EAI Endorsed Transactions
on e-Learning
Research Article EAI.EU

Review on One-Stage Object Detection Based on Deep Learning

Hang Zhang[1,*], Rayan S Cloutier[2]

[1]School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, P R China
[2]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

Abstract

As a popular research direction in computer vision, deep learning technology has promoted breakthroughs in the field of object detection. In recent years, the combination of object detection and the Internet of Things (IoT) has been widely used in the fields of face recognition, pedestrian detection, unmanned driving, and customs detection. With the development of object detection, two different detection algorithms, one-stage, and two-stage have gradually formed. This paper mainly introduces the one-stage object detection algorithm. Firstly, the development process of the convolutional neural network is briefly reviewed, Then, the current mainstream one-stage object detection model is summarized. Based on YOLOv1, it is continuously optimized, and the improvements and shortcomings are summarized in detail. Finally, a summary is made based on the difficulties and challenges of one-stage object detection algorithms.

Keywords: Object Detection, IoT, Deep learning, Computer Vision.

Received on 20 April 2022, accepted on 08 June 2022, published on 09 June 2022

doi: 10.4108/eai.9-6-2022.174181

*Corresponding author. Email: zh@home.hpu.edu.cn

1. Introduction

Object detection is one of the most fundamental and challenging tasks in computer vision. It not only needs to perform image classification on the categories existing in a picture, but also needs to accurately locate objects that may exist in a picture, where classification refers to matching the correct category label, and positioning refers to finding out the corresponding picture frame position. Therefore, the process of object detection is more difficult and more promising. At present, it is closely related to the development of the Internet of Things (IoT), which has been highly recognized by the society in the fields of video surveillance and intelligent transportation [1-3].

According to whether the candidate frame area needs to be generated in advance, the object detection algorithm is divided into two-stage and one-stage detection algorithms. The two-stage algorithm is represented by R-CNN [4], also known as the object detection algorithm based on candidate regions. Simply speaking, first, generates candidate regions in the image, and then performs classification and regression processing on each candidate region [5-7]. The one-stage algorithm is represented by YOLOv1 [8], also known as the regression-based object detection algorithm. It means that the input image is no longer processed by the candidate area, and the object in the image is directly located and classified. In general, the two algorithms have their advantages. The two-stage algorithm has high accuracy, but it takes a long time to pass through the selective search algorithm during detection [9, 10]. Conversely, one-stage algorithms are faster but less accurate. The object detection algorithm based on deep learning is shown in **Figure 1**.

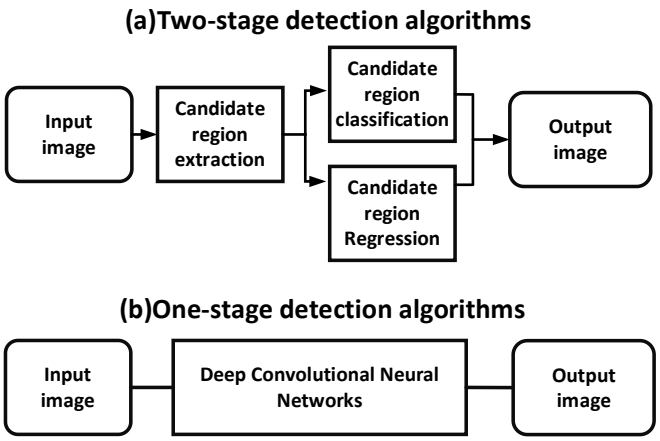

Figure 1 Different Object detection algorithm

The (a) shows the two-stage object algorithm flow, which has a separate candidate region extraction module, not an end-to-end operation. For example, the (b) network structure can be found to be an end-to-end network, and the input pictures can be output directly through the neural network.

In recent years, with the continuous improvement of the YOLO series, not only has the training speed of the one-stage algorithm been improved but also many innovative algorithms and architectures have also been proposed. This paper mainly describes the development process of the one-stage object detection algorithm and conducts an in-depth analysis of the module structure [11] in the development process. Finally, the comparison between the one-stage object detection algorithm and the two-stage object detection algorithm is made, and the existing problems in this field are pointed out.

2. Convolutional Neural Network

The Convolutional Neural Network (CNN) is the most representative model of deep learning. It is composed of the input layer, convolution layer, pooling layer, and full connection layer [12-16]. Most of the current networks are based on a series of improvements made by CNN.

Originally, in 1998, LeCun [17] proposed the LeNet network for handwritten digit recognition and applied CNN to the field of image recognition. As an early neural network, LeNet only contains three full connection layers, two convolution layers, and two pooling layers. Because the model is small, it cannot fit other data well, which limits the development in computer vision fields [18, 19].

In 2012, Krizhevsky proposed the AlexNet network and won the championship in the ILSVRC2012 image classification task, which caused a strong learning upsurge in the field of computer vision. Many researchers [20-23] have also applied it to the object detection task, constructing R-CNN, OverFeat [24], MultiGrasp [25], and other classical object detection algorithms. They applied deep learning to large-scale image classification for the first time and achieved the best results.

