Journal on Artificial Intelligence JAI, vol.2, no.2, pp.59-77, 2020

A Review of Object Detectors in Deep Learning

Chen Song[1], Xu Cheng[1, *], Yongxiang Gu[1], Beijing Chen[1] and Zhangjie Fu[1]

Abstract: Object detection is one of the most fundamental, longstanding and significant problems in the field of computer vision, where detection involves object classification and location. Compared with the traditional object detection algorithms, deep learning makes full use of its powerful feature learning capabilities showing better detection performance. Meanwhile, the emergence of large datasets and tremendous improvement in computer computing power have also contributed to the vigorous development of this field. In the paper, many aspects of generic object detection are introduced and summarized such as traditional object detection algorithms, datasets, evaluation metrics, detection frameworks based on deep learning and state-of-the-art detection results for object detectors. Finally, we discuss several promising directions for future research.

Keywords: Object detection, deep learning, computer vision.

1 Introduction

Object detection [Fischler and Elschlager (1973)] is one of the most fundamental and challenging tasks in the field of computer vision, which aims to recognize object categories and predict the location of each object by a bounding box [Everingham, Van Gool ,Williams et al. (2010); Deng, Dong, Socher et al. (2009)]. At the same time, object detection is the cornerstone of image understanding and computer vision, which plays an extremely vital role in solving complex computer tasks, such as image segmentation, image capture, scene understanding, video tracking and other visual tasks. Therefore, exploring fast and accurate object detection methods is an exceedingly significant and challenging task. This article summarizes recent popular deep learning-based object detection methods and discusses future promising directions. Specially speaking, in Section 2, we explore traditional object detectors. Datasets and evaluation metrics are listed in Section 3. Then detectors based on deep learning are summarized in Section 4 and state-of-the-art detection results for object detectors based on several datasets are listed in Section 5. Finally, we summarize and explore future promising directions in Section 6.

2 Previous work

Traditional detectors in the early stage before the deep learning was mainly divided into three steps: regional proposal generation, feature vector extraction and region

[1] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China.

[*] Corresponding Author: Xu Cheng. Email: xcheng@nuist.edu.cn.

Received: 16 February 2020; Accepted: 18 April 2020.

classification. During regional proposal generation, an intuitive idea was usually to employ sliding windows technique [Vedaldi, Gulshan, Varma et al. (2009); Viola and Jones (2001); Dalal and Triggs (2005); Viola and Jones (2004)] to obtain areas of images that may contain objects. During the second phase, traditional model usually used SIFT(Scale Invariant Feature Transform) [Lowe (1999)], Haar [Lienhart and Maydt (2002)], HOG (Histogram of Gradients) [Dalal and Triggs (2005)], or SURF(Speeded Up Robust Features) [Bay, Tuytelaars and Van Gool (2006)] to extract a fixed-length feature vector in the regional proposal, which was used to capture the semantic information of different positions of the image and express the feature information of the image objects. Finally, the extracted feature vectors were commonly classified by SVM [Hearst, Dumais, Osuna et al. (1998)]. Additionally, some related techniques such as adaboost [Freund and Schapire (1996)], bagging [Opitz and Maclin (1999)] or cascade learning [Viola and Jones (2004)] were employed to further improve detection accuracy.

As the winner of the VOC Object Detection Challenge three times, the Deformable Part-based Model (DPM) [Felzenszwalb, Girshick, McAllester et al. (2009)] was the peak of the traditional detection techniques. As an extension of the HOG detector, DPM combined the idea of dividing and conquering, and has achieved the unprecedented object detection accuracy. Specifically, DPM could be seen as learning a correct method of object component decomposition during training, and combining different object components in reasoning. The improved DPM model [Felzenszwalb, Girshick and McAllester (2010); Girshick, Felzenszwalb and Mcallester (2011); Girshick (2012)] combined some other beneficial strategies, such as hard negative mining [Bucher, Herbin and Jurie (2016)], cascade architecture, and bounding box regression, to handle real-world objects with more sophisticated challenges.

However, between 2008 and 2012, detection accuracy stagnated and stabilized on the canonical PASCAL VOC dataset. The best performing detection models during that time were complex integrated systems, combining multiple low-dimensional image features with high-level context. At the same time, the drawbacks and limitations of traditional detectors were showed, the most prominent of which is that traditional feature extractors failed to capture the high-level semantic features and sophisticated content of the image.

