Progress on Deep Learning-Driven Target Detection Techniques

Dinghuizhi Zhang

School of Computer Science & Technology, Xi'an University of Posts & Telecommunications, Xi'an, 710121, China
zdhz@stu.xupt.edu.cn

Abstract. Deep learning (DL)-driven target detection techniques are one of the key topics in today's research. The target detection algorithms have been updated along with researchers' continuous updating of target detection model libraries and datasets in recent years. However, there is a lack of knowledge about the technical route and key performance of target detection techniques driven by DL. Therefore, this paper explores the research progress of DL-driven target detection techniques by collecting and analyzing the iterative algorithms of DL-driven target detection techniques for the past few years. This paper reviews the latest research advances of this technology to discuss its strengths and weaknesses, as well as a prediction of its challenges and possible future directions. It has been found that the two-stage target detection algorithm is much more accurate in processing images, yet it is too slow and is suitable for use in situations where a high standard of accuracy in image processing is required. The single-stage target detection algorithm is more suitable for applications requiring large-scale pipelined processing. This paper can provide technical guidance or theoretical support for DL applications in areas such as manufacturing.

Keywords: Deep learning, target detection, artificial intelligence, machine vision.

1 Introduction

Recently, the wide application of image data has driven the rapid growth of machine vision technology. Target detection, as a key problem in machine vision, is tasked with finding all the targets in an image and classifying, localizing, detecting, and segmenting them. However, target detection is one of the most difficult problems in computer vision owing to the different appearances, shapes, and postures of diverse types of objects, and the interference of various external factors during imaging. Target detection has undergone a radical revolution from the original methods based on traditional machine learning to today's DL techniques. Early-stage researchers have been relying on hand-designed features and traditional classifiers for target detection, dating back as far as 2001 when the proposed Viola-Jones detector was a weak classifier based on simple features [1]. It was not until the widespread use of convolutional neural networks (CNNs) that the performance of target detection took a qualitative leap

J. Iqbal (ed.), *Proceedings of the 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025)*, Advances in Engineering Research 272,
https://doi.org/10.2991/978-94-6463-821-9_69

forward. Compared to traditional methods, DL has a stronger ability to automatically learn more discriminative deep features. However, looking at the wide variety of algorithmic techniques, it is also a major problem when one should choose the right algorithm based on a particular domain.

The current main technical routes of DL for target detection are based on Convolutional Neural Network (CNN), You Only Look Once (YOLO), Single Shot MutiBox Detector (SSD), Retina Network (RetinaNet), etc., which can realize static target, moving target detection, multi-scale and contextual feature target detection. However, there is a huge difference in the effectiveness of the detection based on different routes, which is due to the difference between single- and two-stage framing principles, feature pyramids, and anchor frame mechanisms. However, the efficiency and accuracy of DL target detection are limited by the fact that the specific principles and mechanisms are not yet clear. Therefore, according to the analysis of the construction principle of the DL target detection model, this paper elucidates how the appropriate DL target detection route should be selected according to the application scenario.

This paper intends to provide a comprehensive and systematic analysis of DL-driven target detection techniques, to supply theoretical guidance and technical support for the further development of this field, and to provide an outlook on its future.

2 Overview of Target Detection Techniques

Since the advancement of target detection technology, from the early simple algorithms to today's DL algorithms, this field has experienced rapid development, and its application scope has been extended from simple image classification to complex three-dimensional spatial recognition. Overall target detection techniques are broadly categorized into traditional target detection techniques and target detection techniques driven by DL.

