Deep Learning-Based Object Detection Algorithms

Linxi Yao*

School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, 611130, Chengdu, Sichuan, China

Abstract. One of the main areas of study in computer vision is object detection. It can identify the type and location of target items and determine whether they are present in pictures or movies. With the development of deep learning, Object detection algorithms have seen significant enhancements in both speed and accuracy, leading to extensive adoption across various domains, including autonomous driving, drone surveillance, and security monitoring. This article examines some of the most well-known algorithms from the deep learning period, classifies them into four types of object identification algorithms—two-stage, one-stage, keypoint-based, and transformer-based—and describes their primary advances, benefits, and drawbacks. Furthermore, this work organizes target detection datasets and performance evaluation indicators that are routinely used in studies and provides detailed explanations of their content and properties. The paper adds to the study and advancement of target detection technology-related domains and serves as a useful resource for practitioners and scholars.

1 Introduction

Target identification technology has advanced quickly since AlexNet's 2012 proposal thanks to numerous deep learning experiments [1]. An important task in computer vision is object detection, which searches for, classifies, and determines the type of objects in a picture. Object detection, an essential part of computer vision, is widely used in many fields, such as autonomous driving, car and pedestrian identification, and more. It provides the basis for tasks in computer vision such as instance segmentation and object tracking.

Traditional target detection algorithms typically combine artificially designed target features with machine learning classifiers. These techniques include region creation techniques like Selective Search, sliding window techniques, and techniques that combine feature extraction and classifiers (such Histogram of Oriented Gradients (HOG) with Support Vector Machines (SVM) and AdaBoost). The sliding window approach detects objects by gradually moving a window across the image, however it is computationally inefficient. Feature extraction techniques such as HOG and Scale-Invariant Feature Transform (SIFT) can effectively capture the edge and shape information of the image and can be used with SVM or AdaBoost classifiers for detection, while DPM improves detection accuracy by

* Corresponding author: 42211168@smail.swufe.edu.cn

modeling the deformable parts of the target. Selective Search generates candidate boxes by merging similar image regions and is often used to improve detection results. Although these traditional methods perform well in some applications, they are often limited by computational speed and accuracy when dealing with complex backgrounds, multi-scales, and multi-categories.

As deep learning technology advances quickly, more effective and reliable deep learning target identification techniques are progressively replacing conventional algorithms. Deep learning-based target identification methods greatly increase the precision and effectiveness of their detection by using deep neural networks to directly train feature representations from a vast amount of annotated data. Numerous technical avenues have been investigated by researchers, primarily the Transformer-based approach that has surfaced in recent years, the single-stage method, the two-stage method, and the key point-based way.

Traditional detection systems include single-stage and two-stage target detection techniques. You Only Look Once (YOLO) and Single Shot MultiBox Detector (SSD) are two examples of single-stage detection techniques that are renowned for their quick inference speed. By using a network forward approach, these techniques circumvent the difficult area formation stage and concurrently forecast the category and position of objects. This enables real-time detection performance. The two-stage approach, on the other hand, uses a more intricate mechanism for region construction, first producing candidate areas and then thoroughly classifying and regressing on them. It is therefore usually more accurate than the single-stage method.

In addition, key point-based detection methods bring new ideas to object detection. By precisely identifying the target's important points, these techniques infer the object's bounding box without using the conventional anchor box approach. These methods have shown advantages in scenes such as irregular shapes and estimation but may have high complexity problems when detecting multi-category objects.

Recently, the introduction of the Transformer architecture has also brought a new development direction for object detection. It processes global features through a self-attention mechanism without needing anchor box setting and candidate region generation steps. This method simplifies the model design process while maintaining accuracy, but it also faces the challenges of long training time and high hardware requirements.

This article outlines target detection techniques based on deep learning and categorizes them into four types: single-stage, two-stage, Keypoint-based, and Transformer-based. It also introduces the development, advantages, and disadvantages of the classic algorithms. This page organizes the most extensively used target detection datasets and evaluation indicators to serve as a resource for researchers and practitioners.

