Research Advanced in the Object Detection Based on Deep Learning

Banghao Chi
School of Advanced Technology
Xi'an Jiaotong-liverpool University
Wuhan, Hubei Province, China
Banghao.Chi21@student.xjtlu.edu.cn

2022 International Conference on Applied Physics and Computing (ICAPC) | 978-1-6654-6303-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAPC57304.2022.00092

***Abstract*—Object detection is one of the most basic tasks in the field of Computer vision, which targets to localize and allocate a wide range of predefined substances from images to their corresponding classification. Thanks to the rapid progress of deep learning, object detection algorithms based on convolutional neural networks have been applied in different fields, and have achieved breakthroughs both in accuracy and efficiency compared to traditional detection schemes. In this paper, based on detailed literature research and analysis, a comprehensive evaluation of object detection research advances are provided and specifically, we divide existing representative algorithms into three main frameworks, including traditional detection algorithms, anchor-based and anchor-free detection algorithms. We then conduct a series of experiments to analyze the achievement of different detection algorithms on some common datasets. Finally, we summarize the main challenges and provide an outlook on the future research directions of object detection.**

Keywords—Object detection, Deep Learning, Anchor-based detection, Anchor-free detection

I. INTRODUCTION

Object detection[1] requires identifying and locating one or more objects in its vision (e.g. cars, pedestrians, road signs, etc.), which forms the fundamentals of computer vision together with other classical tasks like classification, segmentation, motion estimation and scene comprehension. However, while infants as young as a few months can recognize some common objects, it wasn't until a decade ago that machines struggled to learn object detection and gradually matured. At present, object detection algorithms have been successfully applied in various fields and achieved exciting results, such as autonomous driving, industrial detection, smart agriculture, etc.

In the 20-year development process, according to the differences in design ideas, conservational methods and detection methods based on deep learning are the two basic types of target detection algorithms. Normally, traditional target detection algorithms mainly include key steps such as preprocessing, window sliding, feature extraction, feature selection, feature classification and post-processing. Among them, window size, sliding method and strategy exert a great effect on the quality of feature extraction. Deformable part model DPM while its extended models are often used to discriminate sliding windows, like scale invariant feature transform (SIFT) and histogram of oriented gradient (HOG)[2], the efficiency and accuracy of the entire detection process are low. With deep learning's rapid advancement, algorithms for detection have veered from conventional methods to more advanced techniques based on deep neural networks (DNN). Deep learning-based approaches currently combine feature extraction, feature selection, and feature classification into a single model to accomplish end-to-end performance and efficiency optimization, and have gradually become the mainstream framework for object detection.

According to whether generate the candidate regions and how to generate the candidate regions, the existing object detection framework based on deep learning can be further split into two-stage, one-stage and anchor-free object detection algorithms. Specifically, for a detection network, assuming that there is a dedicated module responsible for generating region proposals, the network is defined as an anchor-based detector. Anchor-based detectors further include two-stage detectors and one-stage detectors. Aiming to extract a certain number of object proposals in the first step and then localize and classify them in the second step are what two-stage detectors try to accomplish. Two-stage detectors usually take more time to extract all proposals, which generally have complex structures and lack global information, while the one-stage detectors directly identify and locate objects through dense sampling. Predefined boxes of different scales and sizes are adopted to localize objects. Different from anchor-based detection methods, anchor-free methods don't use anchors and corresponding encoding information to represent detection boxes. According to the difference of bounding box expression, anchor-free methods can be divided into two categories: key-point based detection algorithms and center-based detection algorithms. The former first detects the upper left and lower right corners of the target, and then forms a detection frame by combining the corners. The latter detects the object's center area and border information directly, and decouples the classification and regression into two sub-grids.

Focusing on the three main technical frameworks of the detection methods introduced above, this essay analyzes and sums up the research advance and status quo of deep learning-based target detection algorithms. Secondly, we also introduce the general datasets of object detection and the experimental results of different algorithms on mainstream datasets in detail, and look forward to the field of object detection's potential future development.