2.1 Traditional Target Detection Methods

Traditional target detection algorithms could be separated into three steps: candidate region selection, feature extraction, and classifier. Region selection mainly uses an exhaustive strategy with a sliding window and different sizes and different aspect ratios are set to traverse the image with high time complexity. Feature extraction techniques mainly refer to HOG, SIFT, etc. In the case of HOG, for example, HOG refers to the histogram of the gradient of the direction of the features that can be used to extract a representation of the image. In 2017, Omid-Zohoor et al. proposed and developed a simulation framework based on a customized version of the HOG function with a 2-bit pixel ratio, which has the potential to reduce the I/O energy by a factor of 19 compared to the traditional target detection methods, demonstrating superior object detection performance [2]. However, this approach has poor feature robustness due to morphological diversity, light variation diversity, and background diversity. Classical machine learning methods (e.g., SVMs) play a huge role with small training data [3].

Zhang et al. have then introduced streaming regularized SVM models by integrating local regularizer and global regularizer into SVM models for self-training [4]. This improves the detection performance and significantly improves the accuracy of the pseudo-training samples. The method outperforms 13 cutting-edge methods on five reference datasets, successfully demonstrating its excellent detection performance.

2.2 Evolution of DL Target Detection

Although traditional target detection is usually simple and computationally efficient, the problems of limited accuracy and poor generalization ability have greatly restricted its development. Because of this, target detection techniques based on DL have naturally emerged.

The concept of DL is introduced as a multiple-layer ANN model with significant learning capabilities, as shown in Fig. 1 [5]. DL enables a computational model with multiple processing layers to acquire and represent data with multiple levels of abstraction. It also can capture the complex structure of large-scale data by imitating how the brain perceives and comprehends multimodal information [6]. DL models have made great strides in addressing challenges such as anomaly detection, disease diagnosis, and capturing information from big data.

DL algorithms focus on learning hierarchical feature representations of data by building deep neural networks, which allows them to tackle more complex problems. However, to train the model efficiently, a substantial amount of data is essential, as complex network structures require more data to avoid overfitting.

Before the advent of CNNs, images were a problem for AI, on the one hand, the amount of data that images needed to process was too large, resulting in low efficiency; On the other hand, it is difficult to preserve the features of the image during the digitization process, resulting in less accurate image processing. As the cornerstone of two-stage detection, the R-CNN series incorporates CNN technology into target detection for the first time, improving the accuracy of image processing; Immediately after the introduction of YOLO and SSD signaled the rise of single-stage detection, RetinaNet took over to achieve a breakthrough in accuracy.

DL-driven target detection techniques have significant advantages over traditional methods, and these advantages are not only reflected in performance improvement but also drive a paradigm change in the field of computer vision, enabling automatic feature learning and data-driven optimization based on guaranteed hierarchical feature extraction; Secondly, end-to-end training has been realized, breaking the fragmentation of the various testing phases.

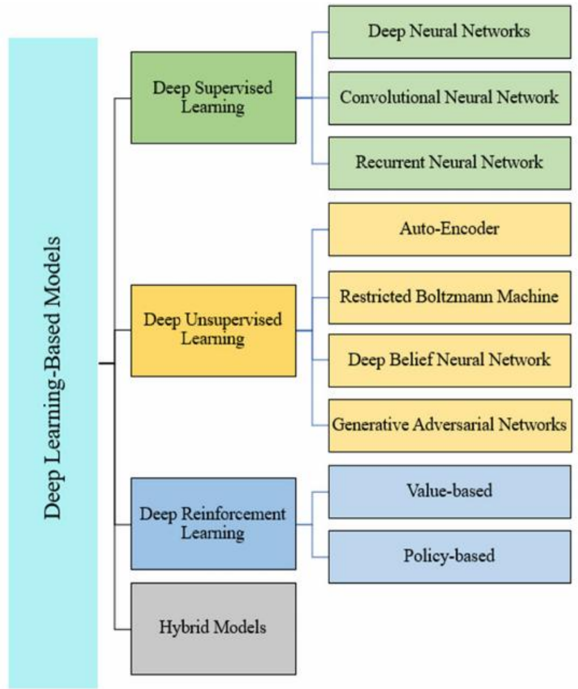

Fig. 1. Schematic review in DL modeling [5].

