Evolution of Object Detection Algorithms Utilizing Deep Learning

Haojun Chen
Faculty of Engineering, Southern University of Science and Technology, Shenzhen, China

Keywords: Object Detection, Deep Learning, Feature Extraction.

Abstract: As one of the key technologies of image processing, object detection has been widely used in autonomous driving, urban planning and medical scenes. Since deep learning has advanced, deep learning-based object detection technology has advanced significantly. Deep learning-based object detection has become the mainstream algorithm in this field due to its high efficiency and accuracy. In this paper, according to the sequence of technology development, from the data, algorithm and other aspects of summary analysis. This paper provides an overview of datasets in the object detection domain and evaluation metrics for object detection algorithms. The algorithms for different categories in object detection are reviewed, including an exploration of traditional object detection algorithms, as well as the development and optimization of single and two-stage object detection algorithms. Brief introductions are provided regarding the characteristics and application scenarios of different algorithms. Finally, the standard performance of object detection algorithms is compared using experimental data. Simultaneously, the core issues of object detection algorithms are highlighted, along with a discussion on future development directions.

1 INTRODUCTION

Object detection is a core component of the visual system in the computer field, utilized to identify and detect objects in images, determine their categories and locations. Object detection finds widespread applications in fields like face recognition, pedestrian detection, vehicle detection, etc. For example, in specific scenarios like facial payment, identity verification, and autonomous driving.

Object detection's primary function is how to classify and locate variously sized and shaped objects. Before the extensive application of deep learning, traditional algorithms for object detection were separated into three phases: region proposal, feature extraction, and feature classification. Generally, manually crafted features were employed. The window sliding algorithm used in region proposal has high complexity of computing, leading to the generation of redundant data. Parameters of manually designed extractors, like Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradient (HOG) which have found extensive applications in the fields of image processing and computer vision used in feature extraction, are limited, resulting in low robustness and suboptimal extraction quality (Juan and Gwun 2013 & Wang et al 2009).

In the developmental history of object detection, the year 2012 marked a significant turning point as deep convolutional neural networks (DCNNs) made groundbreaking progress when classifying images. The effective usage of DCNNs in image classification has been extended to object detection, leading to the milestone Region-Based Convolutional Neural Network (R-CNN) detector (Girshick et al 2014). Since then, there has been a tremendous transformation in the area of identifying objects. Because large-scale datasets like MS COCO and GPU processing resources are readily available, numerous deep learning-oriented algorithms have been developed (Lin et al 2014). Due to the capability of neural networks to extract more robust and semantically meaningful features with an abundance of parameters, and the superior performance of classifiers, object detection with deep learning has developed into a crucial research emphasis on the field of computer vision. It seeks to identify interesting objects in pictures, precisely ascertain the category of each object, and provide bounding boxes for each target.

The second chapter of this paper will introduce commonly used datasets and evaluation metrics in the field of object detection. The third chapter will cover object detection-related algorithms, encompassing

Chen, H.
Evolution of Object Detection Algorithms Utilizing Deep Learning.
DOI: 10.5220/0012837700004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 182-188
ISBN: 978-989-758-690-3

one-stage, two-stage, and conventional object detection algorithms. In the fourth chapter, a comparison of data from different algorithms will be conducted. Finally, the conclusion will summarize the development and improvements in the field of object detection, providing an outlook for future developments.

2 DATASETS AND RELATED EVALUATION INDICATORS

2.1 Datasets

Data serves as the foundation for research, and any learning process is inseparable from the support of data. The widespread development of deep learning is also a result of the emergence of large-scale datasets. Datasets play a crucial role in the training of object detection algorithms. This paper introduces commonly used datasets in object detection, including PASCAL VOC and COCO datasets.

PASCAL stands for Pattern Analysis, Statistical Modeling, and Computational Learning. Although in recent years, object detection has predominantly used the larger COCO dataset, PASCAL, as a pioneer, carries substantial weight in the context of object detection. Researchers commonly use the VOC2007 and VOC2012 datasets in their studies (He et al 2015). VOC2007 comprises 20 categories, containing 9963 images. In contrast to the previous Caltech101 dataset, each image in VOC often contains multiple objects, establishing a standardized precedent. VOC2012 is an extension of the 2007 dataset with an increased image count of 11540.

