Research Advanced in Object Detection based on Deep Learning

Boao Li[1, *], Wangzhixin Qi[2], Xiaoning Zhang[3]

[1]North China University of Technology, Beijing, China

[2]International high school of Yuhuatai, Nanjing, China

[3]King's College London, London, UK

* Corresponding Author Email: liboao@mail.ncut.edu.cn

Abstract. Object detection has always been a fundamental research topic in the computer vision community, which focuses on predicting the category and location of all objects in the scene. In last several years, progressing from the rapid development of deep learning, the speed and accuracy of general object detection methods have also achieved significant breakthroughs. This paper aims to report the latest research progress in the field of object detection based on deep learning to inspire and promote subsequent research. Specifically, this paper systematically introduces the research progress of predecessors from four aspects: dual-stage, single-stage, Transformer-based and key point, including the design ideas and basic processes of representative algorithms. In addition, this paper also quantitatively compares the performance of different methods on common data sets to further distinguish the benefits and disadvantages of different categories of methods. Finally, this paper summarizes the challenges that still exist in the field of object detection and looks forward to future development directions.

Keywords: computer vision, natural language process, Object detection, anchor-free.

1. Introduction

With the fast development and the wide application of artificial intelligence technology, autonomous vehicles are becoming an important part of future transportation in our daily lives. Autonomous driving technology tries to achieve autonomous navigation and control of vehicles by computer vision, sensor technology and deep learning algorithms. In recent years, the research of autonomous driving has made major breakthroughs, which not only promote the transformation and upgrading of the automotive industry, but also bring new opportunities for the development of modern cities and intelligent transportation.

In the process of the realization of autonomous driving technology, environmental perception is a vital link, which is related to the safety directly and reliability of autonomous vehicles. The core task of an environmental awareness system is to accurately detect and identify all kinds of aims on the road, including pedestrians, vehicles, traffic signs, road obstacles and so on. Among them, Object Detection, as the basic link of environment perception, plays a key role. The accuracy and real-time performance of object detection directly affect the decision-making and control ability of automatic driving systems. There is still lots development space and research value in the future technology. To popularize the public's doubts about object detection technology in autonomous driving, this article will show the development process of the technology, analyze the shortcomings, discuss the social significance, and propose current challenges and future directions.

Based on a review of the general and specialized approaches to object detection task modeling in automated driving, we classify the main approaches into four categories according to the evolution route and development time of the model: two-stage, single-stage, key point-based and transformer-based approaches. Through analyzing the representative models of various methods and the pros and cons of existing technologies, we divided the main features of this generation of methods and the improvement of the next generation of methods over the previous generation. At the end of this paper, the current problems and challenges of object detection technology are discussed. With the continuous progress of autonomous driving technology, object detection algorithms still need to be

continuously innovative in the face of complex environments and diverse scenarios. Multi-domain object detection, multi-mode fusion, small object detection, lightweight network architecture, video detection, weak supervision and small sample detection will become the focus of future research. Breakthroughs in these directions will further promote the intelligence and safety of autonomous driving systems, and promote their faster application to reality.

2. Method

2.1. Two-stage Detection Approaches

Two-stage detection methods decompose the object detection task into two stages: region proposal generation, target classification, and bounding box regression. With the refinement of the two phases, the two-phase detection algorithm can identify and localize target objects more accurately, especially when dealing with small and multi-scale targets. Also, two-stage detection possesses greater flexibility. Dual-stage detection allows the algorithm to focus on generating high-quality candidate regions in the first stage, and then focus on classification and precise adjustment of the bounding box in the second stage, a separation that makes the algorithm more flexible. Moreover, two-stage detection also possesses greater scalability and can be applied to different datasets and tasks by simply making appropriate adjustments to the second-stage classifiers.