II. TRADITIONAL DETECTION METHODS

The traditional detection method includes three core segments, i.e., region selection, feature extraction and classification. The first step's goal is to localize the object. Since the position, size and scale of the object are not known, the sliding window strategy is used to traverse the entire image. Although this strategy covers all possible positions of

978-1-6654-6303-4/22/$31.00 ©2022 IEEE
DOI 10.1109/ICAPC57304.2022.00092

the object, its corresponding disadvantage is also transparent. It needs to manually set a series of windows of different sizes and proportions, resulting in redundant windows, which seriously affects the speed and power. Next steps for feature extraction and classification. More importantly, the extraction of object features (using SIFT, HOG, etc.) is quite complicated due to the diversity of morphology, light changes and backgrounds, which also directly affects the accuracy of classification. The representative traditional object detection methods are as follows:

A. Viola-Jones

Proposed in 2001, Viola-Jones[3] was an accurate and powerful detector which at that time was primarily used to verify identification. It combined a great number of technologies such as Haar feature, integral image, AdaBoost, cascade classifier and so on. The initial stage is to look for Haar features by sliding a window across the input image, then calculating using the integral image. Subsequently, it employs a well-trained AdaBoost to identify the classifier of each Haar feature and cascade them, and since Viola-Jones is of great efficiency, it is still used on small devices in the modern time.

B. HOG

Aiming to further enhance the process of extracting features, Dalal and Triggs proposed Histogram of Oriented Gradients (HOG) in 2005, and compared to other detectors, HOG modified the process by extracting the gradient and its edge direction to form a feature table, dividing the image into grids which HOG use to create a histogram for each unit in it, generating the HOG features for Region of Interest (RoI) and inputting the features into the linear SVM classifier for detection. Although it is initially proposed as a pedestrian detection detector, it can also be taught to recognize a variety of other classifications.

C. DPM

Deformable Parts Model (DPM)[4], which was the champion of 2009 Pascal VOC challenge, was proposed by Felzenszwalb and his colleagues. By means of detecting the partial ‘segment’ of the object, its accuracy surpassed the which of HOG while conformed the philosophy of divide and rule. By excluding the impossible combination to generate the ultimate detection, the models based on DMP is the most effective and powerful algorithm prior to the era of deep learning. The upgraded HOG feature, SVM classifier, and Sliding Windows detection notion are all adopted in the DPM. The multi-component strategy is used to solve the target's multi-perspective challenge, and for the deformation problem of the target itself, a multi-component strategy is applied and put into practice. Pictorial Structure is prepared with Component model strategy, and last but not least, Latent Variables include the model category to which the sample belongs, the component model's position, and so on, and Multiple-instance Learning is used to automatically determine.

III. ANCHOR-BASED DETECTORS

The definition of Anchor is easy to understand. Generating regions of different size and scale while regarding every point as the center point, these regions are called Anchor. Anchor-based detectors conform the rule of two points. On one hand it extracts the RoI by sliding windows which are all predefined (which is also called Anchor). On the other hand, it classifies and regresses the Anchor of each point.

A. Two-stage detectors

The concept of two-stage detectors is intelligible, i.e., it first uses anchor to identify the foreground and the background. Then it regresses and classifies the RoI and outputs the ultimate region and its corresponding classification. And among so many of the two stage detectors, R-CNN is one of the most typical examples that can be further illustrated.

Region-based Convolutional Network (R-CNN) is a classic algorithm for object detection. As Figure 1 shown, every image is extracted about 2 thousand region proposals using Selective Search (SS)[5], which locates the region that have a high possibility to possess an object. After that, each of them is warped to the size as the convolutional network requires and is regarded as the input to the convolutional network while the corresponding output is seen as the feature of the region. Third step involves training multiple Support Vector Machines (known as SVM)[6] by using these features to identify objects. Every SVM identify whether there exists a particular object in a certain field. Last but not least, these regional features are further used to train linear regressor to adjust the location of the region. The main difference of the traditional methods and R-CNN is that the latter one applies feature extraction method of deep learning classification model in place of the old algorithm of feature extraction. And the core logic of R-CNN selects several regions of an image, then each of them goes into a convolutional neural network to extract the feature. Nonetheless, pioneer as R-CNN was at that time, it still confronted three significant drawbacks. The individual of region proposals should all be calculated its features by CNN which results in massive amount of calculation. Secondly, the quality of the region proposals is not good enough. Moreover, feature extraction and SVM classifier is trained independently, which is in great need of systematically optimized to avoid its character of time-consuming.