The COCO dataset, with the full name Microsoft Common Objects in Context, is a sizable dataset for keypoint detection, segmentation, object detection, and captioning. COCO comprises a total of 328,000 images, with 80 categories of target objects in object detection. Keypoint detection involves 200,000 images with 250,000 keypoints annotated for human figures. Image segmentation includes 91 categories. Compared to PASCAL, COCO is suitable for more complex scene-based object detection and performs better in recognizing smaller targets.

2.2 Evaluation Indicators for Object Detection

Traditional evaluation metrics for detection machines include the miss rate and false positive rate of windows. But for object detection algorithms that are frequently employed in deep learning, performance metrics include Average Precision (AP), Accuracy, Precision, Recall Rate, Intersection Over Union (IoU), and mean Average Precision (mAP). Accuracy represents the proportion of correctly detected targets in every sample. The percentage of accurately identified positive samples among the positive samples that were detected is known as precision. Recall Rate describes the ratio of successfully detected positive samples among all actual positive samples. IoU, a statistic, quantifies overlapping degree between the bounding box that really exists and the bounding box that the model predicts.

3 INTRODUCTIONS TO ALGORITHMS FOR DEEP LEARNING BASED OBJECT DETECTION

This paper aims to explore the development and classification of object detection algorithms. It discusses the change from conventional object identification techniques to deep learning applications in one-stage and two-stage algorithms. By conducting a thorough analysis of the frameworks, advantages and application scenarios of these algorithms, this paper aims to provide insights into the evolution of object detection technology.

3.1 Traditional Object Detection Algorithms

Conventional object detection is divided into three sections.: region selection, feature extraction, and classifier. Initially, predefined regions are selected in the given sample image. Subsequently, features are extracted from these regions, and finally, classification is performed using a des1ignated classifier.

Region selection aims to find the target's location. As the target's position and shape are uncertain, traditional algorithms typically use a sliding window approach to traverse the entire image (Papandreou et al 2015). However, this generates a large number of redundant windows, impacting the efficiency of subsequent steps. The Selective Search algorithm effectively generates candidate regions with high recall, reducing the number of candidate boxes (Uijlings et al 2013). Traditional algorithms struggle to produce accurate candidate regions, especially in the case of small targets and complex scenes.

After obtaining the target's position, manually designed extractors are utilized for the extraction of features. Starting with SIFT, handcrafted local invariant features have been used extensively in various visual recognition tasks. These include Haar-like features, SIFT, Shape Context, among others. These local features are often aggregated through simple cascades or feature pool encoders. However, these methods have limitations in target feature extraction. In general, feature extraction in traditional object detection algorithms is time-consuming, less robust, and the extracted quality is not consistently high.

After obtaining the features, classifiers such as SVM and AdaBoost are commonly used to classify the features obtained in the previous step.

3.2 One-Stage and Two-Stage Algorithms

Based on the generation of candidate boxes, deep learning-based object detection algorithms are categorized into two groups: one-stage and two-stage. Two-stage approaches utilize methods like Selective Search and anchor-based techniques to generate candidate boxes. The ultimate detection outcomes are obtained based on the selected regions. This strategy accomplishes high accuracy but has a slower detection speed. One-stage methods directly detect results based on the original image, resulting in faster detection but lower accuracy.

3.2.1 One-Stage

One-stage algorithms directly process an image through a single network to obtain detection classification and predict bounding box boundaries, omitting the generation of candidate boxes. Compared to traditional algorithms and two-stage algorithms, one-stage methods offer faster detection, making them more suitable for practical applications with rapid detection requirements.