In 2013, ZFNet [26] made minor adjustments to the AlexNet network, mainly introducing a new visualization technology. In the past, CNN was a black box; there was no corresponding theory or method to explain the optimization and improvement process of the network. ZFNet shows the visualization of the intermediate feature layer through deconvolution [27,28]. They won the ILSVRC championship [29].

In 2014, Simonyan [30] proposed the VGG model, which studies the effect of network depth on accuracy. Unlike AlexNet, VGG uses multiple stacked 3x3-sized convolution layers to replace large-size filters. The advantage of the model is that the structure is simple and effective, and it can be well migrated to other networks, but the disadvantage is that the parameters are too large and easy to fit. Scholars have used VGG in many fields successfully [31-33].

GoogLeNet [34] is the 2014 ImageNet champion, and the network not only studies the impact of depth but also takes into account the breadth of the network. The network removes the last full connection layer and skillfully puts forward the 1x1 convolution operation to reduce the dimension and avoid the over-fitting problem caused by too large network parameters.

In 2015, He *et al.* [35] proposed the ResNet residual network and residual connection. It mainly solves the problems of network degradation caused by increasing the depth or width of the network, and solves the problem of gradient disappearance through residual connection, so that the depth of the network can reach 152 layers. The network uses a small amount of pooling layer and a large number of downsampling, which improves the forward propagation efficiency of the network and achieves the best image recognition effect at that time, which proves the feasibility of residual connection [36-38].

In 2017, Liu et al. proposed DenseNet [39], which won the best paper award at CVPR2017. Drawing on the ResNet network's method of deepening the depth and width of the network can also ensure the accuracy of the model. DenseNet constructed a typing network. One layer of information is concatenated (dimensionally connected) with all the other layers. DenseNet can effectively reduce the number of parameters and enhance the reusability of features between different convolutional layers [40-42].

It is because of the strong feature representation ability of convolution neural networks in deep learning that classical feature extraction networks such as VGG [30], GoogLeNet [34], and ResNet [35] are produced. They can do an excellent job of image extraction. It has been found that they can be used not only for image classification tasks but also for backbone architectures in more complex object detection tasks [43-47].

In 2014, Girshick *et al.* proposed a two-stage object detection algorithm, R-CNN [4], instead of the traditional manual feature selection DPM [48] algorithm, and finally got good results, but it is time-consuming. Therefore, the next part of this paper shows that the single-stage object detection algorithm not only has high speed but also has better accuracy.

3. One-Stage Object Detection Model

3.1. YOLOv1

YOLOv1 was proposed in 2016 and published on CVPR. YOLO is the first one-stage object detection algorithm that achieves good results in both accuracy and speed. The network structure is improved based on GoogLeNet, in which the inception layer is replaced by a 1x1 or 3x3 convolution operation. The core idea is to regard object detection as a regression task.

The algorithm flow is very simple and straightforward: divide the picture into an s×s grid, and each grid cell is only responsible for predicting the object where the central point falls in the grid [49]. At the same time, we also need to predict the **b** bounding box. Each bounding box contains (x, y, w, h) and confidence, as well as the category information N of the specified data set. Then each bounding box needs to predict (4+1+N) dimensional information. The final dimension size generated by YOLOv1 is (s×s, **b**×(4+1+N)).

But once it is set to b = 2, only two rectangular boxes are generated for each square, and finally, a rectangular box with greater confidence is selected as the output. That is, only one object can be predicted for each square in the end. The YOLOv1 model architecture is shown in

Figure 2.

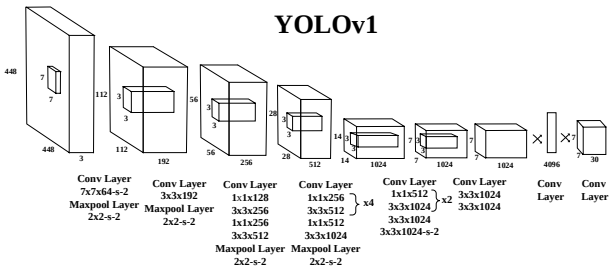

Figure 2 The YOLOv1 model

YOLOv1 does not extract candidate regions, so the detection speed is greatly improved. It can achieve 45FPS in the VOC2007 data set and can reach 63.4 maps. At the same time, YOLOv1 has strong migration ability and can be applied to other new fields (such as flower object detection combined with the IoT). Because each grid can only predict one category, when there are multiple classes in a grid at the same time, all categories cannot be detected, which is not better for small population detection. Moreover, due to the setting of the loss function, there are differences in the processing of large and small objects, and the final detection accuracy is not good enough.