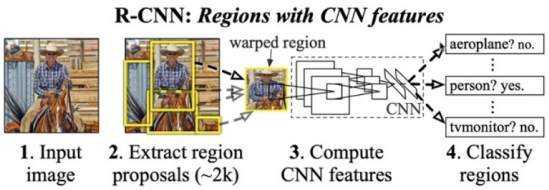

Fig. 1. Framework of R-CNN algorithm

The proposal of the Fast R-CNN (Fast region-based Convolutional Network)[7] is mainly to diminish the time consumed in the process of extracting vector features from region proposals using CNN model. Compared to R-CNN which send every single region proposal into CNN model, Fast R-CNN inputs the full image and combines the methods of RoIs (Region of Interests) pooling and SS to extract the features from the feature maps generated by backbone. Through RoI pooling layers, it obtains a feature map with a fixed length and width. Another immense change between R-CNN and Fast R-CNN is that the latter one uses softmax classifier instead of SVM and since it is single-piped, it can combine the error of classification and positioning together for training, using smooth L1 instead of L2 in R-CNN. Fast R-CNN is chiefly introduced for improve the speed of training and predicting (which is 146 times faster than R-

CNN), and as for accuracy, it is secondary. We can draw the conclusion that Fast R-CNN is way better than R-CNN from Figure 2. However, due to its continuing use of SS, it still costs plenty of time to obtain the region proposals.

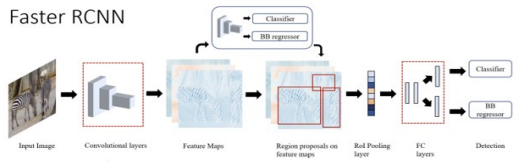

Fig. 2. Framework of Faster R-CNN algorithm

To deal with the complication of the time-wasting process of SS, the Faster Region-based Convolutional Network (Faster R-CNN)[8] was proposed along with the introduction of Region Proposal Network (RPN) to generate the region proposals directly. Faster R-CNN can be thought of as the mix of RPN and Fast R-CNN model. Take a closer look at the Faster R-CNN (See Figure 2), it can be noted that in the first step, RPN introduces the method of Anchor to replace SS in order to speed up the procedure of region proposals selection and extracts the features of Anchor which includes objects in RoI Pooling. Afterwards, it classifies the region proposals and predicts the location of the object. If we want to know the exact structure of RPN, we can first look into an example. For instance, in the last layer of the initial CNN, a 3*3 sliding window moves to the center which is one of the points on the feature map (the output of backbone and input of RPN) and then maps it to lower dimensions (256-d). Then RPN generates several areas on a basis of k fixed scale anchor boxes. The lower dimensions are divided into 2 segments while each region proposal includes 2k score which marks the softmax possibility of the 'object' and 4 coordinates which represents the location of the 'object'. On PASCAL VOC 2007, Faster R-CNN achieved 69.9% mAP, compared to Fast R-CNN with 66.9% mAP at 5 FPS. To some extent, it had reached a balance between accuracy and efficiency simultaneously. Despite the fact that Faster R-CNN seems a little bit complex, the core logic of it is identical to the original R-CNN: predict the location and classify the object.

Proposed by Lin and his teammates, Feature Pyramid Networks (FPN)[9] is a common treat to increase the accuracy when it comes to detect small object. FPN adopts a horizontally connected structure up to down to construct high-level semantic features on various scales and possesses two paths. One is the bottom-up path of feature level calculated by convolutional neural network (ConvNet) on multiple scales, and the other is the top-down path, by which high-resolution features are sampled from a coarse feature map at a higher level. In addition, to improve the semantic information in the feature, these routes are joined horizontally using a 1x1 convolution technique. Under the circumstance of this, the RPN in the Faster R-CNN is taken over by FPN, using ResNet-101 as the backbone. FPN can present provision of high-level semantics on scales and diminish the error rate of detection. It has become an exemplification of the future detection and improved the overall accuracy, which also promotes the strides of modified networks, such as PANet[10], NAS-FPN[11], EfficientNet etc.

Mask R-CNN[12] makes an expansion on the basis of Faster R-CNN by creating a branch for pixel-level object instance segmentation in parallel. It uses a structure similar to the one which Faster R-CNN possesses to extract region proposals, adding a mask head parallel to classification and regression head. Another core idea of Mask R-CNN is to replace RoIPooling with RoIAlign to get a better effect of locating. And for higher accuracy and speed, they chose ResNeXt-101 with FPN as its backbone. Mask R-CNN is now more powerful as compared to the current SOTA one stage framework by adding an extra function of segmentation while the concomitant cost is little. It reached the top when it came to detect the COCO database for both instance segmentation and object detection and presented great universality in detection of key points and the estimation of human pose. Unfortunately, its frame rate is still lower than the real-time one (> 30 FPS).

