

# The Evolution of Object Detection Algorithms Based on Deep Learning

Qiyan Guo <sup>a</sup>

Glasgow College, UESTC, Chengdu, Sichuan, 611731, China

**Keywords:** Object Detection, Deep Learning, Contrastive Analysis.

**Abstract:** Object detection technology is an important research direction in the field of computer vision, which is widely used in automatic driving, security monitoring, medical image analysis and other fields. With the rapid development of deep learning technology, object detection algorithms based on deep learning have made significant breakthroughs in accuracy and efficiency, gradually replacing traditional object detection methods. The object detection algorithm based on deep learning can automatically learn features in images and train and optimize them in an end-to-end manner, significantly improving detection accuracy and robustness. The progress of object detection technology is initially reviewed in this study, with particular attention paid to deep learning-based object identification algorithms and conventional object detection techniques. Next, two common deep learning object identification frameworks—YOLO-v3 and Faster R-CNN—are thoroughly examined and contrasted. Experimental results show that YOLO-v3 has a slightly higher average accuracy (mAP) on the COCO dataset than Faster R-CNN, but performs better in small target detection and dense scenarios. Nevertheless, the Faster R-CNN still has some advantages in terms of overall accuracy.

## 1 INTRODUCTION


As one of the core tasks in the field of computer vision, object detection aims to identify and locate specific categories of objects from images or videos, and its application scope covers many fields such as automatic driving, security monitoring, medical image analysis and so on. Deep learning-based object recognition algorithms have progressively supplanted conventional techniques and gained popularity in recent years due to the quick advancement of deep learning technology. Traditional object detection methods, such as HOG and Haar, rely on hand-designed feature extraction methods and classical classifiers. Although the computational resource requirements are low and the speed is fast, the detection accuracy and robustness in complex scenes are limited.

Deep learning, and convolutional neural networks (CNN) in particular, has significantly accelerated the advancement of object detection technologies. The deep learning-based object identification algorithm can automatically learn the feature representation from a vast amount of labeled data, greatly increasing

the efficiency and accuracy of detection. Two-stage and single-stage algorithms are the two primary groups into which these algorithms fall based on the various detection frameworks. Two-stage detection frameworks, such as Faster R-CNN, which first generates candidate regions, then performs classification and bounding box regression, have higher detection accuracy, but are slower. Single-stage detection frameworks, such as the YOLO family, predict target categories and bounding boxes directly on the feature map, faster but with relatively low accuracy.

Despite significant progress in object detection algorithms based on deep learning, many challenges remain. For example, in extreme weather and light conditions, the performance of the algorithm may be affected; For some specific types of small targets, the detection effect still needs to be improved. How to balance detection speed and accuracy, and how to deal with scale change and target occlusion, still need to be further studied.

The evolution of object detection technology will be reviewed in this paper, along with the analysis of deep learning-based object detection algorithms and

<sup>a</sup> <https://orcid.org/0009-0005-5972-6222>

conventional object detection techniques. Two common deep learning object detection frameworks, Faster R-CNN and YOLO-v3, will be thoroughly discussed and compared. Finally, the challenges and future development direction of target detection technology will be discussed in order to provide reference and inspiration for related research.

## 2 TRADITIONAL OBJECT DETECTION TECHNOLOGY

In contrast to deep learning model-based object detection frameworks, traditional object detection technologies rely more on hand-designed feature extraction methods and classical classifiers (Guo, 2024). These techniques are more effective at capturing information in an image, but are limited to simple image features such as color, texture, and edges to identify the category and location of the object. These manual features, combined with powerful machine learning classifiers, make it possible to distinguish target and non-target regions from feature vectors. According to the method of extracting image features from the model, traditional object detection algorithms can be divided into the following categories: sliding method, region-based method, and model-fitting-based method.

The Histogram of Oriented Gradients (HOG) object detection algorithm, which was put forth by Navneet Dalal and Bill Triggs in 2005, is the most representative of these techniques. This algorithm's core idea is to create a histogram based on gradient characteristics, vote on the image's local gradient amplitude and direction, and then concatenate the local features to create the overall features (Dalal and Triggs, 2005). It is robust to changes such as lighting and local occlusion, and can capture shape and texture information in images well, so it performs well in tasks such as pedestrian detection. However, it is very sensitive to the rotation and scale change of the target, and because of its high computational complexity, the real-time detection of the target is poor.

Haar target detection algorithm is also a very representative traditional target detection algorithm. It is a classical object detection algorithm based on Haar features and a cascade classifier. It represents the texture and edge information of the image by calculating the gray difference of pixels in adjacent rectangular areas in the image. Then, the AdaBoost algorithm is used to select the best subset of Haar features, and a cascade classifier is constructed, which slides on the image with a fixed size window.