3 DL Frameworks and Models

With the development of AI, the algorithmic techniques for DL-driven target detection are constantly iterating.

3.1 CNN

Convolutional neural networks have an irreplaceable and seminal role in target detection. Much of the early success of DL stemmed from a class of feedforward pre-attentive neural networks called CNNs. Their design is influenced by the biological visual system and aims to imitate how human vision is processed. CNNs include three essential parts, they are respectively convolutional, pooling, and fully connected layers. Each layer category serves a distinct function. The CNN uses various kernels to convolve the images to generate various feature maps in the convolutional layer. The pooling layer reduces the data dimension to avoid overfitting. Only data processed by the first two layers can be output. The CNN architecture for object detection in image tasks is shown in Fig. 2 [6]. In recent years, CNNs have made great progress in the field of image recognition, target detection, image generation, and many other areas. It has become an important part of computer vision and DL research.

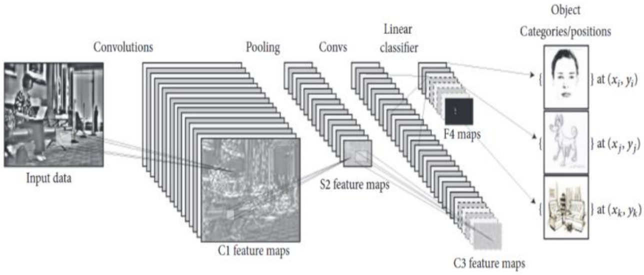

Fig. 2. Architecture of a CNN for an object detection task [6].

R-CNN series. The R-CNN family, proposed in 2014 by Girshick et al., serves as the groundbreaking work of classical target detection algorithms, which significantly improves detection performance and drives the shift from traditional handcrafted features to DL methods [7].

The Fast R-CNN technique proposed in 2015, which is a multi-task learning detector, this technique uses several innovations to increase the speed of testing as well as to improve the accuracy of testing. Redundant computation is reduced by adding a Rol Pooling layer, and the original classification and bounding box regression steps are combined into a single loss function but suffer from a lack of real-time performance and poor quantization time. Wang et al. in 2019 had attempted to acquire an object detector as accurate as fully supervised fast R-CNN. Based on the object detector with less image annotation work, the object detector uses weakly supervised and semi-supervised information to learn in the Fast R-CNN framework, and finally, the new learning framework remarkably reduces the labeling work to obtain a reliable object detector [8].

Ren et al. developed a faster R-CNN target detection method using a deep CNN [9]. It performs region recommendation via the RPN network to fully utilize Graphics Processing Units (GPUs) with high computational complexity, but it is still an excellent target detection framework today. Yang et al. proposed and published a 3D vehicle detection method constructed using multi-learning task modified faster R-CNN named MT-Faster R-CNN. The proposed network not only generates both 2D and 3D images which based on a single image, but it can also be more efficient if it is trained by end-to-end learning, this technique has been tested on a large number of experimental data shows that its performance is better than the baseline [10].

YOLO series. Before the YOLO series was proposed, R-CNN algorithm led target detection. The R-CNN series has high accuracy, however, because of the characteristics of its two-stage network structure, which makes its speed cannot meet the real-time requirement, it has been criticized. There is a trend to design a faster target detector to solve the problem. YOLO was proposed and introduced to the field of computer vision by Joseph Redmon et al. and has been growing in recent years [11]. YOLO means You Only Look Once, which is a single-stage target detection network that uses an end-to-end training and detection method to divide the image into many grids, with each grid predicting a certain number of bounding boxes, as well as the confidence and class of the objects within these boxes. YOLO is capable of completing the detection and categorization in a single pass, it far exceeds the speed of the previous target detection methods, especially in real-time applications. It performs particularly well in real-time applications. Fig. 3 shows the development of the YOLO series of models. It demonstrates the evolution of the technology from simple to complex, from basic to optimized, with each version update bringing performance improvements and extending the potential of YOLO applications in different scenarios.