3 Datasets and evaluation metrics

Datasets play a significant role in the development of object detectors. In this section, several significant datasets for generic object detection task are briefly reviewed and evaluation metrics are summarized in Tab. 2.

3.1 Datasets

From the initial 20 classification of the lesser picture PASCAL VOC [Everingham, Van Gool, Williams et al. (2010)] to the subsequent multi-category of millions of large dataset ImageNet [Deng, Dong, Socher et al. (2009)], the object detection methods based on deep learning have mushroomed. Additionally, the emergence of large dataset such as MS COCO, Open Images etc. [Lin, Maire, Belongie et al. (2014); Kuznetsova, Rom, Alldrin et al .(2000); Papageorgiou and Poggio (2000); Dalal and Triggs (2005)], provided more abundant and finer image features for detection approaches through more sophisticated

image annotation information and finer picture outlines. It was the publication of these datasets that promoted the development of object detection and other computer vision tasks. Therefore, datasets play a significant role in computer vision tasks in the last decade. Some overviews and highlights of significant datasets are listed in Tab. 1.

Table 1: Significant datasets for object detection

Dataset	Total Images	Classes	Highlights
PASCAL VOC (2012)	11540	20	Contains 20 categories of common objects in life. Complete picture annotations.
ImageNet	14 millions+	21841	Ample image samples and rich object classes. More indiscernible than PASCAL VOC.
MS COCO	328,000+	80	The object in image are moderately sized and centered. Relatively fine picture annotations.
Open ImagesV5	2.8 millions+	350	The split mask marks the outline of the object. More accurate object outlines and finer image annotations.
Places	10 millions+	434	The largest labeled dataset for scene recognition
Objects365	6300000+	365	Universal detection dataset with large scale and high quality

PASCAL VOC [Everingham, Van Gool, Williams et al. (2010)] was an excellent dataset that standardized image recognition and classification. From the initial four classifications of 2005 to the 20 classifications of common objects later, the PASCAL VOC dataset provided increasingly rich pictures and categories. At the same time, all previous images were retained to serve as future training sets, which stimulated the development of the model and improves the overall performance and accuracy of the detection methods.

ImageNet [Deng, Dong, Socher et al. (2009)] was a tremendous dataset widely used in the ILSVRC challenging competition, which contained more than 14 million full-size

tagged images and 20,000 image categories. The ILSVRC competition mainly included image classification and object positioning, detection, video detection, and scene classification, and provided different dataset for each part.

MS COCO [Lin, Maire, Belongie et al. (2014)] was considered one of the most authoritative and richest datasets in computer vision. The biggest advantage of dataset was that the objects were moderately sized and centered in the image, with the goal of understanding the scene. In the object detection challenge, the MS COCO provided 200,000 images and more than 500,000 object annotations in 80 categories, making it one of the most widely publicly available detection databases in the world.

Open Images [Kuznetsova, Rom, Alldrin et al. (2000); Ferrari (2018)] (now V5 in 2019) was the largest publicly available object detection dataset publicly available from Google. Open Images V5 contained a split mask of 350 categories and 2.8 million object instances, which differed from the bounding box in which only the object was identified, the split mask marked the outline of the object, characterizing its spatial extent to a higher level of detail. What's even more breathtaking was that the Google team released about 100,000 masked images on the validation and test set, and these masked images were all manually annotated.

Object365 [Shao, Li, Zhang et al. (2019)] was released by Kuangshi Research Institute in 2019 had the characteristics of large scale, high quality and strong generalization ability. Compared with MS COCO dataset, Objects365 contained 630000 pictures, about 5 times as many as MS COCO; contained about 10 million object annotation boxes, about 11 times as many as MS COCO dataset annotation boxes; the average annotation boxes of each picture of Objects365 were 15.8, more than 2 times as many as MS COCO dataset. In addition, Using the Object365 dataset and the previously released large-scale CrowdHuman dataset, Kuangshi technology and Beijing Zhiyuan Artificial Intelligence Research Institute jointly held the field inspection challenge (DIW 2019), and successfully applied for the seminar of computer vision and model recognition conference in 2019.

3.2 Evaluation metrics

Generally, there are three metrics for evaluating detectors performance: speed in Frame Per Second (FPS), precision and recall. Specific details of evaluation metrics can be seen in Tab. 2, and also found in Everingham et al. [Everingham, Van Gool, Williams et al. (2010); Deng, Dong, Socher et al. (2009)].