The formula of YOLOv1 object loss function is shown in(1).

$$
\begin{aligned}
&\lambda_{cord}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]\\
&+\lambda_{cord}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2]\\
&+\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{obj}(C_i-\hat{C}_i)^2\\
&+\lambda_{noobj}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{noobj}(C_i-\hat{C}_i)^2\\
&+\sum_{i=0}^{s^2}1_{i}^{obj}\sum_{c\in classes}(P_i(c)-\hat{P}_i(c))^2
\end{aligned}
\tag{1}
$$

It can be seen from the above that the loss function is composed of three parts. They are confidence loss, class loss, and object loss. The function used to calculate the loss is the sum-squared error. The reason for the square root of w and h is that when calculating the IoU loss, when the prediction boxes of different sizes have the same offset, it is obvious that the IoU of the larger prediction box is larger, which results in particularly poor detection results for small objects. Adding to reduce the impact of w h, which makes the model pay more attention to small prediction boxes.

In addition, the $\hat{C}$i value is 1 in 1_{ij}^{obj}, which means that the prediction frame contains objects, and the $\hat{C}$i value is 0 in 1_{ij}^{noobj}, which means that the prediction frame does not contain objects. The category loss in the last line also uses the mean square error and only performs category prediction for positive samples. $\boldsymbol{\lambda}$ represent the weight balance factor, where $\boldsymbol{\lambda_{cord}}$=5 means that the coordinate loss occupies a larger weight so that the model pays attention to the regression loss. In the early stages of training, many cells generate many low-quality boxes. To reduce the model's learning of these low-quality boxes. Assigning $\boldsymbol{\lambda_{noobj}}$=0.5 reduces the loss of confidence in predictions that do not contain object boxes.

3.2. YOLOv2

In 2017, Redmon *et al*. proposed YOLOv2 [50] based on YOLOv1, focusing on solving the problems of recall rate and positioning accuracy faced by YOLOv1. The test speed is 67FPS on the VOC2007 data set, and the accuracy can reach 78.6% mAP. Specifically, YOLOv1 uses the final full connection layer to predict the bounding box, while YOLOv2 draws lessons from the idea of Faster R-CNN [51] and introduces the anchor mechanism, which can generate a priori frame in advance and use the K-means clustering method to generate an a better anchor template. Among them, YOLOv1 has only 98 bounding boxes, while YOLOv2 can reach more than 1000 bounding boxes [52-54], which is nearly 10 times more than the bounding box, which can significantly improve the recall rate of the algorithm.

The network combines the fine-grained features of the image, which is to fuse the feature maps of different sizes

through a certain technology, which can combine the high-resolution shallow texture features and low-resolution deep semantic features to improve the detection ability of small-sized targets. Different from YOLOv1, this algorithm designs a new full convolution feature extraction network Darknet-19 as the backbone, which includes 19 convolution layers and 5 maximum pool layers. For each layer of convolution, batch normalization is added for preprocessing. It can be seen that YOLOv2 summarizes many deep learning techniques and finally has a high improvement in accuracy and speed. The Passthrough layer used for fine-grained features is shown in

Figure 3.

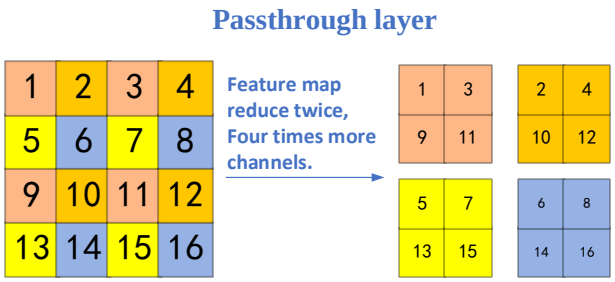

Figure 3 The Passthrough layer

3.3. YOLOv3

In 2018, the author Redmon made a further improvement based on YOLOv2 and proposed YOLOv3 [55]. Using the residual structure of ResNet as a reference, it is proposed that DarkNet-53 make the backbone network deeper (from DarkNet-19 in YOLOv2 to DarkNet-53 in YOLOv3, which is comparable to ResNet-101, ResNet-152 in accuracy and faster in speed) [56, 57]. At that time, it was one of the most classical and popular algorithms for achieving the best tradeoff between accuracy and speed.

Specifically, multiple logical regression classifiers are used instead of softmax classifiers to achieve multi-label classification (in YOLOv2, the algorithm can only determine that the current object belongs to one category, but in some complex scenarios, the object label has the problem of multi-class labeling.

For example, in a fruit transaction scenario, an object belongs to both an apple and a fruit. If softmax is used for classification, the results are mutually exclusive. That is, if it belongs to an apple, it is no longer a fruit, which is not true in some specific data sets and belongs to a single-label classification. If the final output of the network determines that the goal is both apple and fruit, this is the so-called multi-tag classification).

In addition, the feature pyramid network (FPN) architecture is introduced to sample the deepest feature map of the network twice. Combined with the output of the shallow network, different anchors are set on the final three feature maps to predict the object areas of different sizes.

The coordinate prediction method of the bounding box is similar to that of YOLOv2, in which the center point of the bounding box is predicted relative to the coordinates of the upper left corner of the grid (C_x,C_y), and each bounding box is predicted to get five values (t_x, t_y, t_w, t_h) and t_o. At the same time, to limit the center point of the bounding box to the grid, the Sigmoid function δ is used to normalize the (t_x, t_y), and the value is constrained between 0 and 1, and the final prediction result is still within the size of the grid. The stability of the early training of the model is significantly improved, and the coordinate prediction mode of YOLOv3 is shown in

Figure 4.

