.2. Evaluation Metrics

In object detection tasks, a series of evaluation metrics are typically used to assess the performance of models. A key comprehensive performance metric is Average Precision (AP), which combines Precision and Recall. Precision measures the proportion of true positives among all instances predicted as positive; whereas Recall measures the proportion of true positives among all actual positive instances. AP is obtained by approximating the area under the Precision-Recall curve. To further evaluate the performance across multiple categories, the paper calculates the mean of APs across all categories, known as Mean Average Precision (mAP). In this paper, to provide a more detailed assessment of the model's ability to capture global information, special attention is given to separately listing the average precision for large, medium, and small objects. This not only reflects the overall performance of the model but also highlights its characteristics and limitations in handling targets of specific scales, thus providing a more comprehensive perspective for algorithm optimization and application selection.

3.3. Algorithm Performance Comparison

When comparing various object detection algorithms, this paper selects the top-performing and most representative methods within each category for comparison and draws conclusions accordingly. Specific data is presented as shown in Table 1:

Table 1. Algorithm Performance

Algorithm	Category	Dataset	Small Objects AP	Medium Objects AP	Large Objects AP	Overall mAP	Speed (FPS)
R-CNN [2]	Two-stage	PASCAL VOC 2012	-	-	-	58.5	-
Fast R-CNN [3]	Two-stage	PASCAL VOC 2012	-	-	-	70.0	-
Faster R-CNN [4]	Two-stage	COCO	18.2	39.9	49.9	36.2	5.0
Mask R-CNN [5]	Two-stage	COCO	18.3	41.6	50.9	37.1	3.5
Cascade R-CNN [6]	Two-stage	COCO	22.1	45.4	55.2	42.8	4.1
YOLOv1 [7]	One-stage	PASCAL VOC 2007	-	-	-	63.4	45.0
YOLOv3	One-stage	COCO	16.2	37.5	44.8	33.0	30.0
YOLOv5	One-stage	COCO	18.5	40.5	47.3	36.0	70.0
SSD512 [8]	One-stage	COCO	15.3	33.2	43.5	26.8	19.0
RetinaNet (ResNet-101)	One-stage	COCO	26.1	46.7	53.2	39.1	5.0

DOI: 10.54254/2755-2721/80/2024CH0058

Table 1. (continued).

EfficientDet-D0 [9]	One-stage	COCO	21.5	45.1	52.5	37.0	67.0
CenterNet [11]	Anchor-free	COCO	24.2	46.5	52.3	47.0	25.0
CornerNet [10]	Anchor-free	COCO	22.1	44.1	50.4	42.2	15.0
HRNet [12]	Anchor-free	COCO	25.5	47.0	53.6	75.1	12.0
DETR [13]	Transformer	COCO	20.3	45.8	62.4	42.0	28.0
Deformable DETR [14]	Transformer	COCO	22.5	48.0	63.7	43.8	29.0

Through Table 1, the following analysis can be drawn:

- **Trade-off between accuracy and speed:** Two-stage algorithms, such as Faster R-CNN, although highly accurate, consume large amounts of computational resources and struggle to meet real-time requirements. Cascade R-CNN further improves detection accuracy through cascaded detectors, achieving a mAP of 42.8% on the COCO dataset. In contrast, single-stage algorithms like YOLO and SSD stand out in terms of speed but fall short in detection accuracy, especially for small objects. YOLOv5, for instance, has an AP of 18.5% for small objects and an overall mAP of 36.0% on the COCO dataset, demonstrating improvements in both speed and accuracy.
- **Balance of Anchor-free Algorithms:** Algorithms such as CenterNet and CornerNet detect key points (such as the center or corners of objects) to regress bounding boxes, serving as an effective complement to region-based and regression methods. While maintaining relatively high accuracy, their inference speed is typically faster than that of two-stage algorithms. However, in extremely complex scenes or with very small objects, they may not perform as well as two-stage algorithms and transformer-based approaches. These keypoint-based methods seek a balance between detection accuracy and speed through various strategies, progressively enhancing the real-time capability and accuracy of object detection. As these methods continue to evolve, their applicability in practical scenarios is also steadily improving.
- **Global Information Capture by Transformers:** Emerging Transformer methods, such as DETR, excel in capturing global information, overcoming the limitations of traditional Convolutional Neural Networks (CNNs) in capturing long-range dependencies through their self-attention mechanism. DETR achieves an overall mAP of 42.0% on the COCO dataset, with an AP of 62.4% for large objects, 45.8% for medium objects, and 20.3% for small objects. In contrast, RetinaNet demonstrates its superiority in detecting small objects with an AP of 23.1% for small objects, 44.2% for medium objects, and 51.2% for large objects on the COCO dataset. Deformable DETR improves performance in small object detection and training efficiency by incorporating deformable convolutions, achieving an overall mAP of 43.8%, with an AP of 63.7% for large objects, 48.0% for medium objects, and 22.5% for small objects.