4 Detection paradigms in deep learning

Object detectors based on deep learning are mainly divided into two classes: Region Based (Two Stage) Detectors and Unified (One Stage) Detectors. Region based detectors usually involve two steps of regional proposal generation and classification prediction. In contrast, Unified detectors normally employ a complete framework to predict object areas in pictures and classify them. Therefore, region-based detectors could acquire higher accuracy while unified detectors have faster detection speed. Several significant detectors based on deep learning is shown in Fig. 1.

Table 2: Metrics for evaluating detectors performance

Metric	Meaning	Definition and Description	
FPS	Frame per second	The number of images processed per second	
ε	IOU threshold	Standard for assessing location accuracy	
TP	True Positive	Correct predictions from samples	
FP	False Positive	Wrong predictions from samples	
β	Confidence threshold	Indicators used to calculate precision and recall	
P(β)	Precision	The fraction of correct detections out of the total detections	
R (β)	Recall	The fraction of all objects detected by the detector having a confidence of at least β	
AP	Average Precision	Computed over the different levels of recall by varying the β	
mAP	mean AP	Average score of AP across all classes	
TPR	True Positive Rate	The fraction of TP over FP	
FPPI	FP Per Image	The fraction of FP over each image	
		Generic Object Detection	
mAP	mean Average Precision	VOC	mAP at 0.5 IOU threshold over all 20 classes
		ILSVRC	mAP at a modified IOU over all classes
		MS COCO	• APcoco: mAP averaged over ten ε: {0.5: 0.05: 0.95}
			• AP_{50}: mAP at 0.50 IOU threshold
			• AP_{75}: mAP at 0.75 IOU threshold
			• AP_S: AP coco for small objects of area smaller than 32^2
			• AP_M: AP coco for objects of area between 32^2 and 96^2
			• AP_L: AP coco for large objects of area bigger than 96^2