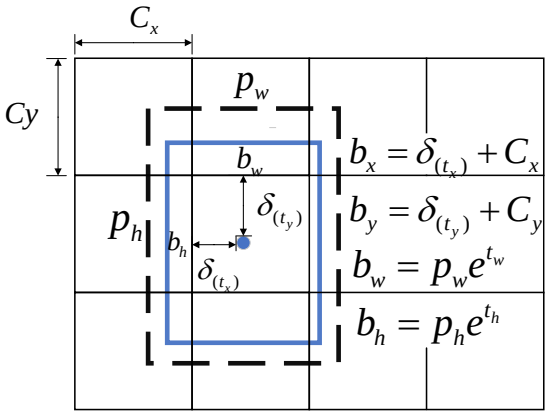

Figure 4 The coordinate prediction mode of YOLOv3

The traditional image pyramid is to extract features from different feature layers. It mainly uses artificial extraction features, and cannot combine the information from the upper and lower feature layers. Since each layer makes predictions, this approach increases the training data in disguise, making the algorithm time-consuming.

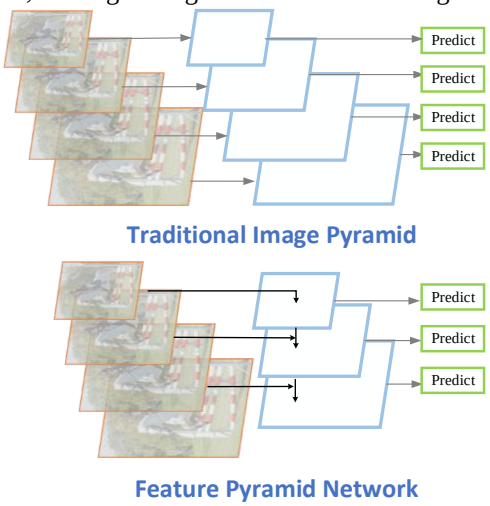

Figure 5 Traditional image pyramid and FPN

When the feature pyramid network (FPN) is used [58], multi-scale information can be used, and the upsampling from the high-resolution feature layer can be combined with the feature layer information of the original scale. To make up for the loss of information caused by multi-scale changes, experiments have shown that the network can

learn stronger feature information through the feature pyramid structure [59-61]. A comparison of the traditional pyramid and FPN is shown in **Figure 5**.

3.4. YOLOv4

In 2020, Alexey Bochkovskiy *et al*. put forward YOLOv4 [62]. The real-time monitoring speed in the MSCOCO data set reaches 65FPS and the accuracy reaches 43.5%AP.

The improvement to YOLOv4 is that the backbone network is CSPDarknet53. The SPP (improved structure inspired by SPP-net [63] and PANet [64] modules are used, and the activation function of the backbone is changed to the Mish activation function. In addition, the SPP module is added to the neck part, which can significantly increase the receptive field of the feature graph, effectively combine the network characteristics of the context, and will not reduce the running speed of YOLOv4. The activation function formulas and images of Mish and LeakyReLu are shown in **Figure 6**.

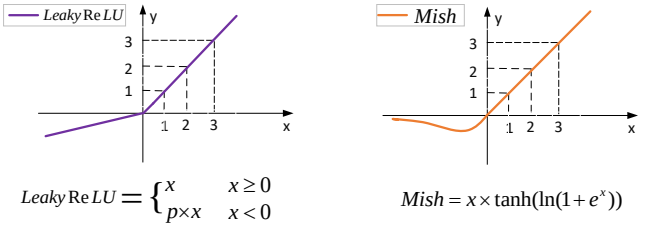

Figure 6 Mish and LeakyReLu formulas and images

The literature [62] uses CSPDarknet53 as the backbone network, and the design inspiration comes from the CSPNet [65], proposed by Chen-Yao Wang *et al*. The CSPNet network proposes an innovative structure from the perspective of network structure design to solve the problem of information redundancy when the gradient is returned and updated, and to reduce the amount of network computation while ensuring that the accuracy does not drop.

The literature [65] found that adopting this new structure can not only enhance the feature learning ability of the backbone extraction network but also reduce the computational cost. The module comparison between CSPDarknet53 and Darknet53 is shown in **Figure 7**.

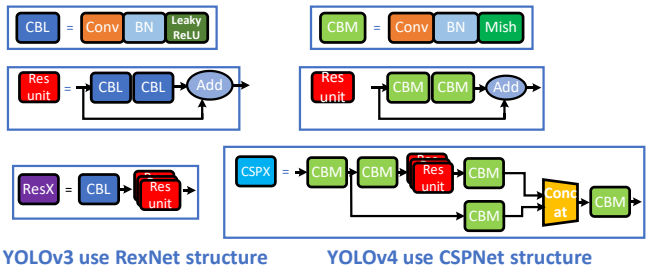

Figure 7 The CSPDarknet53 and Darknet53 module

3.5. YOLOx

The YOLO series is constantly optimizing its speed and accuracy. In recent years, some people have questioned whether YOLO can still be improved. In 2021, Liu *et al.* published YOLOx [66], which is similar to the YOLOv5 model but also uses different network structures such as YOLOx-s, YOLOx-m, YOLOx-l, YOLOx-x, and so on, besides designs a YOLOX-Nano, YOLO-Tiny lightweight network to realize the dynamic selection model according to demand. YOLOx has a simple structure, users can quickly deploy the model architecture, and it has strong flexibility.