Typical one-stage algorithms comprise the SSD series and You Only Look Once (YOLO) series (Redmon et al 2016). The YOLO algorithm (YOLO v1) was put out in 2016 by Redmon et al, pioneering the approach of treating Identification of objects as a regression issue. Fig. 1 depicts the algorithm's network structure (Redmon et al 2016. It integrates feature extraction, classification, and regression within a single deep convolutional network, achieving real-time detection (Redmon et al 2016). The YOLO system generally consists of three parts: preprocessing and resizing of the image, inputting the processed image into a convolutional neural network, and finally selecting detection results based on confidence scores.

In response to this issue, Redmon and Farhadi and others proposed subsequent improvements to the YOLO series with YOLOv2 (Redmon and Farhadi 2017). Due to the simple network architecture of YOLO, as of 2024, the YOLO series has iterated to YOLOv8 through improvements in training strategies, network structures, multi-scale detection, loss functions, label assignment methods, and more.

The YOLO series, known for its fast and accurate characteristics, is generally applicable to scenarios requiring real-time detection, like video surveillance and automatic driving.

The Single Shot Multibox Detector (SSD), proposed by Liu, was initially designed to address the low accuracy of YOLOv1 in detecting small targets and its insensitivity to scale (Liu et al 2016). By combining the idea of extracting multiple candidate regions from Faster R-CNN as Regions of Interest (ROI) with a regression approach, SSD effectively seeks a compromise between one-stage object detection algorithms' speed and accuracy (Ren et al 2018). Moreover, In Fig. 2 (Liu et al 2016), the model diagram of SSD reveals the incorporation of multiple convolutional layers in its network structure to acquire feature maps of distinct scales, addressing the issue of scale insensitivity. Shallow maps with features are used for tiny targets, while deep feature maps are employed for large targets.

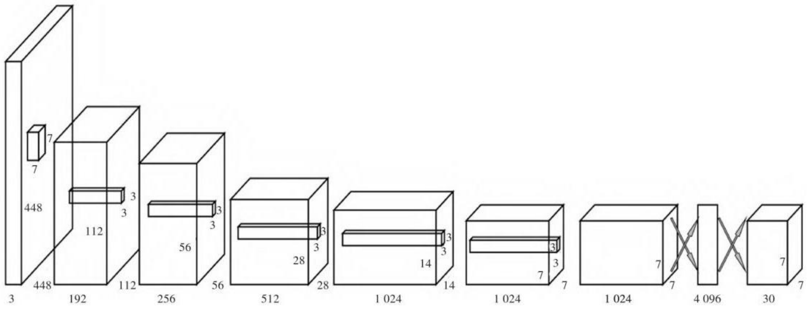

Figure 1. YOLOv1.

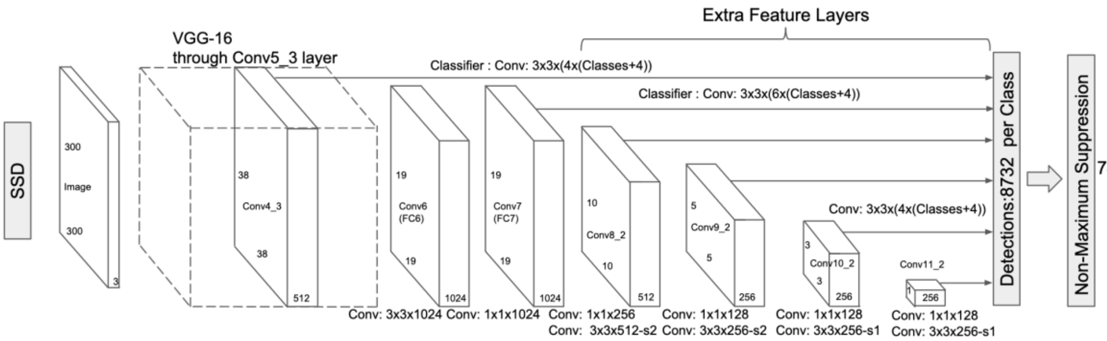

Figure 2. Detection models for SSD.