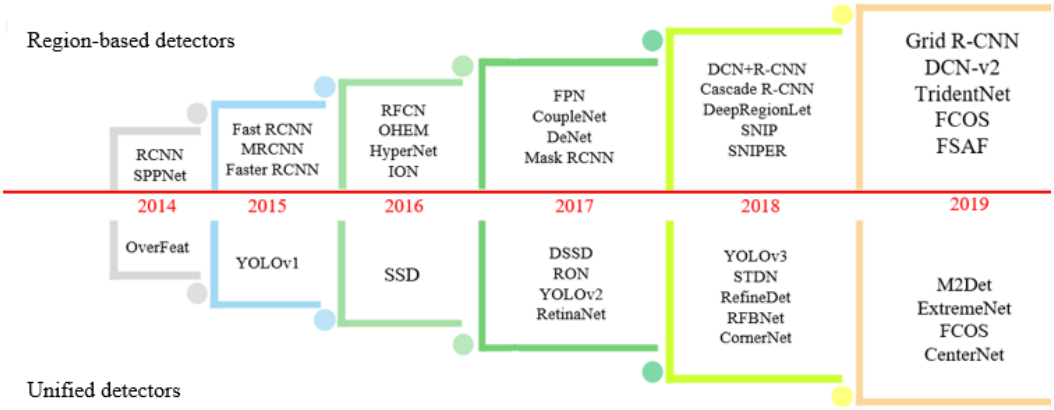

Figure 1: Several typical region based detectors and unified detectors

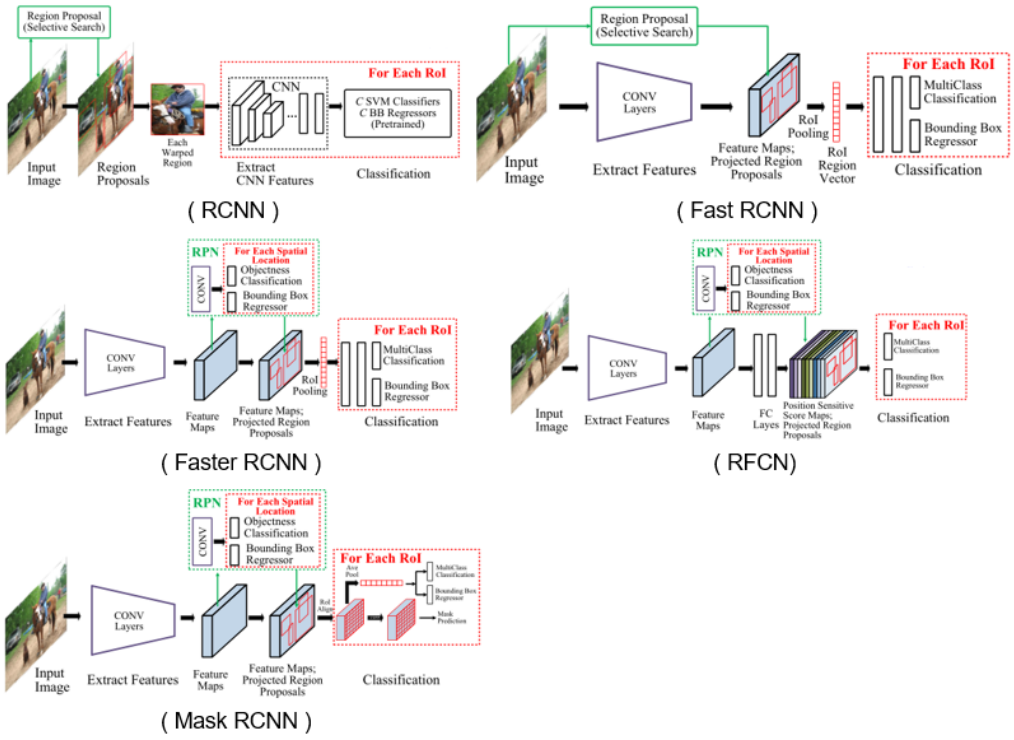

Figure 2: A brief description of several two-stage detector frameworks

4.1 Region based detectors

Region Based Detectors generally split object detection task into two steps, candidate region generation and classification prediction. During candidate region generation phase, an existing search algorithm [Endres and Hoiem (2010); Rahtu, Kannala and Blaschko (2011); Alexe, Deselaers and Ferrari (2012); Uijlings, Van De Sande, Gevers et al. (2013)]

is used to generate a plurality of proposal regions of the image that may contain objects. After that, the deep convolutional neural network is used to extract the features of the object in the candidate regions. Finally, extracted feature vectors in the image is classified and predicted from the predefined category label. Next, some of significant and excellent detectors among Region Based Detectors [Girshick, Donahue, Darrell et al. (2014); He, Zhang, Ren et al. (2015); Girshick (2015); Ren, He, Girshick et al. (2015); Dai, Li, He et al. (2016); He, Gkioxari, Dollár et al .(2017); Cai and Vasconcelos (2018); Ouyang, Wang, Zhu et al. (2017); Li, Peng, Yu et al. (2017); Lu, Li, Yue et al. (2019)] are reviewed and summarized as follows, and an overview of several region based detectors (e.g., RCNN, Fast RCNN, Faster RCNN, RFCN and Mask RCNN) is shown in Fig. 2.

RCNN Series Detectors [Girshick, Donahue, Darrell et al. (2014); He, Zhang, Ren et al. (2015); Girshick (2015); Ren, He, Girshick et al. (2015)]: In the past few years, convolutional neural networks (CNNs) [Krizhevsky, Sutskever and Hinton (2012); Simonyan and Zisserman (2014); Szegedy, Liu , Jia et al. (2015); He, Zhang, Ren et al. (2016); Huang, Liu, Van Der Maaten et al. (2017); Hu, Shen and Sun (2018); Ghiasi, Lin and Le (2019)] had achieved great success in the ImageNet classification task through its powerful hierarchical feature learning ability. After that, Girshirk et al. explored RCNN framework for general object detection and semantic segmentation in 2014, which integrated AlexNet [Krizhevsky, Sutskever and Hinton (2012)] with Selective Search algorithm [Uijlings, Van De Sande, Gevers et al. (2013)]. Apart from that, R-CNN adopted transfer learning and fine-tune techniques to further improve detection accuracy. After that, plenties of models were proposed on the basis of RCNN such as SPPNet, Fast RCNN, Faster RCNN etc.

Inspired by spatial pyramid matching [Kleban, Xie and Ma (2008)], He et al. adopted Spatial Pyramid Pooling (SPP) [He, Zhang, Ren et al. (2015)] into RCNN to accelerate the calculation of spatial feature vectors and solve the problem of fixed-size inputs. Naturally, R-CNN with SPP layer could significantly improve reasoning and detection speed while maintaining comparable or better performance over the no-SPP architecture. However, SPPNet performs an obvious disadvantage that it fails to update the convolutional layer parameters before the SPP layer by back propagation, which greatly limits the performance of deep CNNs. Additionally, feature vectors also require extra disk space to store.