Literature [66] takes into account that YOLOv4, v5 may over-optimize anchor, so a series of improvements have been made under the condition of YOLOv3-SPP. Taking YOLOx-DarkNet53 as an example, there are mainly the following points: not only Mosaic data enhancement but also MixUp data enhancement is used on the input side, and Decoupled Head, anchor free, Multi positives, and other improvement measures are adopted in the prediction module.

The Decoupled Head specifically embodies the splitting of a single output of the original network into three different outputs. The regression parameters of category, confidence and bounding box prediction box were corresponding, respectively. The detection heads used in the original YOLO series may lack the expression ability and the network optimization ability. Using Decoupled Head, the AP value increases from 38.5 to 39.6, and it is found that not only is the accuracy improved, but also the convergence speed of the network is accelerated. The schematic diagram of the Decoupled Head structure is shown in **Figure 8**.

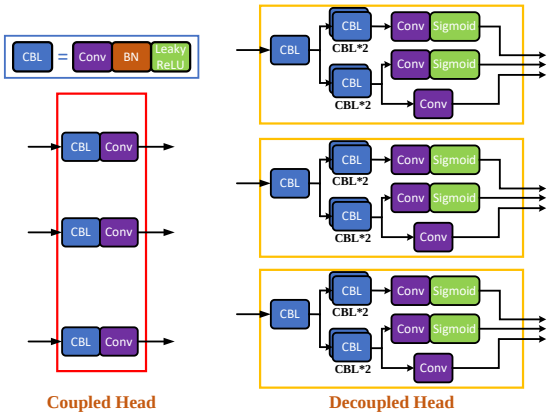

Figure 8 The Decoupled Head and coupled head

3.6. SSD

Liu W *et al*. put forward the SSD [67] algorithm in 2016, which is mainly based on the improvement of YOLO location inaccuracy, insufficient accuracy, and low recall rate at that time.

The improvement of the SSD algorithm mainly has the following points: Feature fusion is carried out for feature extraction of different sizes, which improves the

robustness of network training and enables the learning of more deep contexts; instead of using the operation of YOLO to predict the object after the full connection layer [61, 68, 69], CNN is added to the backbone network to predict directly. Combined with the anchor mechanism in Faster R-CNN, the candidate regions are obtained by using different prior boxes, and the recall rate is improved. But the disadvantage is that the accuracy of the model for small object detection is not high, and the positive and negative samples are extremely uneven. The schematic diagram of the SSD algorithm is shown in **Figure 9**.

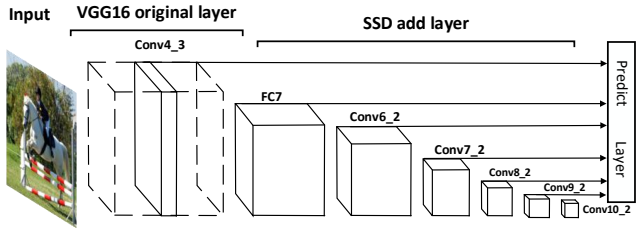

Figure 9 The SSD algorithm

3.7. RetinaNet

In 2018, Lin *et al*. proposed RetinaNet [70] and published it in ICCV2017. They believe that the fundamental reason why the accuracy of the regression-based object detection algorithm (one-stage) is lower than that of the candidate region-based object detection algorithm (two-stage) is the serious imbalance between positive and negative samples in the single-stage algorithm [71-73]. The high accuracy of the two-stage algorithm is due to the existence of region proposal network (RPN) network extraction to filter out a lot of useless background frames and alleviate the problem of category imbalance.

At that time, the one-stage algorithm directly generated the candidate regions in each grid and predicted the regression directly in the results, which contained a large number of redundant candidate boxes, which undoubtedly added great difficulty to the fine classification of network training. Therefore, although the detection speed of the one-stage object detection algorithm is fast, the accuracy is not ideal.

The Focal loss function is proposed in the literature to solve the problem of mismatch between positive and negative samples. Through this method, the proportion of the weight of the samples that are easy to distinguish is low, so the network mainly trains those samples that are difficult to distinguish, towards the correct optimization direction. The Focal loss mathematical formula is shown(2).

$$FL(p) = \begin{cases} -\alpha(1-p)^{y}\log(p) & if\ y=1 \\ -(1-\alpha)p^{y}\log(1-p) & otherwise \end{cases} \quad (2)$$

In the above formula, p represents the probability that the model is a positive sample. In order to solve the problem of sample imbalance, the balance factor α ranges from 0 to 1. The introduction of y makes the range of indistinguishable samples larger and more obvious so that the Focal loss can focus on training indistinguishable samples.