The cascade classifier is used to classify different regions of the image. This algorithm has certain robustness to illumination changes and target attitude, and the detection speed is fast, which is suitable for real-time applications, so it is widely used in face detection, pedestrian detection and other fields (Viola and Jones, 2011). However, it is sensitive to the scale and attitude changes of the target, so the application scenario is relatively limited.

These traditional target detection methods have low requirements on computing resources and relatively fast computing speed, so they still have application value in scenarios with limited conditions or high real-time requirements. However, they require a lot of manual adjustment and optimization in complex or changeable environments. With the progress of technology, traditional target detection algorithms are gradually being eliminated.

## 3 TYPICAL CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE

In computer vision, object detection is a crucial activity that seeks to locate or identify a particular class of objects in a picture or video. By directly learning feature representations from a large number of tags using deep learning neural networks, object detection algorithms based on deep learning have increased the accuracy and efficiency of object detection in recent years. The advancement of object detection technology has benefited considerably from the use of convolutional neural networks (CNN). These algorithms can be divided into two broad categories based on how the detection framework operates and recognizes them. The first type is a two-stage detection framework, which identifies potential target areas, and then performs classification and bounding box regression. The second type is a single-stage detection framework, which recognizes both the image category and the bounding box in one operation (Liu, Ouyang, and Wang, 2020) (Liu, Ouyang, and Wang, 2018).

### 3.1 Faster R-CNN

Faster R-CNN is a very representative two-stage target detection algorithm that represents a substantial improvement in the field of target identification. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Su proposed the third-generation model in the R-CNN series in 2015 (Ren, He, and Girshick, 2017),

This system is composed of a classifier, a region proposal network (rpn), a convolutional layer, and a roi pooling layer. The convolution and pooling layers are utilized to extract the features of the image. The pooling layer uses the size of the image to extract features, and the convolution layer uses the translation invariance of the image. In comparison to earlier extraction networks, the Regional Proposal Network (RPN) significantly increases the speed of regional proposal by using a full convolutional network that can immediately produce candidate regions from feature maps. Its main concept is to separate the object identification work into two phases: the extraction of candidate regions and the categorization of candidate regions. Firstly, selective search is used to select some candidate areas in the input image that may contain the target object. These candidate regions are rectangular regions that are used for subsequent feature extraction and classification. Then feature extraction is carried out for each candidate region, and a classifier is used to classify them to determine whether they contain the target object and the category of the target object. Because RPN replaces the traditional Selective Search in the Faster R-CNN, the time of regional proposal is greatly shortened, making the operation of regional proposal more efficient. At the same time, Faster R-CNN combines RPN and Fast R-CNN into one network. In actual training, the two can share convolution features, which realizes end-to-end training and reasoning, simplifies the training process, and shortens the training time. However, although the speed of regional proposal is accelerated, its detection speed is still slow in real life and can not meet the requirements of real-time detection. At the same time, because it integrates RPN and Fast R-CNN modules, the number of parameters and calculation amount of the model are large. At present, A wide range of target detection tasks, including lesion analysis identification in medical pictures and defect detection in industrial production, also make extensive use of faster R-CNN.

### 3.2 YOLO - v3

The third iteration of the YOLO series, known as Yolo-v3, was put forth by Joseph Redmon and Ali Farhadi in 2018. It keeps the YOLO series' simplicity and effectiveness while enhancing target recognition speed and accuracy. The YOLO series is a common single-stage object detection framework as its main principle is to turn the object identification problem into a regression problem. Its primary capabilities include the ability to recognize targets of varying

sizes using multi-scale feature maps and to anticipate targets directly on the feature map without region proposal (Redmon, Divvala, Girshick, 2015). Darknet-53, an effective convolutional neural network that integrates multi-scale feature fusion and residual linking for enhanced feature extraction capabilities, forms the foundation of YOLO-v3. At the same time, it will predict at 3 different scales, thus improving the detection ability of small targets and making full use of shallow and deep feature information. After that, K-means clustering is used to generate 9 Anchor Boxes for predicting target boundary boxes. Then, the boundary Box coordinates, object confidence and category probability of each Anchor Box are predicted on the feature map of each scale (Redmon, Farhadi, 2018). As a typical single-stage object detection framework, YOLO-v3 has a relatively fast detection speed and can achieve real-time detection. However, compared with the two-stage detection framework, its detection accuracy will be significantly lower due to the defects of the detection steps. At the same time, its multi-scale prediction ability also improves the performance of small target detection, but its performance is still not satisfactory in the scenario of small target density. At present, YOLO-v3 is widely used in a variety of real-time target detection tasks, such as automatic driving, security monitoring and UAV target detection systems have YOLO-v3 framework applications.

### 3.3 Comparison and analysis of the two frameworks

Table 1: Coco Dataset Validation mAP of two Frameworks.