As the most representative method, Faster R-CNN achieves high-precision detection performance on multiple datasets through a two-stage network structure with a region proposal network (RPN) and a classifier, which is easy to migrate to other scenes by changing the target class in the dataset [1-2]. Moreover, by decreasing the number of shared convolutional layers, the amount of computation is reduced, which is widely used in the fields of automatic driving, security monitoring, face detection, object recognition and so on. Despite the above advantages of Faster R-CNN, there are some drawbacks, such as (1) the small resolution of the feature maps of the convolutional extraction network, which may not be conducive to the detection of small and multi-scale objects; (2) NMS may not be friendly to obscured objects, leading to missed detections [3]; (3) The loss of accuracy in RoI Pooling is limited [4]; (4) A large number of parameters and unshared computations in the fully connected layer; (5) the fact that positive and negative sample equalization methods may not be optimal; (6) the relatively slow detection speed..

Mask R-CNN [5] has added a new branch to Faster R-CNN for the purpose of predicting the segmentation mask for each object to achieve accurate instance segmentation. In addition, it inherits the RPN part of Faster R-CNN and integrates a multi-task learning framework, demonstrating the trend of integrating complex models in the field of deep learning. This algorithm is capable of generating a segmentation mask for each detected object to achieve accurate instance segmentation, which gives it a wide potential for practical applications such as human pose estimation, object segmentation, etc., to achieve efficient computation while maintaining accuracy. As a result, Mask R-CNN has achieved remarkable results in instance segmentation tasks through its innovative network structure and training method and has become a significant milestone in the field of computer vision.

2.2. Single-stage Detection Approaches

Single-stage object detection is a class of object detection methods that regresses the target bounding box and categories directly from the input image without going through the Region Proposal step. Compared to two-stage methods, single-stage methods typically have higher detection speeds and are suitable for real-time applications.

In order to cope with the problems of the complexity of the two-stage object detection model, multiple parameters, long training time, and poor real-time detection, Redmon et al. [6] proposed the YOLO (You Only Look Once) algorithm in 2015. This is the first single-stage detector in the deep learning space. the big advantage of YOLO is that its processing speed is very fast, with the enhanced version reaching up to 45 FPS on the GPU, and the fast version even reaching 155 FPS. The YOLO

algorithm directly divides the whole image into $S \times S$ grid cells through a single neural network, and the task of each grid is to detect targets within it and predict the target class, confidence level, and bounding box location [7]. The whole process passes through several layers of convolutional and pooling layers and then two fully connected layers to output the classification result and bounding box of the target. Finally, the algorithm uses threshold filtering to remove low-confidence targets and in order to detect objects, non-maximum suppression (NMS) will remove redundant bounding boxes. Although YOLO significantly outperforms the two-stage Faster R-CNN in terms of detection speed, there are some limitations. First, since YOLO uses an $S \times S$ grid for prediction, if there are multiple targets in the same grid, it may lead to missed detection, especially poor detection of small targets. Second, instead of using the anchor frame mechanism of Faster R-CNN, YOLO directly predicts the absolute position, which increases the difficulty of model training. In addition, the aspect ratio of the prediction frame is preset based on the training set, which is a weak generalization to new datasets. Finally, since YOLO can only output the bounding box that has the highest IoU when dividing the grid, this means that at most only one target can be detected per grid, which is less effective for detecting scenes containing multiple small targets (e.g., flocks of birds).

To address the shortcomings of YOLO in small object detection, Liu et al. [8] provided the SSD (Single Shot MultiBox Detector) algorithm in 2015. SSD combines the advantages of YOLO's rapid detection speed and Faster R-CNN's accurate localization by introducing multi-level feature maps for detecting targets at various scales. Specifically, SSD uses shallow feature maps to detect small targets and deep feature maps to detect large targets, thus dramatically improving the detection of small targets [9-10]. SSD utilizes multi-reference and multi-resolution detection techniques to detect objects of different sizes through different levels of networks. Meanwhile, SSD borrows the Anchor trick from Faster R-CNN to set up a priori frames with different aspect ratios in advance, and predicts the object detection frames based on them, thus reducing the training complexity. In order to further improve the detection effect, Liu et al. also introduced the difficult sample mining technique to address the issue of focusing on difficult samples during the model training process. However, SSD still has some limitations: first, feature maps at different scales are detected independently, which may lead to repeated detection of the same target and increase the computational effort of the model. Second, although the shallow feature map is used for detecting small targets, there is still opportunity for improvement due to less semantic information, resulting in the detection of small-sized objects. Nevertheless, the performance of SSD on the PASCAL VOC and MS-COCO datasets proves that it achieves a good balance between detection speed and accuracy.