3.8. CornerNet

At present, most object detection networks include anchors for regression operations; the final candidate box is screened out, and good results have been achieved. However, the introduction of the anchor mechanism leads to some problems, such as uneven positive and negative samples, poor training in the early stages of the model, and the difficulty of decreasing the loss function.

In addition, because the initial size of the anchor is clustered and screened by the K-means algorithm in advance (different sample distributions of data sets will generate different anchor shapes), these fixed anchors are not suitable for other object detection tasks, which means that they cannot be well migrated to other model tasks. In addition, the size, aspect ratio, and the number of anchors are very sensitive to the detection performance. By adjusting the parameters, the model can improve the AP by nearly 4%.

So in 2018, Law *et al.* proposed CornerNet [74] in ECCV2018. This object detection algorithm is anchor-free and detects objects according to a pair of key points (upper left and lower right coordinates). A new network structure called the CornerNet is proposed, which includes an hourglass network (usually used in attitude estimation tasks) and a new pooling method, Corner Pooling, and finally produces three different outputs: heatmaps, embeddings, and offsets.

The advantages of the model are that there is no anchor box, fast detection speed, and high accuracy, which solves the problems of sample imbalance and adjusting super-parameters caused by the anchor box. The disadvantage is that the information inside the bounding box is not taken into account, and the detection accuracy of complex small objects or multi-object groups is poor. The CornerNet structure diagram is shown in **Figure 10**.

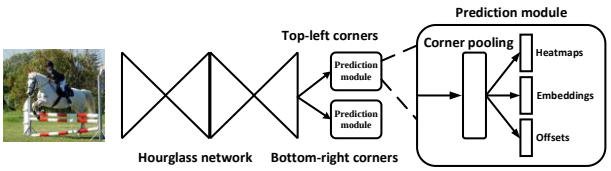

Figure 10 The CornerNet structure

As shown in the final three outputs, heatmaps output the predicted vertices information and are responsible for predicting the position of corners. In the training phase, the corner positions in the area with ground-truth as the radius are set as positive samples [75]. Since heatmaps generate a lot of corner information, how to determine which two points belong to an object, Embedding is useful because it is responsible for minimizing the distance between the two corners of the same object. In the previous object

detection, the offset represents the offset information between the predicted box and the real box, and the offset output by CornerNet represents the accuracy loss information generated during calculation.

3.9. FCOS

In 2019, Chun-Hua Shen *et al*. published FCOS [76], a brand-new pixel-level-based single-stage object detection algorithm that surpassed the state-of-the-art single-stage models at the time.

Overall, FCOS adopts the popular anchor-free algorithm, which reduces the amount of computation and eliminates the influence of unstable network prediction structures caused by adjusting the anchor hyperparameters. At the same time, it is combined with FPN to assign objects of different sizes to different feature layers, which enables FCOS to detect various object overlaps, crowding occlusion [77], small object detection, and other problems, and improves the recall rate.

Since anchor is not used, how does FCOS define positive and negative samples? It generates (x, y) coordinates by mapping each point of the feature map back to the original image size. If the position (x, y) is in the ground-truth, it is considered a positive sample, otherwise it is considered a negative sample. In addition (l^*, t^*, r^*, b^*) is defined, that is, the distance from this point to the left, top, and right, and bottom of the object frame. The specific formula is as follows(3)(4).

$$l^* = x - x_0^{(i)}, \quad t^* = y - y_0^{(i)} \tag{3}$$

$$r^* = x_1^{(i)} - x, \quad b^* = y_1^{(i)} - y \tag{4}$$

Among them, (x0, y0), and (x1, y1) represent the upper left corner and lower right corner information of the ground-truth box.

The YOLO series only selects one bounding box in the corresponding grid cell to participate in the loss function calculation, while FCOS selects many boxes as positive samples, which can speed up the regression. However, since most of the positive samples are low-quality detection frames far from the center point, the loss cannot be reduced. The literature [76] proposes center-ness so that the regression boxes participating in the training are all around the center point. After adopting this method, the AP of larger objects has been significantly improved.

$$centerness^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}} \tag{5}$$

4. Conclusion

This paper briefly summarizes the development process of the classical CNN network, and then describes the development process of the single-stage object detection algorithm [78-81], from the single-stage object detection algorithm based on anchor-based to the popular anchor-free single-stage detection algorithm in recent years. It is mainly pointed out that the improvement is in the following aspects: using a better pyramid structure to extract the feature layer; proposing deeper and wider network agent architecture; improving the anchor-free mechanism [68, 82, 83]; stronger image enhancement strategy; and many other details.

Deep learning based on the single-stage algorithm is developing rapidly, mainly because the single-stage detection algorithm has a simple structure and can be combined with the Internet of Things to deal with real-time application scenarios, such as fire monitoring, online detection, high-altitude work online monitoring, online speed detection on expressways, and so on. Although the single-stage detection algorithm is still in the process of continuous improvement, it is still not accurate enough in location, small object detection, multi-background, and multi-domain detection [84-86], and it still faces many thorny problems. How to reduce the decline inaccuracy caused by complex background or domain differences, low network recall, and other issues will become a hot research direction in the object field.

