Research of object detection algorithms based on CNN

Jingzhe Jiang[1]**, Renkai Liu**[2]**, Xizhe Zhai**[3]**, Ziheng Zhao** [4, 5]

[1] Computer Sciences, Northeastern University at Qinhuangdao, Qinhuangdao, China
[2] Engineering Sciences, University College London, London, United Kingdom
[3] Overseas Education Institutes, Nanjing Tech University, Nanjing, China
[4] School of Automation, Beijing Institute of Technology, Beijing, China

[5]3574238546@qq.com

Abstract. This study first presents a taxonomy of region-based and regression-based target identification methods, and then it evaluates each individual technique in turn. The former can extract picture characteristics from the CNN backbone and apply sliding window procedures to candidate areas to lessen computational complexity. The latter immediately takes data from the convolution layer for feature extraction, target classification, and location regression rather than going via the intermediate layer to retrieve candidate areas. Although the accuracy is only marginally poor, this approach can significantly cut the computation time. Next, we categorize and evaluate the method for detecting Anchor-free targets. First, an appropriate illustration of the target identification technique based on intense prediction is shown. It is also clear that this method, when compared to others, has a high rate of calculation and a low rate of accuracy. The newly suggested corner pooling, Cascade corner pooling, and center pooling have significantly increased the calculation accuracy of the target identification method based on key point estimate. Finally, Transformer, which is very inclusive and powerful, can be effectively used in the field of natural language processing and computer vision.

Keywords: Deep learing, Convolutional neural network, Transformer

1. Introduction

Target detection technology, as a hot research content, has made breakthrough progress in unmanned driving, intelligent security, face recognition, and other fields in recent years. However, in physical truth, the detection effect of target features is affected by many factors, such as scale change, light intensity, local occlusion, etc. There are still some problems with the target detection algorithm, such as incomplete extraction and utilization of feature information, inconsistent classification confidence, and positioning accuracy of the target. This paper classifies the Anchor based Target algorithm, anchor free object detection algorithm and Object Detection with Transformer, and analyses their principles, working characteristics, adaptability, etc. one by one. Each chapter also introduces the basic principle of each algorithm and their respective upgrading process step by step. The limit level of the corresponding model is gradually reached under the fitting and optimization of many parties. For example, the area-based target detection algorithm in the Anchor based target algorithm fully illustrates that the basic principle of the model is to improve the speed and accuracy of model detection respectively through SPPNet and Faster RCNN. And in the following text, it will be compared with the past model to show its advantages and disadvantages. Through these introductions, readers can have a preliminary understanding of the neural networks related to target detection.

2. Anchor-based Target Algorithm

Generally, Anchor based target detection algorithms can be separated into two groups: regression-based and region-based. The region-based algorithm uses two networks to accomplish classification and regression respectively, so it is also called two-level detector, mainly including RCNN series algorithms. The algorithm based on regression is to complete the target classification and location in a network, so it is also called single-stage detector. It mainly includes YOLO series algorithms and SSD series algorithms.

2.1. Region based Target Detection Algorithm

Convolutional neural networks (CNN), which benefit from local connections and weight sharing, are at the heart of area extraction algorithms and exhibit good robustness in object categorization applications. The locale extraction activity starts by removing picture highlights utilizing the CNN spine. Then, at that point, recognizes potential closer view objects from the component map, lastly applies a sliding window activity to the up-and-comer districts to decide the objective classification and area data. This enormously diminishes the time intricacy of computation. Girshickproposed R-CNN (region with CNN characteristics) algorithm, and realized the utilization of depth learning when it comes to target detection for the first time [1]. It extracts a group of candidate regions through selective search, then uses CNN model which has been trained on Image Net to achieve feature extraction, and finally completes target prediction through support vector machine (SVM). Due to the existence of a huge number of superfluous features and potential boxes, the detection time of RCNN is very long.

In order to solve the problems above, He et al. introduced spatial pyramid pooling and proposed spatial pyramid pooling network (SPPNet) [2]. SPPNet allows you to input images of any size, and you can generate features of different scales with only one feature extraction. Although SPPNet's detection rate is quite quick., the training process is still staged. R. Girshick noticed the drawbacks of RCNN and SPPNet and proposed the Fast RCNN algorithm in 2015. Fast RCNN realizes the synchronous training of classification and bounding box regression, and improves the speed of identification and training [3]. However, Fast RCNN still has a problem of slow proposal region proposal. In order to solve the above problems, S Ren introduced Regional Proposal Network (RPN) and proposed Faster RCNN, which greatly improves the speed of regional proposal [4]. Although Fast RCNN has high detection accuracy, it can only predict on a single scale and has poor detection effect on small targets.