To solve these above problems, Girshirk proposed Fast RCNN [Girshick (2015)] to accelerate the calculation of the entire network features and achieve better object detection performance. Specially speaking, Fast RCNN replaces spatial pyramid pooling layer in SPPNet with ROI pooling and changes the final SVM classification to the fully connected layer with SoftMax supervised classification, adding one more connected layer branch for border regression.

Although Fast RCNN can greatly speed up the detection process, it takes more time to process the image in the regional proposal phase than the feature extraction. Additionally, model uses Selective Search [Uijlings, Van De Sande, Gevers et al. (2013)] or Edge Boxes [Zitnick and Dollár (2014)] to generate proposal regions, which are based on low-dimensional image features to produce redundant or repeated candidate regions. In order to generate candidate regions better and faster, Ren et al. proposed Faster RCNN [Ren,

He, Girshick et al. (2015)], which offered a more accurate and faster regional generation proposal called Region proposal Network (RPN) employing anchor boxes mechanism. While Faster RCNN exhibits superior detection accuracy and faster detection speed than previous models, it fails to share feature computation during the classification phase. Such amount of calculation is huge because there are hundreds of candidate regions in each picture.

Fast RCNN and Faster RCNN varients [Bell, Lawrence Zitnick, Bala et al. (2016); Kong, Yao, Chen et al. (2016); Shrivastava, Gupta and Girshick (2016); Lin, Dollár, Girshick et al. (2017); Dai, Li , He et al. (2016); Zhu, Zhao, Wang et al. (2017); Li, Peng, Yu et al. (2017)]: In order to improve the accuracy of detecting small objects, ION [Bell, Lawrence Zitnick, Bala et al. (2016)] adds skip connection and Recurrent Neural Network (RNN) on the basis of Fast RCNN, in which L2 regularized skip connection is used to extract multi-layer features and RNN is used to extract object context information. At the same time, Kong et al. [Kong, Yao, Chen et al. (2016)] proposes a deeply hierarchical structure HyperNet, which is used to generate object proposals and detect objects at the same time. In the same year, shrivast et al. [Shrivastava, Gupta and Girshick (2016)] proposed a new bootstrapping method OHEM based on the Stochastic gradient descent algorithm. The main idea of OHEM is to select only some difficult samples for back propagation training, so it can greatly shorten the reasoning time and the network focuses on distinguishing negative samples. All the detection architectures described before are only predicted on the top-level feature map, but due to the low resolution of the high-level feature map, the detectors have always performed poorly in small object detection. In order to achieve better detection performance, Lin et al. [Lin, Dollár, Girshick et al. (2017)] proposes FPN network, which combines high-level and low-level information and predicts hierarchically. Noticing that the fully connected layers greatly increase the calculation of the feature, Dai et al. [Dai, Li, He et al. (2016)] proposes Region-based Fully Convolutional Networks (R-FCN) to solve this problem, which ensures the entire network sharing feature calculation. Compared to Faster RCNN, R-FCN used the ResNet101 [He, Zhang, Ren et al. (2016)] Benchmark network and replaces the ROI layer and the fully connected layer with Position Sensitive Score Maps and Position Sensitive ROI Pooling Layer. After that, Zhu et al. [Zhu, Zhao, Wang et al .(2017)] proposes CoupleNet network based on global, local and context information, which uses the location sensitive graph in R-FCN to capture the local information of the object, and adopts the candidate region pooling in Fast RCNN to capture the global information and context information of the object. In order to further improve detection speed, Li et al. [Li, Peng, Yu et al. (2017)] proposed Light Head RCNN, which used a large-core separable convolution to generate a feature map with a small number of channels, and then connected an RCNN sub-network that extracted the features of the classification and regression.

Mask RCNN [He, Gkioxari, Dollár et al. (2017)]**:** The instance segmentation task and detection task are interpenetrated, and the only difference is that instance segmentation requires assigning a specific category label to each pixel in the image. Early methods basically classified the images finely, and then used Fast RCNN to classify each region. In order to obtain more precise instance segmentation accuracy and higher efficiency, He et al. [He, Gkioxari, Dollár et al. (2017)] proposed Mask RCNN to handle instance

segmentation by extending Faster RCNN. Compared with Faster RCNN, Mask RCNN replaced ROI Pooling layer with ROI Align layer for preserving spatial correlation at the pixel level.