However, in the detection process of SSD, the detection boxes may repeatedly detect the same target, increasing computational load. Additionally, the representation capability of shallow feature maps is not sufficiently strong. To address the mentioned issues in the SSD algorithm, optimizations have been proposed. Jeong et al. introduced the RSSD algorithm (Jeong et al 2017), which replaces the backbone network of VGGNet with ResNet, resulting in an improvement in detection speed. Fu et al. proposed the DSSD algorithm (Fu et al 2017), based on the ResNet101 network architecture. DSSD incorporates residual modules before classification and regression and adds deconvolutional layers after the auxiliary convolutional layers in SSD to enhance precision of detection for tiny targets. Li et al. drew inspiration from the Feature Pyramid Network (FPN) and proposed the FSSD algorithm (Li and Zhou 2017). FSSD concatenates feature maps from several layers at various scales, creating a new feature pyramid that is fed back to the multi-box detector for prediction. FSSD demonstrates significant performance improvement compared to SSD, even with a slight decrease in speed.

3.2.2 Two-Stage

To address the issue of low accuracy in traditional object detection algorithms when dealing with large amounts of data or features, algorithms for object detection based on deep learning have been introduced into the field of object detection. First, two-stage object detection techniques produce category-agnostic candidate boxes on input images to obtain initial proposed regions. Then, based on these proposed regions, a second localization is performed to determine the detection position. Detection classification is subsequently carried out based on the detected position. This approach enhances detection accuracy and is appropriate for highly accurate detection requirements.

Typical two-step algorithms for detection consist of R-CNN series, Spatial Pyramid Pooling Network (SPP-Net), R FCN, FPN, etc. R-CNN is the earliest object detection algorithm that used selective search to generate candidate boxes, introducing the concept of Region of Interest (ROI). Fig. 3 illustrates the fundamental structure of R-CNN (Girshick et al 2014). It generates candidate boxes in the image using selective search, scales all candidate boxes to a uniform size. Then applies them to a deep convolutional neural network in order to extract features. Finally employs a support vector machine for classification based on the extracted feature vectors.

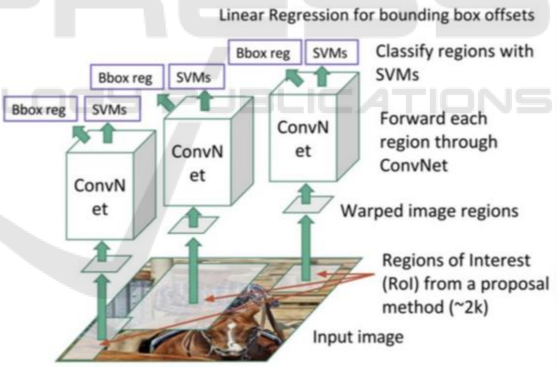
Linear Regression for bounding box offsets
Bbox reg
SVMs
Bbox reg
SVMs
Bbox reg
SVMs
Classify regions with SVMs
ConvNet
ConvNet
ConvNet
Forward each region through ConvNet
Warped image regions
Regions of Interest (RoI) from a proposal method (~2k)
Input image

Figure 3. R-CNN.

However, due to the two-stage training required for each candidate box, the detection efficiency is reduced, leading to slower operation speeds. Therefore, The R-CNN-based SPP-Net was proposed by He. With SPP-Net, not all candidate boxes need to be fed into the neural network; instead, pooling networks are used to extract features only once. Similarly, Girshick et al. inherited R-CNN and adopted the characteristics of SPP-Net (Girshick et al 2015). Based on the method of extracting candidate regions, they modified the network to produce dual-layer outputs, proposing Fast R-CNN (Girshick et al

2015). Both methods involve feeding the image into the network only once, effectively improving training efficiency. It is worth noting that the previous algorithms' testing time did not include the time taken for selective search. In practice, a significant portion of testing time is allocated to selective search. To optimize the efficiency of this stage and address the computational intensity issues with the use of selective search in Fast R-CNN and SPP-Net, Ren introduced the Faster R-CNN algorithm. Building upon the architecture of Fast R-CNN, they added a Region Proposal Network that uses anchors at different scales in order to replace selective search, further enhancing the training speed of the network.