2 Algorithms

2.1 Two-stage Algorithms

2.1.1 R-CNN

Girshick proposed the ground-breaking object detection system known as Region-CNN (R-CNN) in 2014 [2]. From the input image, it uses selective search to provide about 2,000 category-independent candidate regions, commonly referred to as Region of Interest (RoI). Before being fed into a Convolutional Neural Network (CNN) that has already been trained to extract features, each RoI is cropped and scaled to a predetermined size. Following dimension scaling, these features are fed into a series of linear SVMs for object classification.

Similarly, to increase positioning accuracy, R-CNN refines the projected bounding box using a linear regression model.

Although R-CNN achieved high accuracy, it required running CNN for each RoI separately, which was computationally intensive and slow. In addition, the model training process was complex, requiring independent training of CNN, SVM, and bounding box regressors. These problems led researchers to develop more efficient algorithm variants.

2.1.2 SPP-Net

In 2014, He proposed the SPP-Net method [3]. Its primary goal is to resolve R-CNN's efficiency issue when handling inputs of various sizes. By adding a Spatial Pyramid Pooling (SPP) layer to the convolutional network's last layer, constant-size feature extraction is accomplished. The SPP layer uses pooling algorithms of various scales to divide the image into a fixed number of grids. After that, each grid is combined to create a feature vector with a preset length. This eliminates the need to pre-crop input photographs to a predetermined size, allowing the network as a whole to accept images of varying sizes.

This design of SPP-Net reduces repeated calculations of candidate areas and improves calculation efficiency. Simultaneously, the saved features can be applied repeatedly to various ROIs because feature extraction is only done once on the full image, greatly speeding up detection. However, SPP-Net lacks end-to-end training, which lowers accuracy.

2.1.3 Fast R-CNN and Faster R-CNN

Fast R-CNN, which Girshick introduced in 2015, performs better than R-CNN and SPP-Net [4]. Compared with R-CNN, this algorithm integrates the target detection steps into a single, end-to-end trainable framework, greatly reducing repeated calculations. In addition, it incorporates a candidate box mapping function and enhances the more condensed zone of interest pooling layer suggested by SPP. Furthermore, Fast R-CNN enhances detection speed and accuracy by using a multi-task loss function for joint training.

A more advanced version of Fast R-CNN is known as Faster R-CNN [5]. Its primary innovation is the use of the Region Proposal Network (RPN), which simplifies the entire detection procedure. An external selective search technique is used in Fast R-CNN to produce candidate regions, which is a laborious procedure. To solve this problem, RPN is applied in Faster R-CNN, which generates high-quality candidate regions directly in the network by sharing convolutional feature maps, greatly improving the speed.

Faster R-CNN maintains or even increases accuracy in addition to increasing detection speed when compared to Fast R-CNN. From a phased approach to an end-to-end solution, CNNs have revolutionized object detection, greatly streamlining the procedure and increasing object detection efficiency.

2.1.4 R-FCN

In 2016, Dai made the initial introduction of Region-based Fully Convolutional Networks (R-FCN) [6]. Its core innovation is to design the entire detection architecture as a fully convolutional network to make it more efficient. Position-sensitive score maps are used in R-FCN in place of Faster R-CNN's fully linked layers. R-FCN utilizes a number of convolutional layers to produce numerous position-sensitive maps after RPN generates the feature map to produce candidate regions. These maps are position-sensitively averaged and pooled in the ROI area to achieve efficient classification and positioning. Residual networks and other complicated networks benefit greatly from this design's improved detection speed and decreased calculation. Compared with Faster R-CNN, R-FCN achieves similar or even

higher detection accuracy with fewer parameters, reflecting the flexibility and efficiency of CNNs in target detection.

2.2 One-stage Algorithms

2.2.1 You Only Look Once

A detection method based on one stage of regression is called YOLO [7]. It uses a neural network to categorize and forecast the recognized items using bounding boxes, transforming the detection problem into a deep learning regression problem.

Quick and efficient detection is made possible by YOLO's treatment of object detection as a single regression analysis problem. This technique predicts bounding boxes and the associated class probabilities from the input image using a single CNN. Unlike traditional approaches, it is trained on the entire image and interprets the global information of the entire image instead of only the local part. This is one of its features. The YOLO method has the advantage of being rapid to process, making it suited for application scenarios that require real-time response. In addition, because the information of the entire image is considered in training, YOLO shows strong generalization ability when dealing with unfamiliar scenes.