To address the issues mentioned above, in 2017, Kaiming He et al. added a parallel mask branch in Faster R-CNN-Fully Convolutional Networks for Semantic Segmentation (FCN) to generate a pixel level binary mask for each RoI. MaskR-CNN algorithm extends Faster R-CNN and it is suitable for pixel level fine grain image segmentation. In Mask R-CNN, bilinear interpolation is used to solve the problem that pixels cannot be aligned accurately, which is also called Rol Align. After using Rol Align instead of Rol Pooling, Mask R-CNN has achieved outstanding results and surpass Faster R-CNN in the field of target detection. Mask R-CNN model has strong flexibility and can be applied to multiple tasks such as target detection and target segmentation with a little change. However, due to the inheritance of Faster R-CNN's two-stage calculation method, its real-time performance is still not ideal. The above target recognition strategy in light of locale extraction initially creates the proposal box of the district of interest, and afterward orders and relapses the suggestion box. Albeit the discovery precision is improved, the location speed is predominantly sluggish, which isn't suitable for application situations with high ongoing necessities. Its presentation correlation is displayed in Table 1.

Table 1. Performance comparison of region based target detection algorithms.

algorithms	Backbone network	speed/（FPS）	mAP /%	advantages	Disadvantages
R-CNN	AlexNet VGG-16	0.03 0.50	58.5 66.0	CNN for feature extraction	Time consuming, memory consuming, fixed input size
SPP-NET	ZF-5	2.00	59.2	Multidimension convolution of the whole image	Large space cost
Fast R-CNN	VGG-16	7.00	70.0 68.4 19.7	Shared features of classification and regression,implement simultaneously	Select candidate areas and consume time and space
Faster R-CNN	VGG-16 ResNet-101	7.00 5.00	73.2 70.4 21.9 76.4 73.8 34.9	Use RPN to achieve end to end detection	model is complex, spatial quantification is rough, and the small target effect is bad
Mask R-CNN	ResNeXt-101	11.00	78.2 73.9 39.8	Solve the problem that the feature image is not aligned with the original image, and realize detection and segmentation at the same time	Instance splitting costs too much

2.2. Target Detection Algorithm based on Regression

A few specialists have proposed a worked-on calculation model to change target discovery into relapse, which all the while further develops identification precision and speed, to additional upgrade the continuous presentation of target recognition. The regression-based target detection algorithm does not use the middle layer to extract candidate regions; rather, it uses a reverse calculation to obtain target location and category and performs feature extraction, target classification, and position regression across the entire convolution network. Despite the fact that the recognition accuracy is slightly lower than that of the two-stage target detection algorithm, the speed has been greatly improved, making the target detection algorithm based on depth learning useful for a variety of tasks that require rapid reasoning. The YOLO model proposed by R. Joseph et al. in 2015 is the first one-stage target detection algorithm after introducing deep learning, but YOLO actually belongs to the category of non anchored target detection algorithm [5].

R. Josep et al. proposed YOLOv2. YOLOv2 use convolutional kernels for feature compression and global average pooling for prediction, and a batch normalization layer is introduced after each convolution layer [6]. Tackle the issue of angle vanishing and slope blast during the time spent back spread. However, YOLOv2 network has poor detection ability for small targets, and the detection accuracy is not high enough.

In terms of the above problems, YOLOv3 is proposed. YOLOv3 uses deeper DarkNet-53 to extract more fine-grained feature information, uses FPN structure to achieve multi-scale prediction, and uses 1 × 1 Convolution and Logistic activation functions replace the Softmax classification layer to perform data fitting more efficiently [7]. YOLOv3 has made great progress in different kinds of functions, but the detection accuracy and real-time performance are still lacking. Regarding to the lack of real-time performance and high training cost of most target detection algorithms, Alexey et al. proposed YOLOv4, which uses advanced CSP Arknet53 for feature extraction, and uses the SPP+PANet module to further enhance the expression ability of features. At the same time, various newly proposed Tricks are applied to the improvement of the network, making YOLOv4 still obtain better real-time performance and detection accuracy on the basis of reducing training costs.