Cascade RCNN [Cai and Vasconcelos (2018)]: In order to find the appropriate IOU threshold for training and predicting, Cai et al. [Cai and Vasconcelos (2018)] proposed the Cascade RCNN, which consisted of a series of detectors trained by increasing the IOU threshold cascade. Resampling is adopted to make all detectors close to positive sample training set during training to reduce overfitting in the network. In addition, the overall model also adopts the same cascading mode during testing, which promoted a better match between the candidate frame and the detector at each stage.

SNIP and TridentNet [Singh and Davis (2018); Singh and Najibi (2018); Li, Chen, Wang et al. (2019)]: At present, the common object detection datasets mainly include two challenges: small object size and large object size difference. In order to better deal with the problem of image size invariance, Singh et al. [Singh and Davis (2018)] Proposes a new training scheme, Scale Normalization for Image Pyramids(SNIP), which selectively propagates the gradient of object instances of different sizes as the loss function of image size and extracts candidate regions and training according to the setting resolution threshold. After that, SNIPER [Singh, Najibi and Davis (2018)] is an improved version of SNIP. It scales the image to three sizes and extracts a certain size of chip according to certain rules, and then scales the chip to a fixed size for training and detection. In addition, compared with SNIP, which only focuses on the candidate areas obtained by the recommendation box algorithm, SNIPER adopts the negative chip sampling strategy to select the candidate areas that are easy to be misjudged as objects. After that, Li et al. [Li, Chen, Wang et al. (2019)] proposes tridentnet algorithm based on the receptive field, which increases the receptive field of the network by introducing hole convolution, so as to detect targets at different scales. Specifically, tridentnet replaces the convolution layer of Faster RCNN feature extraction network with a convolution layer of three branches equipped with different dilated parameters. So that on the one hand, the time of forward calculation of the network is reduced, on the other hand, the parameters learned by the network have better generalization ability.

4.2 Unified detectors

Compared to region based detectors, unified detectors exhibit faster detection speed. Concretely speaking, unified detectors combine regional proposal and classification prediction, which take all positions in the picture as candidate regions and send entire picture into convolution network to extract feature vectors, and finally output the category and position of the object directly. Therefore, unified frameworks as a whole can be optimized end-to-end manner, and it is foreseeable that the detection accuracy is worse as a result of the extraction of fewer picture feature vectors. Next, some of significant and excellent detectors among Unified Detectors [Sermanet, Eigen, Zhang et al. (2013); Redmon, ivvala, Girshick et al. (2016); Redmon and Farhadi (2017); Redmon and Farhadi (2018); Liu, Anguelov, Erhan et al. (2016); Fu, Liu , Ranga et al. (2017); Lin, Goyal, Girshick et al. (2017); Law and Deng (2018)] are reviewed and summarized as follows, and an overview of unified detectors (e.g., YOLO, SSD, CornerNet and

RetinaNet) are shown in Fig. 3.