References

[1] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[2] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International journal of computer vision*, vol. 38, no. 1, pp. 15-33, 2000.

[3] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212-3232, 2019.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[5] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, 1998: IEEE, pp. 555-562.

[6] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 528-537.

[7] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *2009 IEEE Conference on computer vision and Pattern Recognition*, 2009: IEEE, pp. 1271-1278.

[8] 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.

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1532-1545, 2014.

[10] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *CVPR 2011*, 2011: IEEE, pp. 1585-1592.

[11] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," *Advances in neural*

information processing systems, vol. 26, 2013.
[12] Y.-D. Zhang, "A five-layer deep convolutional neural network with stochastic pooling for chest CT-based COVID-19 diagnosis," *Machine Vision and Applications*, vol. 32, 2021, Art no. 14.
[13] S.-H. Wang, "Covid-19 Classification by FGCNet with Deep Feature Fusion from Graph Convolutional Network and Convolutional Neural Network," *Information Fusion*, vol. 67, pp. 208-229, 2020/10/09/ 2021.
[14] A. K. Sangaiah, "Alcoholism identification via convolutional neural network based on parametric ReLU, dropout, and batch normalization," *Neural Computing and Applications*, vol. 32, pp. 665-680, 2020.
[15] S.-H. Wang and J. Sun, "Cerebral micro-bleeding identification based on a nine-layer convolutional neural network with stochastic pooling," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 1, p. e5130, 2020.
[16] K. Muhammad, "Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3613-3632, 2019.
[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.
[18] A. Cuesta-Infante, F. J. Garc?a, J. J. Pantrigo, and A. S. Montemayor, "Pedestrian detection with LeNet-like convolutional networks," (in English), *Neural Computing & Applications*, Article; Proceedings Paper vol. 32, no. 17, pp. 13175-13181, Sep 2020.
[19] V. C. Swetha, D. Mishra, and S. S. Gorthi, "Eigenvector Orientation Corrected LeNet for Digit Recognition," in *3rd International Conference on Computer Vision and Image Processing (CVIP)*, PDPM Indian Inst Informat Technol, Design & Mfg, Jabalpur, INDIA, 2018, vol. 1022, 2020, pp. 313-324.
[20] S. Lu, "Detection of abnormal brain in MRI via improved AlexNet and ELM optimized by chaotic bat algorithm," *Neural Computing and Applications*, Accessed on: 2020/06/13. doi: 10.1007/s00521-020-05082-4 [Online]. Available: https://doi.org/10.1007/s00521-020-05082-4
[21] S. Lu, "Pathological Brain Detection based on AlexNet and Transfer Learning," *Journal of Computational Science*, vol. 30, pp. 41-47, 2019.
[22] S. Xie, "Alcoholism Identification Based on an AlexNet Transfer Learning Model," (in English), *Frontiers in Psychiatry*, Original Research vol. 10, 2019-April-11 2019, Art no. 205.
[23] V. V. Govindaraj, "High performance multiple sclerosis classification by data augmentation and AlexNet transfer learning model," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 9, pp. 2012-2021, 2019.
[24] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
[25] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*, 2015: IEEE, pp. 1316-1322.
[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014: Springer, pp. 818-833.
[27] A. M. C. Antioquia, D. S. Tan, A. Azcarraga, W. H. Cheng, and K. L. Hua, "ZipNet: ZFNet-level Accuracy with 48 x Fewer Parameters," in *33rd IEEE International Conference on Visual Communications and Image Processing (IEEE VCIP)*, Taichung, TAIWAN, 2018: IEEE, 2018, pp. 1-11.
[28] R. S. Sinha and S. H. Hwang, "Comparison of CNN Applications for RSSI-Based Fingerprint Indoor Localization," *Electronics*, vol. 8, no. 9, Sep 2019, Art no. 989.
[29] V. Makde, J. Bhavsar, S. Jain, and P. Sharma, "Deep Neural Network Based Classification of Tumourous and Non-tumorous Medical Images," in *2nd International Conference on Information and Communication Technology for Intelligent Systems (ICTIS)*, Ahmedabad, INDIA, 2017, vol. 84, 2018, pp. 199-206.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[31] M. A. Khan, "VISPNN: VGG-Inspired Stochastic Pooling Neural Network," *Computers, Materials & Continua*, vol. 70, 2, pp. 3081-3097, 2022.
[32] S. Fernandes, "AVNC: Attention-based VGG-style network for COVID-19 diagnosis by CBAM," *IEEE Sens. J.* doi: 10.1109/JSEN.2021.3062442
[33] Q. Zhou, "ADVIAN: Alzheimer's Disease VGG-Inspired Attention Network Based on Convolutional Block Attention Module and Multiple Way Data Augmentation," *Front. Aging Neurosci.*, vol. 13, 2021, Art no. 687456.
[34] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[36] S. Lu, "Detecting pathological brain via ResNet and randomized neural networks," *Heliyon*, vol. 6, no. 12, p. e05625, 2020.
[37] X. Yu, "ResNet-SCDA-50 for Breast Abnormality Classification," *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 94-102, Jan 2021.
[38] Y. T. Xiao, "TReC: Transferred ResNet and CBAM for Detecting Brain Diseases," *Front. Neuroinformatics*, vol. 15, Dec 2021, Art no. 781551.
[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.
[40] X. Yu, "Utilization of DenseNet201 for diagnosis of breast abnormality," *Machine Vision and Applications*, vol. 30, no. 7-8, pp. 1135-1144, 2019/10/01 2019.
[41] S.-H. Wang, "DenseNet-201-Based Deep Neural Network with Composite Learning Factor and Precomputation for Multiple Sclerosis Classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 2s, p. Article 60, 2020.
[42] S. C. Satapathy, "Covid-19 diagnosis via DenseNet and optimization of transfer learning setting," *Cognitive Computation*. doi: 10.1007/s12559-020-09776-8
[43] Y. Yan, "A survey of computer-aided tumor diagnosis based on convolutional neural network," *Biology*, vol. 10, no. 11, 2021, Art no. 1084.
[44] Z. Zhu, "PSCNN: PatchShuffle Convolutional Neural Network for COVID-19 Explainable Diagnosis," *Frontiers in Public Health*, vol. 9, 2021, Art no. 768278.
[45] K. Wu, "SOSPCNN: Structurally Optimized Stochastic Pooling Convolutional Neural Network for Tetralogy of Fallot Recognition," *Wireless Communications and Mobile*