In response to the status quo that single-stage object detection algorithms were generally lower in accuracy than two-stage detection algorithms at that time, Lin et al. [11] found that the principle reason for this problem was the imbalance of positive and negative samples, which made it difficult for the model to learn difficult samples effectively during training. To solve this problem, they proposed the RetinaNet algorithm in 2017. RetinaNet combines ResNet and FPN as feature extraction networks to obtain a multi-scale feature map of an image, and two fully convolutional networks (FCNs) to perform classification and regression tasks, respectively. In order to overcome the problem of positive and negative sample imbalance, RetinaNet introduces a focal loss function, which serves as a replacement for the traditional cross-entropy loss function [12]. Focal loss addresses the issue of instance sample imbalance through increasing the weight of difficult samples so that the model pays more attention to these sparse and hard-to-classify samples during training, which significantly improves the accuracy of the single-stage detector to a level equivalent to the performance of the two-stage detector. Despite RetinaNet achieves breakthrough in accuracy, its detection speed is slower than other single-stage detection algorithms, which limits its performance in real-time detection applications. Overall, RetinaNet introduces focal loss while guaranteeing accuracy, making it an important innovation in single-stage object detection algorithms.

EfficientDet [13] is suggested by the Google team in 2019 to address the problem of low detection accuracy in regression-based detection algorithms. The EfficientDet family consists of 8 models (D0-D7), which are capable of achieving the then optimal detection results while satisfying different

resource constraints. As the model number increases, the model's detection accuracy steadily increases, while its inference speed slows down [14-15]. The main innovations of EfficientDet include: first, the introduced Weighted Bidirectional Feature Pyramid Network (BiFPN), capable of rapidly and effectively fusing multi-scale features, integrate more features by connecting input and output nodes at the same level, while combine bottom-up and top-down paths to form a reusable base layer. BiFPN improves accuracy by 4 percentage points over traditional Feature Pyramid Networks (FPNs) with the same backbone, EfficientNet, and fewer parameters. Second, EfficientDet uniformly scales the resolution, depth, and width of the model through a composite scaling method, resulting in a reduction of the number of parameters in the model by a factor of 4-9, and a reduction of FLOPs by a factor of 12-42 compared to traditional object detection algorithms. In addition, EfficientDet uses EfficientNet as the backbone network, allowing the model size to be controlled by adjusting the version of EfficientNet (B0-B6), while different EfficientDet structures can be flexibly constructed by modifying the quantity of channels and the repetition of layers in the BiFPN, as well as the input image's resolution [16]. EfficientDet-D7 reached a state-of-the-art AP of 55.1% on the MS-COCO dataset with a single model and single scale. However, EfficientDet's pre-training is costly, and the authors used 32 TPUs to train on a large dataset to accomplish SOTA (state-of-the-art) performance, with high time and hardware costs required to train the full model. Overall, EfficientDet has the advantages of fewer parameters, faster inference, and higher accuracy, making it an ideal choice for efficient object detection in resource-constrained situations.

2.3. Key point-based Detection Approaches

Key point-based methods detect the key points of a target for localization and tracking, whose key innovation is the anchor-free design. Earlier generations of object detectors, such as the methods of two-stage and one-stage, relied heavily on the use of anchor boxes to propose and locate objects, which improved efficiency and accuracy. However, the need for numerous anchor boxes to ensure accurate detection led to substantial resource consumption, as well as hyperparameter tuning and computational challenges. Additionally, both one-stage and two-stage detectors suffer from significant localization inaccuracies, especially when detecting dense and small objects [17]. This motivated the development of anchor-free models to address these issues, leading to the proposal of keypoint-based approaches. This approach can be primarily divided into two subcategories: corner-based methods and center-based methods [18]. In this section, we will introduce two representative models, CornerNet and CenterNet, to analyze the significant advancements, features, strengths, and limitations of these methods.