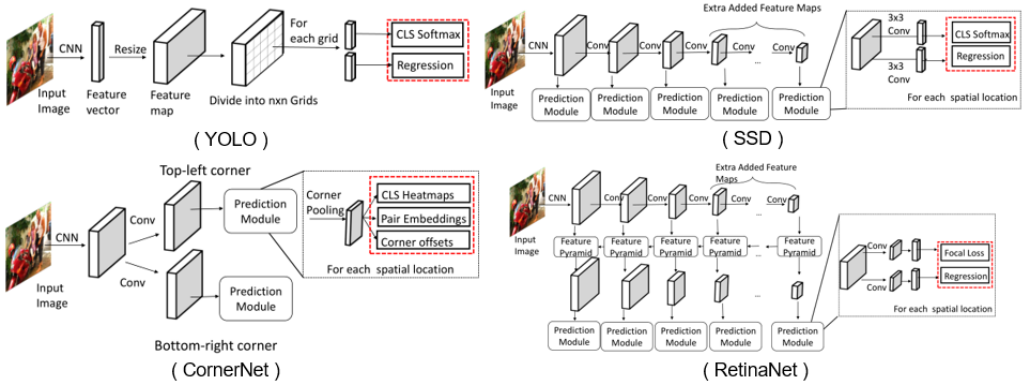

Figure 3: A brief description of several one-stage detector frameworks

Overfeat [Sermanet, Eigen, Zhang et al. (2013)**]**: As one of the earliest unified detectors, Sermanet et al. **[**Sermanet, Eigen, Zhang et al. (2013)**]**proposed a feature detector in 2013 called Overfeat using a fully convolutional deep network, which trained multiple tasks and assisted each other to improve accuracy through a single shared network, thus it could success to apply in classification, location and detection tasks. Concretely speaking, Overfeat employs multi-scale images and sliding window method to generate candidate regions and replaces final pooling layer with offset max pooling for obtaining more robust image features. Apart from this, not suppressing bounding boxes, Overfeat employs a cumulative bounding box approach to increase detection confidence. Nevertheless, Overfeat still has some places such as coordinates loss function, bounding box prediction etc. to optimize and improve.

YOLO Series Detectors [Redmon, ivvala, Girshick et al. (2016); Redmon and Farhadi (2017); Redmon and Farhadi (2018)]: Similar to the thought of Overfeat, Rendom et al. [Redmon, Divvala, Girshick et al. (2016)] developed a unified real-time detection framework called YOLO, which handled detection as a regression problem. The overall model employed a single deep convolutional network, in which input image was send to extract features and directly output the object classification and coordinates, thus the overall architecture could be optimized end-to-end manner and performed faster reasoning and detection speed compared with region based detectors. Despite all this, there are also some limitations of the YOLO, which yields more location errors and fails to recognize clusters or small objects, irregular or different aspect radio objects, untrained object categories etc. After that, Redmon et al. [Redmon and Farhadi (2017)] proposed YOLOv2 (better and faster) by mending the shortcomings of YOLO architecture and employing more excellent optimization methods and network architecture. Concretely speaking, model mainly focuses on improving model recall, object location and classification accuracy compared to YOLO, which applies some novel and beneficial methods such as multi-scale training, k-means clustering algorithm, joint optimization technique etc. to address the limitations of the original model. Additionally, YOLOv2 replaces the original VGG network [Simonyan and Zisserman (2014)] with Darknet19 benchmark network, and adds Batch Normalization after each convolution layer to

optimize the deep network, allowing the model to converge faster. Although YOLOv2 improved in the original model by a large margin, it failed to break the limit of inaccurate detection of small objects. Two years later, some novel ideas and continuous advancement of the deep network prompted Redmon and Farhadi to propose a better unified detector called YOLOv3 [Redmon and Farhadi (2018)], which was elevated on the basis of YOLOv2. Different from YOLOv2, model employed darknet53 as backbone network and resNet shortcut connection to avoid gradient disappearance. In addition, YOLOv3 referenced Feature Pyramid Network (FPN) [Lin, Dollár, Girshick et al. (2017)] to adopt three scale feature maps and replaced the original SoftMax with sigmoid and entropy on the loss function to support multi-tag prediction.

SSD Series Detectors [Liu, Anguelov, Erhan et al. (2016); Fu, Liu and Ranga et al. (2017)]: To optimize the problem of inaccurate position in the YOLO model, Liu et al. [Liu, Anguelov, Erhan et al. (2016)] proposed a faster and more accurate unified detector called SSD, which combined the regression idea in YOLO with the anchor box mechanism of Faster RCNN. The overall model architecture was fine-tuned on the VGG16 network, where the altrous algorithm was used to convert the last fully connected layers into convolutional layers and added several additional convolutional and pooling layers at the end of the network. By employing picture features at different resolutions extracted from multiple convolutional layers to predict Bounding boxes of different scales and sizes, the model could handle multiple sizes of objects even in low resolution pictures. Nevertheless, SSD model fails to solve the problem of low accuracy on detecting small objects, the dominating reason is that there are few features of small objects in the high-level feature maps. After that, Fu et al. [Fu, Liu, Ranga et al. (2017)] proposed the DSSD model for more accurate detection of small objects in the picture, which added a series of deconvolution layers after the original SSD architecture for more complete image information interaction. Specially speaking, DSSD model improves the feature representation ability of low-level high-resolution images for small objects detection, which takes account of high-level feature information and low-level spatial resolution simultaneously. However, due to the addition of deconvolution module and increase in calculation, the model demands more reasoning and detection time.