With the development of technology, YOLO has gone through multiple versions of updates. YOLOv1 greatly simplified the detection process by dividing the image into grids and predicting bounding boxes and categories for each grid; nonetheless, it did not perform well in situations with small and dense objects. YOLOv2 introduced anchor boxes and batch normalization to improve positioning accuracy and training stability. Because multi-scale training improves the network's capacity to adapt to diverse input sizes, YOLOv3 employs Darknet-53 as a feature extractor and incorporates multi-scale prediction technology, which significantly improves accuracy and detection capabilities for small objects. YOLOv4 further improves speed and accuracy by integrating new technologies such as the Mish activation function, CSPNet, and CIoU loss. YOLOv5 and subsequent versions (YOLOv6, YOLOv7, YOLOv8, YOLOv9, YOLOv10, YOLOv11) continue to optimize lightweight design and hardware adaptation. The YOLO algorithm is able to better adapt to the demands of various scenarios and hardware platforms thanks to the introduction of anchor-free detection, improved network architecture, new loss functions, efficient feature extraction technology, and more effective training strategies. These improvements also increase the model's generalization ability and flexibility.

2.2.2 Single Shot MultiBox Detector

The SSD aims for high detection speed by using single-stage detection to predict item bounding boxes and class labels directly from input photographs [8]. SSD is based on the VGG16 network and extracts multi-scale information via multiple convolutional layers [9]. These feature maps assist SSD in detecting and recognizing objects of various sizes at various scales. The use of a default box system, which generates many default boxes with different scales and aspect ratios for each feature map point, is a noteworthy enhancement. To improve detection accuracy and variety, these default boxes are coupled with ground-truth bounding boxes.

Position regression loss is used to accurately change the bounding box's position, whereas classification loss is used to correctly determine the target object's category. This allows SSDs to maintain high detection accuracy while remaining fast. Through a reasonable hierarchical structure, SSD can effectively handle multi-target detection tasks in different scenarios.

Compared with the two-stage detection method, SSD eliminates the complex region proposal step. As a result, it offers a significant speed advantage and works well for applications like video surveillance and autonomous driving that demand high real-time performance. However, SSD's performance might be marginally worse than some two-stage approaches when it comes to detecting small or extremely dense objects. Despite this, SSD's basic and efficient architecture remains a vital foundation and motivation for the future development of target detection technology.

2.3 Keypoint-based Algorithms

2.3.1 CornerNet

In contrast to the conventional bounding box approach, CornerNet is a novel object detection technique that locates an object by detecting its upper corners [10]. The algorithm is divided into two parts: identifying and appropriately pairing the object's top left and bottom right corners.

CornerNet predicts the locations of every corner at once using a single CNN. The network produces a range of feature maps, such as feature embedding maps to help the system pair the upper left and lower right corners to build a complete bounding box and heat maps to recommend possible corner placements. The heat map sampling also uses a focal loss function to train corner recognition, which enhances the resilience of objects of different scales.

This method avoids the region proposal step, making the detection process more simplified and efficient. CornerNet is designed to capture more complex object shapes and provide higher detection accuracy. However, because CornerNet only looks at edges and corners, the corners of the bounding box may be outside the semantic information, resulting in false positive results.

2.3.2 CenterNet

The CenterNet algorithm improves the problem of CornerNet’s lack of attention to the global nature of objects [11]. It infers the size and category of the object by explicitly predicting its center point on the feature map.

Using a CNN, the technique creates a feature map that includes an offset, size, and heat map. The heat map forecasts the location of the object's center, the size map provides the object's width and height, and the offset map helps more accurately regress the exact coordinates of the center point. By including key point detection, CenterNet enables the model to concurrently determine the category and placement of several objects.

The advantage of CenterNet lies in its end-to-end simplified structure, which abandons complex steps such as RPN and non-maximum suppression, thereby improving detection efficiency and speed. Furthermore, CenterNet can adjust to target identification tasks of different sizes and attain greater accuracy with fewer parameters.