CornerNet [19] was the first model to frame object identification as the process of identifying and classifying corners using embeddings. In order to calculate an object's bounding box, this model first locates two keypoints: the top-left and bottom-right corners. Following detection, redundant bounding boxes are removed using Non-Maximum Suppression (NMS), leaving just the bounding boxes with the highest confidence ratings. Furthermore, CornerNet presents a new pooling layer called corner pooling that improves corner point localization by merging nearby contexts and aggregating data along the corner's two sides. CornerNet offers several strengths due to its innovative design. Firstly, as a one-stage approach, CornerNet operates faster and more simply compared to two-stage methods like DeNet, as it bypasses the need for regional proposals and separate classification and regression stages. Secondly, reducing the dependency from four sides to two sides simplifies the localization of the bounding box center. Despite its high accuracy, the introduction of corner points increases computational complexity. Moreover, detecting small objects remains challenging for CornerNet, an issue that has been partially addressed in its successor, CenterNet.

CenterNet, introduced by Zhou et al. [20], differs from CornerNet primarily through the introduction of the center point for detection. CenterNet utilizes three key points (center, top-left, and bottom-right) to detect objects, integrating object size to perform object detection and human pose estimation. The model typically uses ResNet or Hourglass as its backbone network. During training, the Center Point Offset and Keypoint Heatmap Loss are crucial for refining the model. The use of

heatmaps, in particular, ensures that key points receive strong feedback when correctly positioned. CenterNet offers several advantages over CornerNet. First, by locating objects via the center point, CenterNet improves the accuracy of detecting multiple and small objects. Second, CenterNet eliminates the corner pooling and embedding vectors used in CornerNet, thereby simplifying the model structure and reducing the complexity of training the Fully Convolutional Network. Third, in terms of efficiency and real-time performance, CenterNet is faster because it detects center points and size features directly, while CornerNet requires additional calculations for corner matching. Finally, CenterNet reduces reliance on hyperparameters, facilitating faster development across different datasets and application scenarios.

However, CenterNet is not without challenges. Its performance may degrade on high-resolution images due to the higher spatial resolution required for accurate center point detection. Additionally, keypoint accuracy can be problematic in complex scenarios.

2.4. Transformer-based Detection Approaches

Transformer is a model based on an attention mechanism suitable for processing sequential data while being able to compute in parallel. The advantage of Transformer is not only that it can efficiently perform parallel computation and increase the training speed, but also that it can better capture the global dependencies in the sequence data.

DETR (Detection Transformer) is the initial end-to-end object detection algorithm utilizing Transformer architecture, which models object detection as an ensemble prediction problem, combining a hybrid structure of CNN and Transformer to simplify the object detection process [21]. Below is the framework of DETR: first, DETR uses CNN to extract image features, and these features are summed up with the position encoding and passed to the Transformer encoder. The output of the encoder, together with a set of learnable Object Queries, serves as the input to the decoder, which is processed by a Feed-Forward Network (FFN) and finally decoded into bounding box coordinates and category labels. The detection results are matched one-to-one by the Hungarian algorithm to acccess the loss, thus avoiding the traditional non-maximum suppression (NMS) step. DETR achieves an average precision (AP) of 42.0% on the COCO dataset, which is comparable to the optimized Faster R-CNN. However, DETR has some limitations. On the one hand, it performs poorly in small object detection due to the high computational complexity associated with high-resolution images. On the other hand, However, due to the computational intricacies of the attention mechanism of the Transformer, the convergence speed of DETR is slower, 10~20 times slower than Faster R-CNN. Attention weights of pixels on the feature map are almost uniformly distributed when the Transformer is initialized, requiring more training rounds to focus attention on sparse meaningful locations, while high computational complexity and memory complexity are also major issues. DETR, as the first Transformer-based object detection algorithm, provides an anchor-free detection framework that simplifies the overall process of object detection. Although it is more excellent in speed and accuracy, it still needs further improvement in small object detection and training efficiency. These defects also become the main improvement direction for subsequent research.