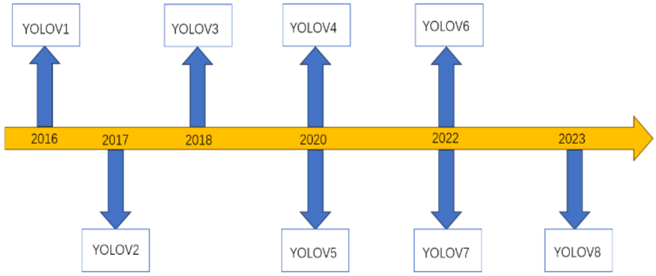

Fig. 3. YOLO evolutionary timeline (Picture credit: Original).

Liu et al. proposed an improved metal-based defect detection method for packaged semiconductor laser tubes based on the YOLO algorithm called YOLO-SO, which is 5.5 % better than the YOLO-V5, and the experiments proved that the YOLO-SO model has significant advantages in real-time monitoring of metal TO-base appearance defects [12]. Betti et al. developed a lightweight, high-accuracy YOLO network. The network, called YOLO-S, can be used to detect small targets in aerial images [13]. The algorithm is 25% to 50% faster than YOLO-V3 and 15% more accurate on the dataset than YOLO-V3, which can be done to truly meet the real-time detection requirements while measuring accurately.

Other popular target detection framework. In addition to the two mainstream families, R-CNN and YOLO, some other target detection frameworks are popular in the market today, such as SSD, RetinaNet, and so on. RetinaNet typifies single-stage object detection that addresses sample imbalance more flexibly. Liang et al. used various RetinaNet optimization methods to make the accuracy of target detection is improved, and the data showed that the model achieved a mAP value of 40.8, which is 4.3 higher than the baseline, which ensures the speed and also improves the target detection accuracy [14]. SSD is another single-stage detector different from YOLO, it differs from YOLO mainly in doing target detection, not only using the last obtained feature map, but also in the first few layers of the network 5 feature maps are also predicted, that is, multi-scale fusion, and then all of them are inputted to the last layer, which ensures that the accuracy of the recognition of small objects. In addition to this, SSD has been improved for multi-scale target detection networks (SSD-MSN) [15], for context-informed scene-aware target detection (CP-SSD) [16].

4 Application of DL Target Detection

In the current wave of digitalization sweeping, DL target detection technology is penetrating every aspect of thelives with its powerful functions, reshaping the future pattern of various industries.

4.1 Application of DL to Target Detection in Physiological Images

Sun et al. pointed out the great potential of deep neural networks for face recognition, and the proposed DeepID3 architecture achieved 99.53% face verification accuracy and 96% first-level face recognition accuracy [17]. Fig. 4 shows the experimental results.

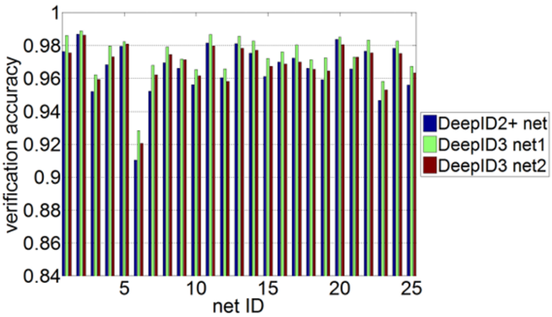

Fig. 4. LFW face verification accuracy of a single different network trained on the same face region [17].

The classification model of CNN-ML proposed by Buyukarikan et al. has strongly demonstrated the great potential of this model for classifying physiological disorder problems with an average accuracy of up to 96% in the dataset in the classification experiments performed on physiological disorder images [18]. Fig. 5 shows the experimental data.