2.4 Detection Transformer

Detection Transformer (DERT) is a novel object identification algorithm proposed by Facebook's AI research team that combines the benefits of CNN and Transformer models [12]. Conventional object detection techniques often use non-maximum suppression and region proposal. DETR simplifies the detection process by converting the detection problem into a sequence-matching task.

Initially, DETR uses a CNN (such as ResNet) to extract visual information. The Transformer encoder then receives these feature maps and uses a multi-layer self-attention mechanism to record global information. During the decoding phase, the decoder is made up of N learnable query vectors that use an attention mechanism to interact with the encoder's output features and forecast the object's kind and location. The decoder's output immediately generates each object's category and bounding box coordinates after passing through a linear layer.

DETR does not require non-maximum suppression to handle redundant outputs, thus simplifying post-processing. The Hungarian matching approach is used to match the projected outputs with the genuine labels, and the goal is to train the complete network end-to-end by combining the bounding box regression loss and classification loss.

DETR's innovation is to eliminate redundant steps and adopt the Transformer architecture to achieve global information modeling, which enhances the detection ability of complex scenes. However, its computational complexity is high, especially when processing high-resolution images, which requires more computing resources.

3 Experimental Datasets and Evaluation Indicators

3.1 Datasets

The PASCAL VOC dataset is primarily used to assess tasks related to semantic segmentation, object detection, and image classification. Its advantages are clear evaluation indicators and standardized image annotations. It contains images of various real-world scenes, which are annotated in detail and cover multiple categories. VOC2007 and VOC2012 are the two mainstream versions.

The Microsoft COCO dataset contains rich and diverse images of everyday life, emphasizing the context of objects. The COCO dataset has detailed annotations, including bounding boxes for target detection, category labels for objects, semantic segmentation, and fine-grained instance segmentation. Covering up to 80 categories of objects, the dataset is known for its challenges with complex backgrounds, high-density targets, and small-sized objects. One of the fundamental datasets in computer vision, COCO is used extensively to assess and advance the effectiveness of deep learning models in picture understanding because of its variety and difficulties.

Google released the Open Images dataset, a large collection of images that may be applied to various computer vision tasks. With over 9 million images in thousands of categories, each image has extensive annotations, including bounding boxes for object detection, image-level labels, instance segmentation, multi-label tasks, and visual relationship detection.

Compared with other datasets, Open Images has a larger scale and richer annotation information, highlighting the multi-object relationships in complex scenes.

Table 1 shows general datasets and specialized information in the subject of target detection.

Table 1. General datasets.

Datasets	Publication Year	Number of Images	Features
VOC2007	2007	9963	There are large changes in size, direction, posture, lighting, position, occlusion, etc.
VOC2012	2012	11540	
MS COCO	2014	328000	There are many small objects, many objects in a single image, and most categories of objects have many instances.
Caltech 101	2006	9146	Most images contain only one object and have little image variation.
Caltech 256	2007	30607	It provides online annotation tools, more open.
LabelMe	2008	187240	
ImageNet	2009	14M	The dataset is large in scale and has many image categories, which is very challenging.
SUN	2010	131072	This is a multimodal dataset that combines multiple modal information such as region description, relationship, question-answer pairs, etc.
YFCC100M	2014	99.2M images + 800K videos	A multimodal dataset containing images and videos, with a very large dataset size.
VQAv1	2015	254721	An image dataset based on MS COCO that focuses on visual question answering.
VQAv2	2017	286046	
Visual Genome	2016	108249	Multimodal dataset, integrating multiple modal information such as region description, relationship, question-answer pair, etc.
Open ImagesV4	2018	1.9M	In manual annotation, images often show complex scenes with multiple objects.

3.2 Evaluation Indicators

Precision indicates the proportion of test frames (TP) predicted as positive and correct among all test frames (TP+FP) predicted as positive. Precision shows how well the model can anticipate the right object. All of the model's projected positive cases are incorrect when the precision is 0. All of the positive cases that the model predicted are accurate when the precision is 1. A greater precision means that only a tiny percentage of non-target objects are recognized as targets, and the majority of all positive examples that the model predicts are accurate targets.