Liu et al. put forward the SSD (single shot multibox detection) model. After VGG-16, SSD added multiple convolution layers to obtain multi-scale feature maps for prediction, and used Faster RCNN's anchor frame concept to set a prior frame with different aspect ratios for feature maps of different scales to better detect objects of different sizes, reduce training difficulty, and have better detection effect on overlapping areas or objects closer to each other. However, SSD has many repeated frames and is not robust to small target detection.

The one phase relapse-based target location calculation begins later than the two phase locale based target recognition calculation, which enjoys the benefit of newbie and can all the more likely ingest the upsides of the previous and beat its disadvantages. Albeit the mid one phase target location calculation has a quick recognition speed, there is as yet a huge hole in discovery precision contrasted and the two phase identification calculation. With the quick advancement of target discovery innovation, the speed and exactness of target identification models at the ongoing stage have been incredibly moved along. The presentation correlation is displayed in Table 2.

Anchor based target detection algorithm has always occupied a dominant position in the field of target detection. From the original RCNN to the latest YOLOv4, the number of parameters of the algorithm is decreasing, and the detection speed and accuracy are improving.

Table 2. Performance comparison of region-based target detection algorithms.

algorithm	Backbone network	Speed/(FPS)	mAP /%	advantages	Disadvantages
YOLO v1	GoogLeNet	45.0	63.4 57.9	Divided into fixed grids,speed fast	Low positioning accuracy
YOLO v2	Darknet-19	40.0	78.6 73.5 21.6	The anchor box is introduced to improve the speed and accuracy	Calculation of Pool Layer Multi influence Gradient
SSD	VGG-16	19.3	79.8 78.5 28.8	Fusion of multi-layer convolution features to improve the detection accuracy of large and medium-sized targets	Difficult convergence and limited improvement of small target detection accuracy
YOLO v3	Darknet-53	51.0	33.0	Convolution replaces pooling, and residual module is introduced	Complex model, poor detection effect of medium and large scale targets
YOLO v4	CSPDarknet-53	23.0	43.5	Comprehensively improve training skills and improve training accuracy and speed	Data enhancement and anchor box setting to be optimized

3. Anchor-free Object Detection Algorithm

The Anchor-free target detection algorithm can be classified as density-based prediction and key-point estimation.

3.1. Target Detection Algorithm based on Dense Prediction

The target detection algorithm based on dense prediction uses pixel by pixel prediction to complete the detection. FCOS (fully convolutional One-stage) object detection algorithm framework, which uses pixel-by-pixel prediction method (Figure 1).

1. Multilevel prediction through FPN to improve recall and solve ambiguity caused by overlapping bounding boxes.

2. Low quality prediction boxes are suppressed through centre-ness branch, which greatly improves the performance.

The network structure of FCOS is shown below. Obviously, it contains the following three parts,

(1) Backbone network;

(2) Feature Pyramid structure;

(3) the output section (classification/Regression/Center - ness is);

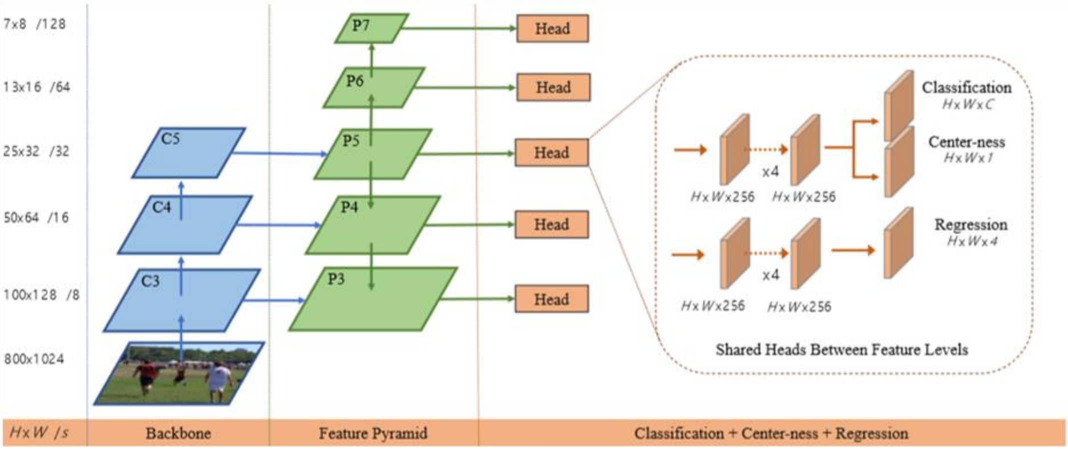

Fig. 1. Structure of FCOS.