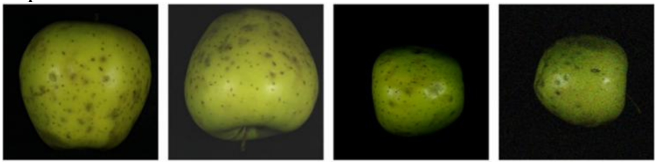

Fig. 5. Photographs of physiological disorganization as exemplified by apple bitter pit [18].

The development of physiological image detection technology based on DL has so far significantly improved the detection accuracy, and with the emergence of lightweight detection models, real-time monitoring can be achieved; however, due to the singularity of the experimental material and the limited funding, resulting in insufficient generalization ability and low universality. In the future, DL algorithms that are more suitable for face recognition should be further studied in larger-scale training data.

4.2 Application of DL in Motion Target Detection

Zhu et al. proposed a method called YOLOV3-SOD, which combines background compensation and object matching for motion target detection [19]. The algorithm has a recall of 86% and a mAP value of 91%, which is similar to the values produced by YOLOV3, but the size of the model is only half of YOLOV3. Various experimental data show that the algorithm performs better and can accurately detect moving targets in a variety of scenarios compared to other methods. Fig. 6 shows the test results.

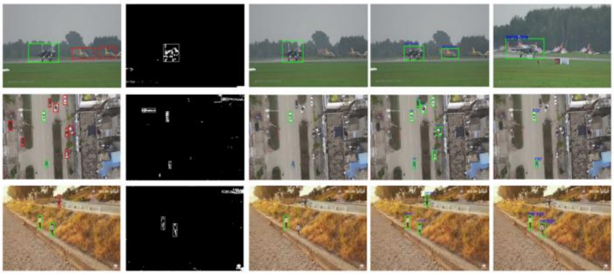

Fig. 6. Motion target detection results [19].

Zhao et al. put forward a DL algorithm based on wireless sensor networks for motion target detection [20]. The algorithm achieves target detection and localization by extracting visual latent representations and using a hierarchical clustering algorithm to cascade low-precision histograms with high-precision histograms to identify the target location. The algorithm has low energy consumption, high accuracy, and robustness. Fig. 7 illustrates the flow of the algorithm.

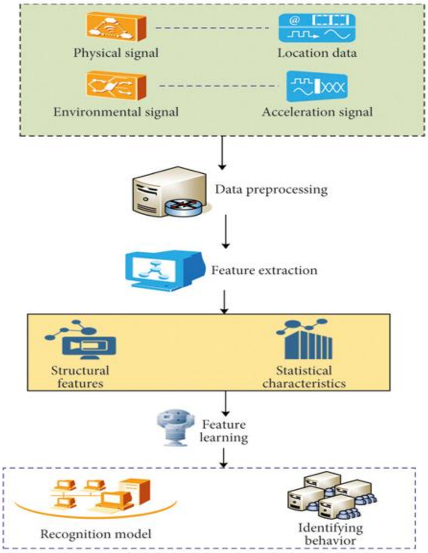

Fig. 7. Moving target recognition flowchart [20].

Detection of moving targets is often more difficult than detection of static targets, but despite this, researchers have made some achievements, realizing real-time detection and tracking as well as robustness enhancement in complex motion scenarios. However, due to the difficulty of the research problem, objectively there are still problems such as lower detection accuracy and speed, interference from complex backgrounds, and flexibility of the detection object; in the future, people should continue to research better performance algorithms to obtain motion information and use multimodal data and data preprocessing algorithms to build end-to-end networks for detection.