B. One-stage detectors

One-stage detectors' mission is simple, extracting the features and classifying while locating. Instead of generating region proposals, it directly accomplish the mission. The ideology of one-stage detectors was first introduced by YOLO. You Only Look Once (YOLO)[13], was proposed by Ross Girshick after the emergence of R-CNN, Fast R-CNN and Faster R-CNN. The previous methods invariably generate huge numbers of frames that may contain the object. Then the classifier identifies whether the frames contain the object or not and predict the possibility while adjusting its frame and delete the frames with high-overlapping to get the final region. Although it is more accurate, it is rather time-consuming as well. Under such circumstance, YOLO innovatively regard the assignment of object detection as a regression problem, combining two stages of region proposals and detection into one stage (See Figure 3). At a mere sight of the image, YOLO knows what objects are in the image and the position of them. But actually, YOLO doesn't remove the concept of region proposals. It adopts a method called predefined region proposals, that is to say, it divides the image into 7*7 grids as the inputs and every grid predicts 2 boxes as a total of 49*2 bounding boxes, which are called 98 region proposals roughly covering the whole image. As a result, YOLO makes a vast expansion on efficiency at the cost of reducing mAP.

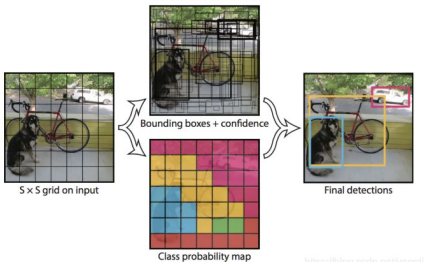

Fig. 3. Framework of YOLO algorithm

YOLOv2[14] is modified based on YOLO, reaching a balance between speed and accuracy. YOLOv2 replace GoogleNet which is the backbone of YOLO with DarkNet-19 and can predict 9000 classes of objects in real time. Compared with original YOLO algorithm, for accuracy, certain sorts of training techniques are used; for speed, new backbone is adopted; for classification, the jointing training method is adopted, combined with wordtree and other methods, contributing to the astonishing expansion of

detection types to more than 9000. At 67 FPS, YOLOv2 achieves 76.8 mAP on the VOC 2007 database. At frame rates of 40, it achieves a score of 78.6 mAP, more remarkable than Faster R-CNN and other pioneer methods which use ResNet while maintaining at a relatively fast speed.

The improvement made by YOLOv3[15] is not all contributive: some of them are positive while others are negative. Among these tentative changes, two of them are worth illustrating, one is the residual model and the other is using FPN to accomplish multi-scale detection. In terms of the backbone, YOLOv3 adopts DarkNet-53 rather than Darknet-19, which introduce residual module that deepen the network. From YOLOv1, YOLOv2, YOLO9000 to YOLOv3, YOLO series maintained its advantage of speed and improve its network structure simultaneously, absorbing other leading trick like anchor box mechanism and the introduction of FPN. YOLOv3 consider both accuracy and speed, and when detecting COCO datasets, its mAP is identical to SSD's with a triple speed; its mAP is slightly lower than that of RetinaNet but the speed of YOLOv3 is 3.8 times the speed of RetinaNet.

The proposal of YOLOv4[16] is a breakthrough in the YOLO series. YOLOv4 combines many effective and useful ideas, creating an object detector that is rather easy to train within a minor amount of time in the existing system. Adopting the 'Bag of Freebies' approach, only the time for training increases while having no effects on the inference time. YOLOv4 combines the techniques and methods of Data Augmentation[17], Label Smoothing[18], CIoU-loss, Cross mini-Batch Normalization (CmBN), Self-adversarial-training (SAT)[19] and Cosine Annealing[20] to improve the process of training. Another method which merely influence the inference time is also introduced, called 'Bag of Specials', including Mish Activation Function[21], Cross-stage partial connections (CSP), SPP-Block, Path Aggregation Network (PAN)[22], Multi-input weighted residual connections (MiWRC) while uses Genetic Algorithm to accomplish hyper-parameters search. In addition, YOLOv4 is composed of several parts: CSPNetDarkNet-53 as its backbone, SPP and PAN blocks as its neck and the same head as YOLOv3. Most detection algorithms need more than a single GPU to steel their models, but YOLOv4 is capable of doing the same thing on one GPU with ease. With the similar performance as EfficientDet, its efficiency doubles, reaching SOTA in the mean time.