FCOS adopts FPN architecture and adds a center-ness branch to exclude prediction boxes that are far away from the target Center. P6 and P7 are obtained by convolution with step size 2 after P5, instead of from C5, so as to obtain stronger feature meaning and reduce the number of parameters. FCOS has lightweight structure, fast detection speed and high detection accuracy.

The baseline of the FSAF improvement is RetinaNet (Figure 2), to which a branch of Anchor free has been added. In each stage of FPN structure, a prediction head in Anchor free output format is added to the original classification and box regression prediction head in parallel, and the number of channels is adjusted by a 3×3 convolution. In the training process, the scale loss value of different branches is calculated for each sample, and then the scale branch of the sample is determined according to the size of the loss value, so that the sample can automatically learn and select the optimal scale branch for training. On the basis of fast detection speed, FSAF further improves the accuracy of multi-scale detection.

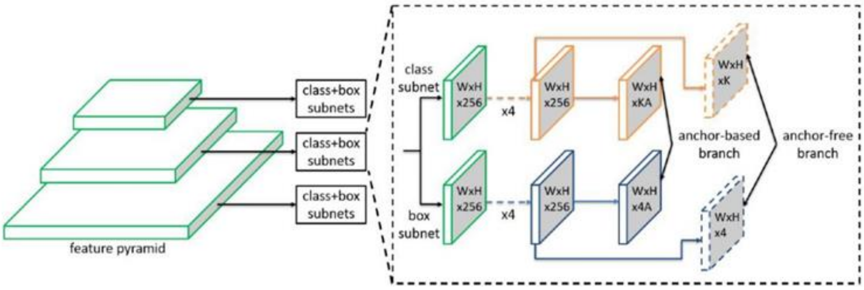

Fig. 2. Structure of FSAF.

Inspired by the structure of human eyes, Kong et al. paid more attention to the central area of the target in the training and detection process, and then corrected the central area by the offset to obtain the final prediction box. Foveabox defines positive and negative samples by shrinking and expanding the points on the real box. Different feature layers detect the target of the corresponding scale, and solve the problem of prediction overlap by controlling the scale coefficient. Foveabox has the advantages of small number of parameters and fast detection speed, but the detection accuracy is not high enough.

3.2. Target Detection Algorithm based on Key-point Estimation

The target detection algorithm based on key-point estimation realizes the target detection by estimating the corner point or center point. CornerNet eliminates the requirement of anchor frame in the existing detection algorithm and reduces the training requirements of the entire detection network. Researchers can select and design different feature extraction networks. Meanwhile, in order to better locate the corner points, a new corner pooling method is also proposed, which effectively improves the detection accuracy.

In addition, CenterNet also proposed Cascade corner pooling and Center pooling, which effectively improved the generation of each key point, thus improving the detection effect. The detection accuracy of CenterNet has been effectively enhanced, but the inaccurate matching of key points in dense target is still a problem. CentripetalNet, after generating candidate corner points, introduces the centripetal offset method to match the diagonal points, and the centripetal offset alignment, to obtain high quality corner points; Then, a new Cross-star deformable convolution module was used to learn the offset field by the offset from the corner point to the center point, and the feature adaptation was carried out to enhance the visual features of the corner position, so as to enhance the accuracy of the centroid migration module. Finally, a segmentation mask module is added, which can be directly applied to the instance segmentation task after simple improvement. CentripetalNet enables end-to-end training with high detection accuracy.

The standard RPN in Two-stage algorithm cannot infer the likelihood of target-background well, while One-stage algorithm has a good effect. Therefore, CenterNet2 by probability explanation is derived (Figure 3), the above two kinds of fusion algorithm, detection algorithm design more efficient two stage, among them, the first stage is the goal of using single stage detector forecast the unknown categories likelihood probability, the second stage is to use two levels of second part of the detector conditions are classified, in order to obtain accurate category of conditional probability, Finally, the probability scores of the two stages are united to acquire the final prediction result. CenterNet2 optimizes the network from a relatively new idea, which further improves the detection accuracy of the algorithm.