Deformable DETR is proposed for the problems of slow convergence of DETR training and poor performance of small object detection [22]. The algorithm integrates the advantages of deformable convolution and Transformer, and solves the defects of DETR by introducing a deformable attention module to replace the original Transformer attention module. The deformable convolution adapts to the shape of the object and improves the model's capacity to adjust to object deformation, while the deformable attention module significantly reduces the computational complexity and accelerates the model convergence by sparse sampling and focusing only on the key sampling points around the reference point. In Deformable DETR, the attention module only interacts with key sampling points, avoiding the intensive computation of global features and dramatically reducing the computation of the original self-attention module. This sparse spatial sampling enhances the performance of small object detection and also makes the merging of multi-scale information more organic and reduces the dependence on FPN. On the COCO dataset, Deformable DETR has 10 times fewer training cycles

than DETR and improves the AP value of small object detection by 5.9%. Although Deformable DETR improves the training speed and small object detection accuracy, the number of tokens increases compared to the original DETR due to the use of multi-scale features.

UP-DETR (Unsupervised Pre-trained DETR) [23] is an unsupervised pre-trained object detection model designed to enhance the effectiveness and convergence rate of DETR. Encouraged by the effectiveness of unsupervised pre-training in NLP, the model employs a "random query patch detection" pre-training task for unsupervised pre-training of the Transformer in DETR. In this process, UP-DETR devised methods for freezing the CNN backbone and image block feature reconstruction to ensure the categorical discriminative nature of the features, and introduced object query shuffling and attention masking mechanisms to support multi-query block localization. The pre-training process of UP-DETR allows the network to learn to detect the location of an image block without human annotation by feeding the original image into the encoder and randomly cropping a number of image blocks from the original image into the decoder. The pre-trained DETR model is able to efficiently localize the position of the image blocks as they are fed to it. In the Pascal VOC dataset, UP-DETR achieved 56.1% AP using 150 training rounds, which is a 6.2% improvement over DETR's 54.1% AP. In the COCO dataset, UP-DETR achieves 42.8% AP with 300 training rounds, showing higher accuracy and faster convergence than DETR. This demonstrates the effectiveness and feasibility of the unsupervised pretraining strategy in object detection.

3. Experiment

3.1. Datasets and Metrics

To evaluate the performance of target detection algorithms, researchers have used a variety of datasets and evaluation metrics. Commonly used public datasets include Pascal VOC (Visual Object Classes) and COCO (Common Objects in Context). The Pascal VOC dataset is one of the most famous target detection datasets in the field of computer vision, which contains two versions, 2007 and 2012, consisting of 5,000 and 21,000 images, respectively. The Pascal VOC dataset provides detailed annotation information for each image, including the category of the object, and the location of the bounding box. The COCO dataset was created by Microsoft Research in 2014 to provide a larger, more complex, and more challenging dataset. The COCO dataset contains 328,000 images, covering 91 categories of objects, and contains more than 2 million annotated instances. Among them, the size and posture of the objects vary significantly, which increases the difficulty of detection.

Commonly used evaluation indicators in target detection research include average precision (AP), mean average precision (mAP), precision, recall, and detection speed (FPS). AP is a performance indicator for measuring the target detection algorithm in a certain category, usually obtained by calculating the area under the precision curve (AUC) under different recall rates. mAP is a comprehensive performance indicator for measuring the target detection algorithm in multiple categories, usually obtained by calculating the average AP of all categories. Precision refers to the proportion of positive examples detected that are truly positive, while recall refers to the proportion of all actual positive examples that are correctly detected. FPS is an indicator to measure the real-time performance of the detection algorithm, which describes the number of image frames processed per unit time.

3.2. Performance Comparison

To quantitatively compare the advantages and disadvantages of different methods, this paper selects the most commonly used VOC2007 and COCO dataset as the basis and reports the performance of existing detection methods respectively in Table 1. It can be clearly observed that the detection method based on the Transformer can achieve significant improvement in detection accuracy. For instance, on the COCO dataset, the Transformer-based method can obtain an mAP of over 52%, which is an average improvement of over 8% in accuracy compared to the key point-based method. We believe that the possible reasons for the above results come from the global attention

mechanism and multi-scale detection design in the Transformer architecture. The DETR-based method does not use traditional anchor boxes, but directly predicts the bounding box and category of the object from the image. By setting a fixed set of queries, each query is responsible for detecting a potential object. This method avoids post-processing steps such as non-maximum suppression (NMS) and simplifies the overall process.