Framework	Faster R-CNN	YOLO-v3
mAP	0.42	0.45

mAP of the two frameworks verified using the coco dataset is as shown in table 1, that is, the average test accuracy, which represents the average value of the detection accuracy under different IoU thresholds (Neha, Bhati, Shukla, 2024). According to the experimental results, YOLOv3 achieves the best performance on the COCO validation set, which may be because it adopts more advanced network structure and optimization methods, which makes the model perform better in feature extraction and classification. In this test, the YOLO-v3 was slightly more accurate than the Faster R-CNN, but there are several versions of YOLO-v3, such as the YOLO-v3 tiny, which is typical of models that ensure sufficiently high test speeds at the expense of some accuracy. So overall, the accuracy of Faster R-CNN is still slightly higher.

Furthermore, the RPN, or regional proposal network, is not very sensitive to the identification of small and dense objects because the candidate region it generates is typically huge. In contrast, YOLOv3 works much better. At the same time, the two stages of the Faster R-CNN network require a large amount of data, so the training of Faster R-CNN is relatively difficult.

## 4 CHALLENGES AND FUTURE DEVELOPMENT

The performance of the algorithm may be affected when processing images under extreme weather and illumination conditions, and the detection effect of some specific types of small targets still needs to be improved. Further research can be carried out to improve and optimize the above limiting factors to further improve the practicality and robustness of the big data target detection algorithm. When working with a high number of candidate regions, the R-CNN series technique requires a lot of processing, which leads to inadequate real-time performance (Zou, Chen, Shi, 2023). The YOLO algorithm's accuracy has to be increased while working with small, dense targets. Therefore, how to balance the detection speed and accuracy is still a problem to be solved. In addition, how to deal with scale change and target occlusion still to be further studied.

From conventional manual feature approaches to deep learning-based CNN and Transformer architectures, object identification technology has advanced tremendously in the field of computer vision. Accuracy and real-time detection have also increased significantly. Nevertheless, challenges such as small target detection, domain adaptation, and adversarial attacks remain. In the future, multi-modal fusion, self-supervised learning and open-world object detection will provide new impetus for the development of object detection technology.

## 5 CONCLUSIONS

This study compares the performance of two well-known deep learning object detection frameworks, Faster R-CNN and YOLO-v3, by examining the evolution of object recognition technology from the manual, feature-based approach to the deep learning-based convolutional neural network (CNN) architecture. To assist readers in understanding the advancement of technology in this area, a detailed

summary of the object detection technology's evolution from the old method to the deep learning method is provided. By reviewing the traditional methods such as HOG and Haar, we reveal their advantages in computational efficiency and real-time performance, and point out their limitations in complex scenarios. Next, two common deep learning object identification frameworks – YOLO-v3 and Faster R-CNN – are thoroughly examined and contrasted. The experimental results show that YOLO-v3 has a slightly higher average accuracy (mAP) on the COCO dataset than Faster R-CNN, and performs better in small target detection and dense scenes, while Faster R-CNN still has some advantages in overall accuracy. This comparative analysis provides a valuable reference for researchers and helps to choose a suitable detection framework for practical applications. This paper not only summarizes the current status of object detection technology, but also points out the future research direction and challenges. For example, how to improve the robustness of the algorithm under extreme weather and lighting conditions, how to optimize the detection accuracy of small target detection and dense scenes, and how to balance the detection speed and accuracy. These research directions provide new ideas and impetus for the further development of object detection technology. In conclusion, this paper not only provides valuable reference and inspiration for researchers in this field, but also provides an important basis for technology selection and optimization in practical application. The findings of this study will support the advancement of target detection technology and enhance its viability and resilience across a range of application domains.

## REFERENCES

- Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, pp. 886-893 vol. 1
- GUO, H., 2024. Object detection: From traditional methods to deep learning. *Emerging Science and Technology*, 3(2), pp.128-145.
- Liu, L., Ouyang, W., Wang, X. et al., 2020. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision*, 128, pp.261–318. <https://doi.org/10.1007/s11263-019-01247-4>
- Liu, L., Ouyang, W., Wang, X., Fieguth, P.W., Chen, J., Liu, X., & Pietikäinen, M., 2018. Deep Learning for

Generic Object Detection: A Survey. *International Journal of Computer Vision*, 128, pp.261-318.

- Neha, F., Bhati, D., Shukla, D.K., & Amiruzzaman, M., 2024. From classical techniques to convolution-based models: A review of object detection algorithms. *ArXiv*
- Redmon, J., Divvala, S.K., Girshick, R.B., & Farhadi, A., 2015. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.779-788.
- Redmon, J., & Farhadi, A., 2018. YOLOv3: An Incremental Improvement.
- Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), pp.1137-1149. doi: 10.1109/TPAMI.2016.2577031.
- Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*
- Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J., 2023. Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, 111(3), pp.257-276.