IV. ANCHOR-FREE DETECTORS

Although anchor-based detectors have attained breakthroughs in both accuracy and speed of object detection, they still have the following limitations. (1) The settings of Anchor need to be manually designed (aspect ratio, size and scale as well as the number of Anchor), and particular datasets need corresponding settings, which is rather troublesome. (2) Ancho's matching method devastatingly lower the frequencies that object at an extreme scale is detected. It is hard for Deep Neural Networks (DNN) to learn from these samples. (3) The enormous number of Anchors accounts for an imbalance when sampling and every single of them needs IoU calculation, leading to tremendous calculation and reducing the efficiency.

Then how can object be located and classified if there is no anchor to express the bounding box? There are two sorts of anchor-free algorithms: keypoint-based and center-based. The top-left corner and bottom-right corner of the item are detected by keypoint-based algorithms, which then combine the corner points to build a detection box. Center-based algorithms detect the center of the object, dividing classification and regression into two sub-grids.

A. Keypoint-based detection methods

Among so many Anchor-free detectors, CornerNet[23] is the symbolization of milestone between Anchor-based and Anchor-free detectors, which is also one of the algorithms that are keypoint-based. CornerNet presents a provision of a brand-new way to detect objects. The author of CornerNet coincidentally found that the kernel of the logic to handle tasks such as human pose estimation and segmentation is adding labels to the dots to form high-level semantics. In light of this, he thought that finding the frame of the object was equal to detecting the corners of the frame and combining them. One thing should be noted is that keypoint-based algorithms are easier to be trained compare to the center-based ones, which is also another reason why the author chose keypoint. For instance, the top-left point is only related to two sides while the center is pertinent to four sides. The process of locating the object is further explained as follow. First, it uses bottom to extract features from the image, then two individual prediction modules are introduced to locate two corners respectively. In every prediction module, there exist two branches: Heatmaps is responsible for locating the point and classifying it and Embeddings' duty is to match the points that are both the corners of the same object. Offsets are the error between the original coordinates and the ones that Heatmaps get. For distinct objects, there can be a huge variation in their Embeddings and for the same object, the Embeddings are quite similar.

The accuracy of CornerNet is magnificent, similar to the mAP of Anchor-based detectors. However, it can make mistakes when detecting objects of the same classification and since it adopts Hourglass-104 as its backbone, the speed of detecting still has a lot of space for improvement. CornerNet-Squeeze[24], is a modified version of CornerNet. It reduces the amount of process of every pixel and combines the idea of SqueezeNet[25] and MobileNet. The performance of CornerNet-Squeeze is remarkable, surpassing one stage detector YOLOv3 no matter in the aspect of accuracy and efficiency.

B. Center-based detection methods

Center-based detectors are another type of Anchor-free algorithms and there are two typical detectors that are most reputed, one of which is FCOS. Fully Convolutional One-stage Object Detection (FCOS) makes regression for every position on the feature map using FCN, which means FCOS treats every single point as a training sample, getting similar effects as Anchor-based methods do. The reason why FCOS uses FCN is that it is hard to tell a certain point belongs to which one if two objects overlap in the image. Therefore, multi-level prediction with FCN is adopted, which can effectively solve the problem. It can be obtained that there are three branches in FCOS, classification, regression and center-ness. The branch of classification treats pixels as training samples using Focal loss; the branch of regression the regression regards the distance of a point to four sides as object value to be trained; the branch of center-ness assesses

the distance between pixel point and the center. It achieves a score of 42.1 in mAP, greater than the performance of CornerNet.

Another typical center-based detector is CenterNet[26]. The core idea of CenterNet is concise, which composes of the center pluses the scale. On one hand, it uses an approach similar to one-stage method, i.e., every center can be seen as an Anchor without the information of the shape. On the other hand, it adopts similar predicting method, using Heatmap along with Focal loss to be trained and replacing NMS which is rather time-consuming. It also possesses a branch similar to offset, which compensates for the error of the position of the sample point. Its architecture is rather simple and it has a great flexibility and expansion in other visual tasks. Though, the time for training is at an enormous amount and the regression supervised information is merely generated by the location of the center. Thus, a modified version of CenterNet, TTFNet was proposed.