5 Conclusion

Target detection algorithms have been iterated to date with increasing accuracy and continuous improvement for small target detection and have evolved towards multimodal fusion and adaptive learning. Through the research, this paper reviews some of the most representative literature and elucidates their contributions to this important topic systematically and analyzes and compares their data to sort out the research progress of DL-driven target detection technology, clarify the development history of target detection technology and the application scenarios, advantages and disadvantages of each technology: two-stage target detection algorithms are more accurate but too slow for processing images and are suitable for processing images of small objects with high accuracy. The two-stage target detection algorithms represented by R-CNN are more accurate in processing images, but the speed is too slow, which is suitable for the occasions that require higher accuracy in image processing. In contrast, the single-stage target detection algorithms are more suitable for the occasions that require large-scale pipelined processing, because they do not need to select the candidate region. The operation rate is faster, but the processing is not accurate enough. However, DL-driven target detection technology still has shortcomings, such as small target detection in the localization difficulties, inaccurate feature extraction; multiple detections of the target resulting in the occlusion problem, etc., future research should continue to build on the existing algorithmic model optimization, towards the direction of the lightweight model, and continue to optimize the loss function. Finding solutions to these problems will greatly facilitate the development of DL-driven target detection methods.

References

1. Viola P, Jones M.: Rapid object detection using a boosted cascade of simple features, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, **1**, p. 1 (2001).
2. Omid-Zohoor, A.: Analysis of performance tradeoffs for embedded hog feature extraction, Stanford University (2017).
3. Hearst, M., Dumais, S., Osuna, E.: Support vector machines, in: IEEE Intelligent Systems and Their Applications, **13**(4), pp. 18-28 (1998).

4. Zhang, L., Zhang, D., Sun, J., et al.: Salient Object Detection by Local and Global Manifold Regularized SVM Model. Neurocomputing, **340**, pp. 42-54 (2019).
5. Talaei K., Ould S., Kaabouch, N.: Deep learning: systematic review, models, challenges, and research directions. Neural Computing and Applications, **35**(31), pp. 23103-23124 (2023).
6. Voulodimos, A., et al.: Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience, **2018** (1), p. 7068349 (2018).
7. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587 (2014).
8. Wang, X., Wang, J., Tang, P., et al.: Weakly- and semi-supervised fast region-based CNN for object detection. Journal of Computer Science and Technology, **34**, pp. 1269-1278 (2019).
9. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, **39**(6), pp. 1137-1149 (2017).
10. Yang, W., Li, Z., Wang, C., et al.: A multi-task Faster R-CNN method for 3D vehicle detection based on a single image - ScienceDirect. Applied Soft Computing, (95), p. 106533 (2020).
11. Redmon, J., Divvala, S., Girshick, R.: You only look once: unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 779–788 (2016).
12. Liu, J., Zhu, X., Zhou, X., Qian, S., Yu, J.: Defect Detection for Metal Base of TO-Can Packaged Laser Diode Based on Improved YOLO Algorithm. Electronics **11**, p. 1561 (2022).
13. Betti, A., Tucci, M.: YOLO-S: A lightweight and accurate YOLO-like network for small target detection in aerial imagery. Sensors, **23**, p. 1865 (2023).
14. Liang, Y., Wang, G., Li, W., et al.: A new object detection method for object deviating from center or multi object crowding. Displays, **2021**(6), p. 102042 (2021)
15. Chen, Z., Wu, K., Li, Y., et al.: SSD-MSN: An improved multi-scale object detection network based on SSD, IEEE Access, **7**, pp. 80622-80632 (2019).
16. Jiang, Y., Peng, T., Tan, N.: CP-SSD: context information scene perception object detection based on SSD. Applied Sciences. **9**, p. 2785 (2019).
17. Sun, Y., Liang, D., Wang, X., et al.: DeepID3: face recognition with very deep neural networks. **1502**, p. 00873 (2015).
18. Buyukarikan, B., Ulker, E.: Classification of physiological disorders in apples fruit using a hybrid model based on convolutional neural network and machine learning methods. Neural Computing and Applications, **34**(19): pp. 16973-16988 (2022).
19. Zhu, J., Wang, Z., Wang, S., et al.: Moving object detection based on background compensation and deep learning. Symmetry, **12**, p. 1965 (2020).
20. Zhao, L., Huang, Z.: A moving object detection method using deep learning - based wireless sensor networks. Complexity, (1): p. 5518196 (2021).