Table 1. Accuracy (mAP) comparison between different representative methods

Method	Category	VOC2007	COCO
Fast RCNN	Two-stage method	70.0	19.7
Faster RCNN		73.2	21.9
SSD	Single-stage method	81.6	26.8
YOLOv4		81.8	43.5
Retina-Net		79.5	36.2
RefineDet		79.2	39.1
CornerNet	Key point-based method	79.5	42.5
CenterNet		78.0	41.8
FCOS		78.8	44.7
Deformable DETR	Transformer-based method	79.5	52.3
Swin Transformer		81.0	57.7

4. Discussion

Target detection algorithm invention and optimization play a critical role in the swift advancement of autonomous driving technologies. This technology still faces the following problems, despite the fact that current techniques have significantly improved in terms of speed and accuracy of detection.

(1) Multi-domain target detection and diversified datasets. The current mainstream target detection models are mostly concentrated in specific fields, and the categories of most datasets are relatively single, which makes the model prone to missed detection or false detection when encountering unfamiliar scenes or unknown targets. Autonomous vehicles need to drive in various towns, cities, and weather situations in order to be used in driving applications. The target detection model must be able to accurately identify pedestrians, vehicles, traffic signs and other targets in various environments. Therefore, the development of a general detection model that can be applied across fields is crucial to improving the reliability and safety of autonomous driving systems. To solve this problem, future research will focus on creating more diversified datasets that cover more categories of targets and cover scenes in various complex environments. This will help train more general target detection models that can adapt to the needs of different fields.

(2) Small object detection. Target detection based on deep convolutional neural networks (CNNs) has become mainstream, but deep networks often face the problem of semantic loss when extracting features for small targets. This is because the features of small targets are easily over-compressed after multiple convolution and pooling operations, resulting in decreased detection accuracy. Small targets such as road signs, pedestrians or bicycles in the distance may be directly related to driving safety. Improving the detection ability of small targets can help autonomous vehicles detect potential dangers earlier and respond in time. To address this challenge, shallow semantic information can be combined with traditional image detection algorithms to better retain the features of small targets and improve detection performance.

(3) High-precision lightweight network architecture. Autonomous vehicles have limited computing resources, especially in actual driving, where real-time requirements are extremely high. By using a lightweight network architecture, the inference speed of the model can be improved without sacrificing accuracy, ensuring that the autonomous driving system can respond quickly in various situations. The existing target detection model architecture is usually complex and has many parameters, making it difficult to achieve real-time detection on resource-constrained edge devices.

Lightweight model design has therefore become an important development direction. In recent years, Google has proposed MobileNet, which significantly reduces the number of model parameters and computational complexity by replacing traditional convolution with deep separation convolution. In 2020, Huawei proposed GhostNet, which further reduces the computational overhead while maintaining model accuracy by optimizing the number of convolution channels.

(4) Video detection. The autonomous driving system needs to detect and track targets in real time while processing video streams. However, video detection faces challenges such as time series feature processing, redundant information of adjacent frames, and occlusion, resulting in computational redundancy and low detection accuracy. To this end, studying target detection algorithms based on video sequence data, optimizing the use of time series information, and reducing redundant features will become a research hotspot in the future. In dynamic driving scenarios, continuous video frames provide rich information that can be used to predict the motion trajectory and behavior of the target. By effectively utilizing this information, the autonomous driving system can better cope with complex driving environments, such as identifying and tracking other vehicles and pedestrians in dense traffic.

(5) Multimodal detection. Currently, most target detection models rely only on single-modal data, such as RGB images. This single-modal data may not be sufficient to provide sufficient information in some scenarios. For example, in low light or bad weather conditions, relying solely on RGB images may not accurately detect targets. Future research will focus on fusing data from multiple modalities (such as lidar images and 3D images) to improve the robustness and accuracy of detection.

(6) Insufficient oversight and limited detection. Large amounts of labeled data are frequently needed for object identification model training, and labeling data takes a lot of effort and time. Weak supervision and small sample detection systems have emerged to lower the cost of data labeling. Through efficient training with a small amount of labeled data, or using transfer learning technology to migrate labeled data from related fields, the detection effect of the model can be improved while reducing data requirements.