Recall can be defined as the proportion of objects (TP) that the model anticipated to be positive instances out of all positive examples (TP+FN). The model's recall is its capacity to identify every correct object. When the recall is 0, the model does not find any correct positive examples, and when the recall is 1, all correct positive examples are found. The more accurate positive instances the model can identify, the higher the recall.

The F1-score represents the harmonic mean of the precision and recall markers. It creates a more stable model that is impartial by balancing the Precision and Recall metrics.

Average Precision (AP), or the average of the accuracy at different recall points, is represented by the area under the PR curve on the PR curve. The model's average accuracy increases with the AP number.

Since a model in object detection typically detects a wide variety of items, an AP value can be computed once a PR curve has been built for each category. AP values are averaged across multiple categories to determine mean average precision (mAP). Table 2 displays the calculations and justifications for pertinent evaluation indicators.

Table 2. Evaluation Indicators.

Keywords	Formula	Explanation
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$	a measurement of the percentage of all samples that have the right classification.
Precision	$\frac{TP}{TP + FP}$	the proportion of correctly predicted positive samples to all projected positive samples.
Recall	$\frac{TP}{TP + FN}$	the ratio of all true positive cases to the samples that were accurately predicted to be positive examples.
F1-Score	$\frac{2Precision * Recall}{Precision + Recall}$	The harmonic mean of the two variables is used to analyze the precision and recall of the model in detail.
AP	$\int_0^1 precision(recall)d(recall)$	Measures the change in precision over recall. A high AP indicates good model performance.
mAP	$\frac{1}{N}\sum_{i=1}^{N} APi$， N is the number of categories.	The target detection task's overall performance is assessed using the average AP of several categories.
IOU	$\frac{Intersection\ Area}{Union\ Area}$	By calculating the overlap between the expected and actual boxes, it is commonly used to evaluate the accuracy of item detection. The detection result is often regarded as accurate when IOU $\geqslant$ 0.5.

4 Conclusion

Several deep learning-based target detection methods are presented in this article and are separated into four categories: transformer-based target detection algorithms, key point-based target detection algorithms, two-stage target identification algorithms, and single-stage target detection algorithms. This makes it possible to talk about the evolution, breakthroughs, benefits, and drawbacks of many traditional algorithms. The article also organizes the general target detection data set and various evaluation indicators and briefly introduces them. A basic method in computer vision, target identification has been significant in many fields. Even though a lot of the algorithms in use today can handle tasks that are moderately challenging, it is still important to think about how to improve detection efficiency and accuracy. In the future, target detection technology needs to be combined with the development of emerging technologies to continuously improve its performance to cope with more complex scenarios.

References

1. A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks. Commun. ACM. **60**, 84-90 (2017)
2. R. Girshick, J. Donahue, T. Darrell, et al., Rich feature hierarchies for accurate object detection and semantic segmentation, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, (2014), 580-587

3. K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 1904-1916 (2015)
4. R. Girshick, Fast R-CNN, in Proceedings of the 2015 IEEE International Conference on Computer Vision, IEEE, (2015), 1440-1448
5. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in Proceedings of the IEEE International Conference on Computer Vision, IEEE, (2015), 91-99
6. J. Dai, Y. Li, K. He, et al., R-FCN: Object detection via region-based fully convolutional networks. Adv. Neural Inf. Process. Syst. **29,** (2016)
7. J. Redmon, S. Divakaran, R. Girshick, A. Farhadi, You Only Look Once: Unified, real-time object detection, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, (2016), 779-788
8. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: Single Shot MultiBox Detector, in Proceedings of the European Conference on Computer Vision, Springer, (2016), 21-37
9. K. Simonyan, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. H. Law, J. Deng, CornerNet: Detecting objects as paired keypoints, in Proceedings of the European Conference on Computer Vision, Springer, (2018), 740-755
11. K. Duan, S. Bai, L. Xie, et al., CenterNet: Keypoint triplets for object detection, in Proceedings of the IEEE/CVF International Conference on Computer Vision, (2019), 6569-6578
12. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in Proceedings of the European Conference on Computer Vision, Springer, (2020), 213-229