TTFNet[27] mainly changes three places as compared to CenterNet. Firstly, ellipse Gaussian kernel is adopted to generate negative sample signals to better locate the center and obtain the regression branch supervised information. Moreover, it removes the offset branch, predicting the distance using regression branch. TTFNet proposes a new method to locate the center and it reduces the time for training in an immense degree while maintaining powerful performances synchronously.

V. Experiments

A. DatsSet

This branch contains two common datasets, which are usually used to measure the performances of different detectors.

(1) PASCAL VOC[28]. Pascal Visual Object Classes (VOC) is a challenge which supply the detectors with a standard dataset of images. There are two major versions of it, one is VOC2007 which possesses 5,000 images and more than 12,000 labeled objects, the other is VOC2012 which acquires 1,1000 images, more than 27,000 labeled objects and 20 types of objects, adding tasks of semantic segmentation and action recognition. It introduces mAP@0.5IoU as an evaluating indicator to assess the performance of the model.

(2) MS-COCO[29]. The Microsoft Common Objects in Context (MS-COCO) is one of the most intricate datasets, including 91 types of common objects discovered in the nature and easily identified by four-year-old child. Proposed in 2015. It now possesses over 2000,000 numbers of images with every single of them in 3.5 sorts including multiple perspectives and introduce a more accurate method to assess the detector, calculating its mAP every 0.5 ranging from 0.5 to 0.95 and even classifying its AP into small, medium-sized and large objects to present its accomplishment from various aspects.

TABLE I. Commonly used datasets for face recognition

Method	backbone	AP	AP50	AP75	APS	APM	APL
R-CNN	AlextNet	/	/	/	/	/	/
Fast R-CNN	VGG-16	/	/	/	/	/	/
Faster R-CNN wTDM	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	52.1
Mask R-CNN	ResNeXt-101-FPN	39.8	62.3	43.4	22.1	43.2	51.2
Faster R-CNN wFPN	ResNeXt-101	36.2	59.1	39.0	18.2	39.0	48.2
YOLO	(Modified)GoogLeNet	/	/	/	/	/	/
YOLOv2	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
YOLOv3	DarkNet-53	33.0	57.9	34.4	18.3	25.4	41.9
YOLOv4	CSPDarknet-53	43.5	**65.7**	47.3	26.7	46.7	53.3
CornerNet	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
CornerNet-squeeze	Hourglass-54	34.4	/	/	13.7	36.5	41.9
FCOS	ResNeXt-64x4d-101-FPN	44.7	63.1	48.4	27.6	47.5	55.6
CenterNet	Hourglass-104	**47.0**	64.5	**50.7**	**28.9**	**49.9**	**58.9**
TTFNet	DarkNet-53	39.3	56.8	42.5	20.6	43.3	54.3

B. Performance analysis

As presented in Table I, we report various results on the COCO dataset and several conclusions can be drawn as follow:

(1) It can be apparently drawn from the table that Anchor-free algorithms, all of them expect CornerNet-squeeze present powerful performance when detecting medium and large objects. When it comes to small object detection, YOLOv4, FCOS and CenterNet are excellent.

(2) Detectors with latest backbone such as ResNeXt-101, CSPDarknet-53 or Hourglass-104 invariably acquire more accuracy as compared to old-fashioned backbone like DarkNet-19 and DarkNet-53.

(3) Anchor-free algorithms have pervasive advantages from all aspects as compared to Anchor-based algorithms.

(4) CenterNet can be seen as the best detectors from all aspects while the changes of TTFNet reduces the time for training, followed by FCOS and YOLOv4. Therefore, we can draw the conclusion that CenterNet outperform the field of object detection, dwarfing all the other detectors to reach the top.

VI. Discussion

As a significant part of computer vision, object detection has been developing at a tremendous speed to catch up with the rapidly changing world with the assistance of deep learning. However, there are still some problems that we should care for, like small object detection.

(1) How to increase the accuracy of small object detection has become a hotspot nowadays and the process of improving it struggles along the road. Mainstream two stage detectors such as Faster R-CNN is typically used in small object detection, however, since the backbone (a tool that can extract the features from an image) of Faster R-CNN has a top-down structure whose deep and shallow feature maps do not achieve a satisfactory balance in semantics and space, the performance seems mediocre. As for one stage mainstream detectors, for instance SSD[30], despite the usage of a multi-

layer feature map, the semantic information of shallow feature map is insufficient and feature fusion is not performed, resulting in poor tiny item recognition performance.