5. Conclusion

Thanks to the rapid development of deep learning technology, the accuracy and speed of object detection have achieved remarkable results. This paper extensively reviews the representative frameworks in the field of object detection research, and introduces representative detectors from four aspects: dual-stage, single-stage, keypoints, and Transformer, including their design ideas and basic processes. In addition, this paper also quantitatively compares the results of different methods on common datasets to analyze the advantages and disadvantages of various methods. Finally, this paper summarizes promising future directions in the hope of providing inspiration for subsequent research.

Authors Contribution

All the authers contributed equally and their names were listed in alphabetical order.

References

[1] Girshick R. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1440–1448.

[2] Ren SQ, He KM, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv: 1506.01497, 2015.

[3] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression [C]//Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, China, Aug 20- 24, 2006. Piscataway: IEEE, 2006: 850-855.

[4] QIN Y, HE S, ZHAO Y, et al. RoI pooling based fast multi domain convolutional neural networks for visual tracking [C]//Proceedings of the 2016 International Conference on Artificial Intelligence and Industrial Engineering, Beijing, Nov 20-21, 2016: 198-202.

[5] He KM, Gkioxari G, Dollár P, et al. Mask R-CNN. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2980–2988.

[6] J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. Proceedings of the 2016 Unified, real-time object detection. Proceedings of the 2016 Recognition (CVPR). Las Vegas: IEEE, 2016. 779–788.

[7] LI G, SONG Z, FU Q. A new method of image detection for small datasets under the framework of YOLO network [C]//Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference, Chongqing, Oct 12-14, 2018. Piscataway: IEEE, 2018: 1031- 1035.

[8] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C]//LNCS 9905:Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Oct 11-14, 2016. Cham: Springer, 2016: 21-37.

[9] CHEN H J, WANG Q Q, YANG G W, et al. SSD object detection algorithm with multi-scale convolution feature fusion [J]. Journal of Frontiers of Computer Science and Technology, 2019, 13 (6): 1049-1061.

[10] HOU Q S, XING J S. SSD object detection algorithm based on KL loss and Grad-CAM [J]. Acta Electronica Sinica, 2020, 48 (12): 2409-2416.

[11] LIN T Y, GOYAL P, GIRSHICK R, et al. focal loss for dense object detection [J]. arXiv:1708.02002, 2017.

[12] HO Y, WOOKEY S. The real-world-weight crossentropy loss function: modeling the costs of mislabeling [J]. IEEE Access, 2019, 8: 4806-4813.

[13] TAN M, PANG R, LE Q V. EfficientDet: scalable and efficient object detection [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 10781-10790.

[14] AO L, ZHANG X, PU J, et al. The field wheat count based on the EfficientDet algorithm [C]//Proceedings of the 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education, Dalian, Sep 27-29, 2020. Piscataway: IEEE, 2020: 557-561.

[15] FA Z W, YAN W Y, SHUIYUAN D, et al. Research on location of chinese handwritten signature based on EfficientDet[C]// Proceedings of the 2021 IEEE 4th International Conference on Big Data and Artificial Intelligence, Qingdao, Jul 2- 4, 2021. Piscataway: IEEE, 2021:192-198.

[16] MEKHALFI M L, NICOLÒ C, BAZI Y, et al. Contrasting YOLOv5, Transformer, and EfficientDet detectors for crop circle detection in desert [J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-15.

[17] Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey [J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[18] Wu X, Sahoo D, Hoi S C H. Recent advances in deep learning for object detection [J]. Neurocomputing, 2020, 396: 39-64.

[19] Law H, Deng J. Cornernet: Detecting objects as paired keypoints [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 734-750.

[20] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv preprint arXiv:1904.07850, 2019.

[21] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C]//European Conference on Computer Vision. Cham: Springer International Publishing,2020: 213-229.

[22] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks [C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 764-773.

[23] DAI Z G, CAI B L, LIN Y G, et al.UP-DETR: Unsupervised pre-training for object detection with transformers [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 1601-1610.