In summary, this paper outlines the advantages and disadvantages of different types of algorithms, as shown in Table 2:

Table 2. Summary of Algorithm Advantages and Disadvantages

Category	Advantages	Disadvantages
Two-stage	High accuracy	Slow
One-stage	Very fast speed	Insufficient accuracy
Anchor-free	Balances accuracy and speed	Insufficient accuracy in some scenarios
Transformer	Capable of capturing global information	High computational resource requirements

DOI: 10.54254/2755-2721/80/2024CH0058

4. Challenges and Prospects

In this paper, the paper has reviewed various implementations of object detection technologies and their performance. After evaluating traditional methods, single-stage, two-stage, keypoint-based algorithms, and emerging Transformer methods, it is evident that each type of algorithm exhibits unique advantages in specific application scenarios. With the advancement of technology, the field of object detection is facing a series of new research directions and challenges.

Firstly, the extension of object detection into three-dimensional space is gradually becoming a hot topic in research. This trend is driving the demand for 3D detection technologies, which involve the acquisition and processing of depth information, as well as handling the overlap of objects in three-dimensional space. The development of this technology will further enhance the performance of applications in fields such as autonomous driving and robotic navigation. Secondly, the integration of object detection with textual information introduces new multimodal learning strategies. Through these strategies, models can not only recognize objects in images but also understand associated text descriptions, thereby enhancing the model's ability to comprehend complex scenes. This is particularly useful in areas such as image retrieval and automatic annotation. Furthermore, open-set object detection demonstrates a model's generalization capabilities on unseen categories, which is crucial for dynamic environments commonly encountered in real-world applications. Weakly supervised and unsupervised learning methods reduce training costs by decreasing the reliance on large amounts of annotated data, making object detection technology more viable and economical. These methods are especially applicable when data annotation is challenging or too expensive.

In summary, although current object detection technologies have made significant progress, new technological developments and application demands still present numerous challenges. Future research should continue to explore how to integrate different detection technologies and how to optimize algorithm performance under specific scenarios, balancing speed and accuracy to meet the needs of practical applications. Such research will not only drive further technological advancements but also expand the application domains of object detection technologies.

5. Conclusion

This paper has reviewed the main algorithms in the field of object detection, categorizing and comparing them according to their technical characteristics and implementation methods. It also discussed the challenges and future directions faced by the field of object detection. When evaluating various algorithms, the following conclusions were drawn: two-stage algorithms excel in accuracy, while single-stage algorithms have an advantage in speed; keypoint-based methods serve as a complement, striking a relative balance between speed and accuracy; and emerging Transformer-based methods demonstrate powerful capabilities in capturing global information.

In future research, investigators can focus on integrating object detection technology with other techniques to enhance the system's ability to understand complex scenes. By combining the advantageous features of multiple methods, researchers can aim for both lightweight and high-performance development, further enhancing the wide applicability and practical effectiveness of object detection technologies. Additionally, to address the issue of high training costs for many models, weakly supervised and unsupervised learning methods can be introduced to make object detection technologies more economically viable.

References

[1] Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).

[2] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition.

[3] Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision.

DOI: 10.54254/2755-2721/80/2024CH0058

[4] 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.
[5] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision.
[6] Cai, Z., & Vasconcelos, N. (2019). Cascade R-CNN: Delving into High Quality Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
[7] 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.
[8] 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.
[9] Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and Efficient Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
[10] Xu, Y., He, Z., & Tang, X. (2016). CornetNet: A Deep Learning Model for Object Detection in Aerial Images. IEEE Transactions on Geoscience and Remote Sensing, 54(1), 307-320.
[11] Yu, Z., Nair, V., Paluri, M., Han, K., He, K., & Girshick, R. (2019). CenterNet: Geometric Features for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[12] Chen, L.-C., Wu, Y.-L., & Chang, W.-C. (2019). HRNet: High-Resolution Network for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[13] 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.
[14] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv preprint arXiv:2010.04159.
[15] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. European Conference on Computer Vision.
[16] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision.
[17] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition.
[18] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. IEEE Conference on Computer Vision and Pattern Recognition.
[19] Xia, G. S., Bai, X., Ding, J., Zhu, Z., Wang, S., & Belongie, S. (2018). DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. IEEE Conference on Computer Vision and Pattern Recognition.