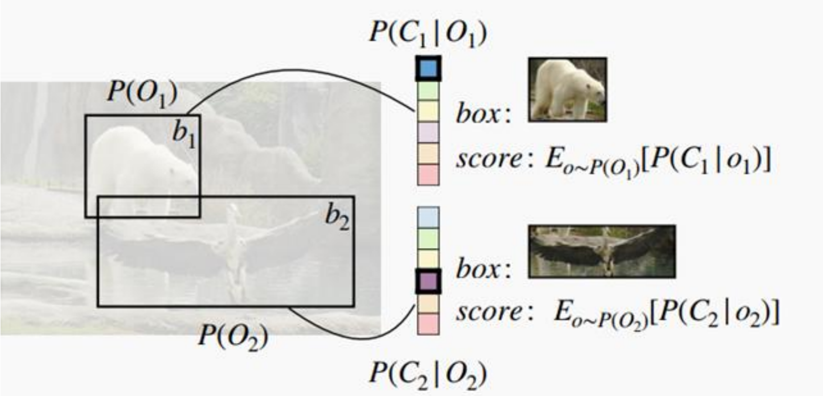
$P(C_1 | O_1)$
$P(O_1)$
b_1
box:
score: $E_{o \sim P(O_1)}[P(C_1 | o_1)]$
b_2
box:
score: $E_{o \sim P(O_2)}[P(C_2 | o_2)]$
$P(O_2)$
$P(C_2 | O_2)$

Fig. 3. Structure of CenterNet2.

4. Object Detection with Transformer

Transformer is a new CNN model proposed by Google in June 2017. It has got great results with low calculation amount and high efficiency of parallel computation in the fields of natural language processing (NLP). Meanwhile, in the fields of computer vision (CV), there have also been many studies on the applications of transformer in object detection, such as DETR and its derivative models.

DEtection TRansformer (DETR) is the first object detection model based on transformer which was proposed by a team of Facebook in 2020 [8]. Its structure is shown in Figure 4. DETR uses the image features extracted from CNN network to produce a one-dimensional set of features. Before entering the encoder-decoder converter, the features are then encoded at predefined places to maintain the feature map shape. The decoder generates the same number of outputs by simultaneously decoding several objects at each decoding layer using a multi-head attention method. Finally, feedforward neural networks are used to perform bounding-box regression and picture classification (FFNs). In contrast to RNN, DETR approaches object detection as a direct set prediction task. Specifically, to forecast the full set simultaneously without being influenced by earlier outcomes (Figure 4). Thanks to the characteristics of end-to-end detection, the encoder-decoder framework in DETR is more consistent with the paradigm of object detection in comparison with traditional anchor-based algorithms. Moreover, it can avoid non-maximum suppression (NMS) to remove redundant boxes. According to an experiment based on a popular object detection dataset, COCO, DETR performed comparably to the competitive Faster R-CNN, and more importantly, its performance on large objects is significantly better than Faster R-CNN. As a result, DETR is suitable to be extended to some complex tasks like panoptic segmentation.

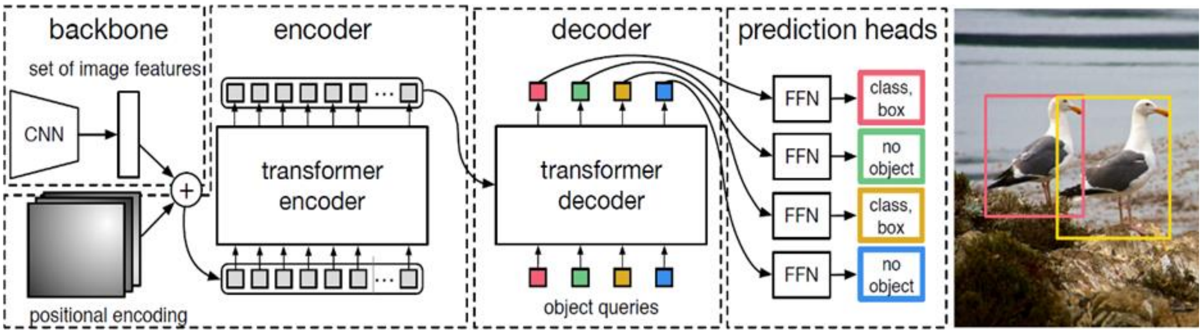

Fig. 4. Structure of DETR.