The development of deep learning has exposed the issue of repeated calculations for each Faster R-CNN's ROI, leading to a continuous increase in computational load. Dai discovered that after ROI pooling, the network's layers lose their translational invariance. This means addressing the issue that changes in the image do not alter the image properties, allowing for weight sharing. Furthermore, the quantity of layers after ROI pooling directly impacts detection efficiency. Therefore, they proposed the Region-based Fully Convolutional Network (RFCN) (Dai et al 2019). This approach addresses the issue by utilizing position-sensitive score maps.

4 ALGORITHM PERFORMANCE EVALUATION AND COMPARISON

Performance metrics in object detection primarily consist of mAP, AP, recall, accuracy, and precision. mAP involves first calculating the AP for every single class, and after that computing the average of these AP values. mAP is the primary performance metric used in object detection algorithms. Below is a performance comparison of different algorithms on various datasets (in this paper, algorithms using the VOC dataset are assumed to be tested on VOC2007 unless otherwise specified).

Table 1. Comparison of algorithms for object detection (Zhao et al 2020 & Cai 2023).

Algorithm	**Backbone Network**	**Databases**	**Detection Speed/(frame/s)**	**mAP/%**
R-CNN	AlexNet	ILSVRC 2012+VOC 2007	0.03	58.5%
R-CNN	VGG-16	ILSVRC 2012+VOC 2007	0.5	66.0%
SPP-Net	ZF-5	ImagNet2012	2	59.2%
Fast R-CNN	VGG-16	VOC2007+VOC2012	3	70.0%
Faster R-CNN	ResNet101	VOC2007+VOC2012	5	76.4%
Faster R-CNN	VGG-16	VOC2007+VOC2012	7	73.2%
MaskR-CNN	ResNet101	MSCOCO	4.8	33.1%
R-FCN	ResNet101	VOC2007+VOC2012	5.8	79.5%
YOLOv1	VGG-16	VOC2007+VOC2012	45	66.4%
YOLOv2	DarkNet-19	VOC2007+VOC2012	40	78.6%
YOLOv3	DarkNet-53	MSCOCO	51	33.0%
YOLOv4	CSP-DarkNet53	MSCOCO	66	43.5%
YOLOx	Focus+DarkNet-53	MSCOCO	57.8	51.2%
FPN	ResNet-50	MSCOCO	5.8	35.8%
SSD321	ResNet101	VOC2007+VOC2012	11.2	77.1%
SSD513	ResNet101	VOC2007+VOC2012	6.8	80.6%
RSSD300	VGG-16	VOC2007+VOC2012	35	78.5%
RSSD512	VGG-16	VOC2007+VOC2012	16.6	80.8%
DSSD321	ResNet101	VOC2007+VOC2012	9.5	78.6%
DSSD513	ResNet101	VOC2007+VOC2012	5.5	81.5%
FSSD300	VGG-16	VOC2007+VOC2012	65.8	78.8%
FSSD513	VGG-16	VOC2007+VOC2012	35.7	80.9%

5 CONCLUSION

One essential component of computer vision is object detection. This paper provides a comprehensive review of object detection algorithms, including traditional detection methods, one-stage YOLO series algorithms, SSD series algorithms, and two-stage R-CNN series algorithms. Throughout the development of object detection algorithms, researchers continuously optimize algorithms by improving network architectures, enhancing original data, and optimizing loss functions, leading to significant improvements in both accuracy and speed. With deep learning's ongoing advancement, the application scope of object detection is becoming increasingly widespread.

As algorithms for object detection grounded in deep learning continue to develop and be applied, the domain of object detection has made significant progress. However, numerous challenges remain unresolved, including the detection of tiny objects, insufficient robustness, and model architecture optimization.