Computing, vol. 2021, p. 5792975, 2021/07/02 2021, Art no. 5792975.
[46] S. C. Satapathy and D. Wu, "Improving ductal carcinoma in situ classification by convolutional neural network with exponential linear unit and rank-based weighted pooling," *Complex Intell. Syst.*, vol. 7, pp. 1295-1310, 2020/11/22 2021.
[47] D. S. Guttery, "Improved Breast Cancer Classification Through Combining Graph Convolutional Network and Convolutional Neural Network," *Information Processing and Management*, vol. 58, 2, 2021, Art no. 102439.
[48] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627-1645, 2009.
[49] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 516-520.
[50] 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.
[51] 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.
[52] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 17-24.
[53] H. Schneiderman and T. Kanade, "Object detection using the statistics of parts," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 151-177, 2004.
[54] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Hoggles: Visualizing object detection features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1-8.
[55] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
[56] F. S. Khan, R. M. Anwer, J. Van De Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, "Color attributes for object detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: IEEE, pp. 3306-3313.
[57] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe, "Scalable, high-quality object detection," *arXiv preprint arXiv:1412.1441*, 2014.
[58] 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.
[59] K. Chen *et al.*, "Towards accurate one-stage object detection with ap-loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5119-5127.
[60] Y. Chen, C. Han, N. Wang, and Z. Zhang, "Revisiting feature alignment for one-stage object detection," *arXiv preprint arXiv:1908.01570*, 2019.
[61] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021: IEEE Computer Society, pp. 3490-3499.
[62] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[63] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.
[64] 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.
[65] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390-391.
[66] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.
[67] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer vision*, 2016: Springer, pp. 21-37.
[68] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 913-922.
[69] K. Fujii and K. Kawamoto, "Generative and self-supervised domain adaptation for one-stage object detection," *Array*, vol. 11, p. 100071, 2021.
[70] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.
[71] V. Y. Mariano *et al.*, "Performance evaluation of object detection algorithms," in *Object recognition supported by user interaction for service robots*, 2002, vol. 3: IEEE, pp. 965-969.
[72] R. Girshick, P. Felzenszwalb, and D. McAllester, "Object detection with grammar models," *Advances in neural information processing systems*, vol. 24, 2011.
[73] Y. Yang, M. Li, B. Meng, J. Ren, D. Sun, and Z. Huang, "Rethinking the Aligned and Misaligned Features in One-stage Object Detection," *arXiv preprint arXiv:2108.12176*, 2021.
[74] 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.
[75] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational visual media*, vol. 5, no. 2, pp. 117-150, 2019.
[76] 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.
[77] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3611-3620.
[78] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-11, 2021.
[79] J. Ding *et al.*, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[80] J. Mao *et al.*, "Voxel transformer for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3164-3173.
[81] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *arXiv preprint arXiv:2101.07663*, 2021.
[82] X. Zhang, F. Wan, C. Liu, X. Ji, and Q. Ye, "Learning to match anchors for visual object detection," *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 2021.
[83] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, "Deep learning in multi-object detection and tracking: state of the art," *Applied Intelligence*, vol. 51, no. 9, pp. 6400-6429, 2021.
[84] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Systems with Applications*, vol. 172, p. 114602, 2021.
[85] M. Xu *et al.*, "End-to-end semi-supervised object detection with soft teacher," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3060-3069.
[86] Z. Fan, Y. Ma, Z. Li, and J. Sun, "Generalized few-shot object detection without forgetting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4527-4536.