As the attention modules in traditional DETR need to apply nearly the same number of weights to all the pixel in the high-resolution feature maps, which requires long training epochs and high computational complexity, Zhu et al. proposed D-DETR (Deformable-DETR), which uses deformable attention module to replace multi-head attention mechanism in DETR [9]. Since the deformable attention module only samples a small number of key points around a reference point, its computational complexity is effectively reduced and the convergence of training is accelerated. Besides, the deficiency in detecting small objects can be mitigated by its strong ability to fuse multi-scale features. As is shown in Table 3, compared to the original DETR algorithm, the number of epochs for training D-DETR is 10 times smaller than that for training the original DETR, which greatly reduces the training time on GPU, and the speed of detection is 1.6 times higher compared to DETR-DC5.

Table 3. Comparison of DETR with D-DETR on COCO 2017 dataset.

Method	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	Training GPU hours	Inference FPS
DETR-DC5	500	43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	7000	12
Deformable DETR	50	43.8	62.6	47.7	26.4	47.1	58.0	40M	173G	325	19

Adaptive clustering transformer (ACT), a new enhanced model developed by Zheng et al., can lower pre-computing training's cost without requiring any training [10]. ACT uses locality sensitive hashing

(LSH) to cluster features adaptively, and uses prototype key interactions to cluster near query key interactions. It replaces the self-attention module of the pre-trained DETR model and saves the training process of 500 epochs in the original DETR. This approach can effectively reduce the calculation cost with only a small amount of accuracy loss. Furthermore, multi-task knowledge distillation (MTKD) is used to mitigate the degradation of performance: Using DETR to extract the ACT module and fine-tuning to further improve the performance of ACT. Thus, a better trade-off between performance and computational cost can be achieved.

Using transformer in object detection seems to be a good idea as it performs better than other models in some aspects. However, new designs are required to address the problems of long training epochs and low performance in the detection of small objects.

5. Conclusion

Target detection has significant research value in both the academic and practical worlds. Target detection technology built on deep learning has made several advancements thanks to the ongoing development and innovation of the field. This paper presents the collection of existing techniques through the research of the target identification algorithm, including the Anchor-free target detection algorithm, which can address the issue of slow training speed. The most traditional depth learning-based target detection algorithms include the Faster RCNN, YOLO V3, and SSD methods. Although faster RCNN is unquestionably the most accurate algorithm, it is slower than the other two. Although SSD is not slow or inaccurate, it is not suitable for small target identification. The YOLO V3 industrial detection system is quick and simple to operate.

Many new algorithms are being developed right now, but most of them are just enhanced versions of older ones. Although target recognition has advanced quickly, some issues remain difficult to resolve: Consider the detection of small target objects. There is no algorithm with a good detection effect for small target objects due to the poor resolution and limited amount of information on small target objects in the image. In addition, issues like obscured objects, target detection with interference from the surroundings, etc. continue to exist. When target detection is used in practical situations, it can be difficult to find clear, comprehensive data sets with high quality and quantity to satisfy the needs of training models. The target detection technology's direction for future development is: 1) How to lower energy usage in both production and daily living; 2) How to apply integration and customisation to industries with various demands; 3) The goal of target detection technology is to create an optimal balance between detection accuracy and speed in real-world applications.

Reference

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 *IEEE Com. Soc. Conf. Com. Vis. Pat. Rec.* 580-587.

[2] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. 2014, *Euro. Conf. Com. Vis.* 346-361.

[3] Girshick R B. Fast R-CNN 2015 *IEEE Inter. Conf. Com. Vis.* 1440-1448.

[4] Ren S Q, He K M, Girshick R, et al, Faster R-CNN: Towards real-time object detection with region proposal networks Annual Conference on Neural Information Processing System. 2015 *NIPS,* 91-99.

[5] Redmon J, Divvala S, Girshickr, et al. You only look once: unified, real-time object detection 2016 *IEEE Com. Soc. Conf. Com. Vis. Pat. Rec.* 779-788.

[6] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger, 2017 *IEEE Com. Soc. Conf. Com. Vis. Pat. Rec.*: 6517-6525.

[7] YOLOv3: An Incremental Improvement, 2018 *arXiv preprint arXiv*: 1804.02767.

[8] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. 2020 *Euro. Conf. Com. Vis.* 213-229.

[9] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection 2020 *arXiv preprint arXiv:* 2010.04159.

[10] Zheng M, Gao P, Wang X, et al. End-to-end object detection with adaptive clustering transformer, 2021 arXiv preprint arXiv: 2011.09315.