Small object detection is a critical aspect of object detection, as realistic scenes from the real world involve detecting objects of different scales, especially small objects. Due to the small size, indistinct features, and low contrast of small objects, accurately detecting small targets becomes challenging. Therefore, one of the key future approaches is to further optimize small object detection by using attention processes, multi-scale detection methods, and feature enhancement techniques.

In real-world scenarios, real images are prone to occlusion, blurring, changes in lighting, noise, and other external variations that can hinder effective object detection. Addressing how to make models more adaptable to specific real-world scenarios is a significant challenge. Therefore, continually improving model performance through methods like incorporating contextual information, selective parameter sharing, and complementary feature fusion is crucial to adapt to specific scene-based object detection requirements.

The underlying network architecture is the foundation of object detection algorithms, and optimizing the network architecture has always been an important area of study for object detection. Currently, the selection of network architectures has some randomness, displaying different performances for different detection tasks. Therefore, enhancing the processing efficiency of network architectures is an important future direction.

There has been considerable research on 3D object detection, but most algorithms are not yet mature. Conducting precise 3D object detection using high-precision LiDAR point clouds is expensive and sensitive to weather conditions. Therefore, how to elevate 2D images to 3D for detection has become a research direction. One approach is to address this problem by using methods such as inverse perspective mapping (IPM) and orthogonal feature transformation (OFT) to convert perspective images into bird's-eye views (BEV). Another approach involves obtaining relationships through overall size and inter-keypoint size.

REFERENCES

L. Juan and O. Gwun, "A comparison of SIFT, PCA-SIFT, and SURF," in *International Journal of Image Processing* (IJIP), vol. 3, no. 4, 2013, pp. 143-152.

X. Y. Wang, T. X. Han, and S. C. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proceedings of the 12th IEEE International Conference on Computer Vision*, Kyoto, Japan: IEEE, 2009）, pp. 32-39.

R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *in Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA: IEEE, 2014, pp. 580-587.

T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al. "Microsoft COCO: Common Objects in Context," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740-755.

K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2015, 37(9)), pp. 1904-1916.

G. Papandreou, I. Kokkinos, P.-A. Savalle, "Modeling Local and Global Deformations in Deep Learning: Epitomic Convolution, Multiple Instance Learning, and Sliding Window Detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 390-399.

J. R. R. Uijlings, K. E. A. van de Sande, Gevers T, Smeulders A. W. M., "Selective Search for Object Recognition," in *International Journal of Computer Vision* (2013, 104(2)), pp. 154-171.

J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, 2016, pp. 779-788.

J. Redmon, A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, HI, USA: IEEE, 2017, pp. 6517-6525.

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 14th European Conference on Computer Vision (ECCV)*, Amsterdam: Springer, 2016, pp. 21-37.

Mengye Ren, et al., "Meta-learning for Semi-Supervised Few-Shot Classification," arXiv preprint arXiv:1803.00676 (2018).

J. Jeong, H. Park, N. Kwak, "Enhancement of SSD by Concatenating Feature Maps for Object Detection," *arXiv preprint arXiv:1705.09587* (2017).

C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, "DSSD: Deconvolutional Single Shot Detector," *arXiv preprint arXiv:1701.06659* (2017).

Z. Li, F. Zhou, "FSSD: Feature Fusion Single Shot Multibox Detector," *arXiv preprint arXiv:1712.00960* (2017).

R. Girshick, F. Iandola, T. Darrell, J. Malik, "Deformable Part Models are Convolutional Neural Networks," in *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Boston, MA, USA: IEEE, 2015, pp. 437-446.

J. Dai, Y. Li, K. M. He, J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," [Online]. Available: https://arxiv.org/pdf/1605.06409.pdf. Accessed on June 20, 2019.

Y. Q. Zhao, Y. Rao, S. P. Dong, J. Y. Zhang, "Survey on Deep Learning Object Detection," in *Journal of Image and Graphics* (2020, 25(4)), pp. 629-654.

C. Jialei Cai, "A Review of Deep Learning-based Target Detection Algorithms and Applications," in *Network Security Technology and Applications* (2023, 11), pp. 41-45.