(2) Low resolution, blurred image and less concomitant information are all examples of the obstacles that is blocking the way to the accomplishment of the flawless object detection. As a result, the ability of feature expression is weak, meaning that in the process of feature extraction, very few features can be extracted, which is not conducive to the detection of small objects. Moreover, due to its small size, the available features of small objects are limited, which contributes to its detection more sophistication. Contemporarily, the kernel problem of small object detection based on deep learning is how to enhance the feature expression of small object, making it contain prolific semantic information, which is also the key to reinforce the achievement of small object detection. Mainstream detection algorithms are not friendly to small object detection, therefore, the modified version comes into vision.

(3) The introduction of FPN is a miracle treat to increase the accuracy when it comes to detect small object. The architecture of it coincidentally meets the need, which adopts a horizontally connected structure up to down to form semantic features of high-level on different scales. And provided that it is carried out ubiquitously, this can bring about myriad of great practical value and application prospect ranging from aviation, automatic driving to industrial automation and satellite remote sensing images. For example, it’s unavoidable that there will be small objects on the airport runway, such as nuts, screws, washers, nails and fuses. Accurately detecting these small objects on the runway can avoid devastating aviation accidents and immense economic losses. For automatic driving, it is rather necessary to accurately detect small objects like distant signals or remote objects at a tremendous speed from high-resolution scene photos of cars to avoid traffic incidents. For industrial automation, small object detection is also in urging need to locate small defects visible on the material surface. For satellite remote sensing images, the targets in the them, such as cars and ships, may only have dozens or even a few pixels, thus accurate detection of small object in this field will help the government agencies curb drug and human trafficking, finding illegal fishing vessels and better enforcing the prohibition of illegal transshipment of goods. Henceforth, the accomplishment of small object detection possesses pervasive application value and important research significance from an enormous number of perspectives.

(4) There are also other future trends for our society. 3D object detection in real time is a significant issue in autonomous driving since the performance can sometimes be lower than that of human. Object detection in video also attracts lots of attention, which induce the relationships of the images from the perspectives of time and space. The urging need for lightweight detectors indicates the demand for small, effective, using fewer sources and powerful models in the same time.

VII. Conclusion

The main concerns and important technical hurdles in the domain of target detection are discussed in this essay, along with the introduction of the main technical framework and representative algorithms in the discipline of object detection research, chiefly including conventional target detection approaches, anchor-based and anchor-free Detection method. An additional contrast is made between the experimental results of common datasets and related algorithms on mainstream datasets, and predict the future development direction of this research field while summarizing the major challenges in the realm of target detection.

References

[1] N. Zheng, G. Loizou, X. Jiang, X. Lan, and X. Li, "Computer vision and pattern recognition," ed: Taylor & Francis, 2007.

[2] T. Lindeberg, "Scale invariant feature transform," 2012.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 2005, vol. 1: Ieee, pp. 886-893.

[4] Y.-Q. Wang, "An analysis of the Viola-Jones face detection algorithm," Image Processing On Line, vol. 4, pp. 128-148, 2014.

[5] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in 2008 IEEE conference on computer vision and pattern recognition, 2008: Ieee, pp. 1-8.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, pp. 142-158, 2015.

[7] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," International journal of computer vision, vol. 104, no. 2, pp. 154-171, 2013.

[8] I. Steinwart and A. Christmann, Support vector machines. Springer Science & Business Media, 2008.

[9] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440-1448.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[12] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9197-9206.

[13] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7036-7045.

[14] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International conference on machine learning, 2019: PMLR, pp. 6105-6114.

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961-2969.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[17] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263-7271.

[18] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[20] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of big data, vol. 6, no. 1, pp. 1-48, 2019.

[21] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," Advances in neural information processing systems,

vol. 32, 2019.

[22] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018: IEEE, pp. 17-30.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101, 2017.

[24] D. Misra, "Mish: A self regularized non-monotonic activation function," arXiv preprint arXiv:1908.08681, 2019.

[25] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759-8768.

[26] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 734-750.

[27] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "Cornernet-lite: Efficient keypoint based object detection," arXiv preprint arXiv:1904.08900, 2019.

[28] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," arXiv preprint arXiv:1602.07360, 2016.

[29] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9627-9636.

[30] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6569-6578.