RetinaNet [Lin, Goyal, Girshick et al. (2017)]: Generally, unified detectors exhibit worse detection performance than region based detectors because former fails to process the imbalance of positive and negative samples. In order to solve this problem, Lin et al. [Lin, Goyal, Girshick et al. (2017)] proposed a unified detector called RetinaNet using Focal loss replace original cross entropy loss, which could reduce the weight of easily categorized samples during training and focused on those samples that are difficult to classify. The overall model architecture adopts FPN as backbone network to extract picture features, followed by two sub-networks for classification and border regression. Nevertheless, the fly in the ointment is that RetinaNet spends more reasoning and detection time than other state-of-the-art unified detectors. After that, Zhang et al. [Zhang, Wen, Bian et al. (2018)] proposes a single-stage detector RefineDet with balanced speed and accuracy, which consists of anchor refinement module, transfer connection block and object detection module. In addition, the feature fusion of feature pyramid network is also used in the architecture to effectively improve the detection performance on small objects. Specifically, the anchor refinement module filters the negative prediction box to reduce

the search range and slightly adjust the position and size of the anchor box. The transform connection block uses deconvolution to match the dimensions of the upper and lower modules, and adopts the way of bitwise addition to increase the advanced features.

CornerNet [Law and Deng (2018)]: Many of the state-of-the-art detectors have recently used the anchor boxes mechanism, but there are also some apparent flaws in the anchor boxes such as imbalanced positive and negative samples, excessive introduction of hyperparameters etc. Therefore, Law et al. [Law and Deng (2018)] proposed a unified detector called CornerNet, which considered bounding box as a pair of key corner detections (top left and bottom right). At the same time, it also proposed a new pooling method corner pooling to help the model better locate the corners. In addition, model's backbone network consists of two hourglass networks [Newell, Yang and Deng (2016)], which no longer employs maximum pool and adopts intermediate supervision method. After that. Zhou et al. [Zhou, Wang and Krähenbühl (2019)] proposes a single-stage detector CenterNet [Duan, Bai, Xie et al. (2019)] using corner and center point prediction, which is modified on the CornerNet architecture. Similar to CornerNet, CenterNet uses thermodynamic diagram to realize and introduce the Gaussian distribution area of prediction point to calculate the real prediction value, and the loss function is also fine-tuned on the original basis. Compared with CornerNet, the model directly regresses the target frame size, and the predicted value can be obtained according to the target frame size and the coordinates of the center point. In addition, two additional modules, center pooling and cascade corner pooling, are introduced into the model. Center pooling is used to predict the key point branches of the center to help the center key point obtain more visual recognition information of the target, while cascade corner pooling is used to obtain more robust corner information.

5. State-of-the-art for Generic Object Detection

PASCAL VOC and MS COCO are the most commonly used datasets for object detection competitions, where former is a small dataset containing approximately two objects per image and latter is a large dataset including of multiple small objects per image. Therefore, it is more challenge to identify objects employing MS COCO dataset. Next, in Tab. 3 and Tab. 4 we list some excellent detector results on PASCAL VOC and MS COCO over the recent few years. In Tab. 3, for VOC2007, the models are trained on VOC2007 and VOC2012 trainval sets and tested on VOC2007 test set. For VOC2012, the models are trained on VOC2007 and VOC2012 trainval sets plus VOC2007 test set and tested on VOC2012 test set by default. In table 4, Detection results on MS COCO test-dev data set.

Table 3: Some of state-of-the-art detectors result on PASCAL VOC

Detector	Backbone Network	Year	VOC07 (IOU=0.5)	VOC12 (IOU=0.5)
Region Based Detectors:				
RCNN	VGG-16	2014	58.5	-
SPPNet	VGG-16	2014	59.2	-
Fast RCNN	VGG-16	2015	70.0	68.4
Faster RCNN	VGG-16	2015	76.4	73.8
OHEM	VGG-16	2016	74.6	71.9
CRAFT	VGG-16	2016	75.7	71.3
HyperNet	VGG-16	2016	76.3	71.4
ION	VGG-16	2016	79.2	76.4
R-FCN	ResNet-101	2016	80.5	77.6
DeNet512	ResNet-101	2017	77.1	73.9
CoupleNet	ResNet-101	2017	82.7	80.4
FPN-Reconfig	ResNet-101	2018	82.4	81.1
DCN+R-CNN	ResNet-101	2018	84.0	81.2
Unified Detectors:				
YOLO	VGG-16	2016	66.4	57.9
SSD512	VGG-16	2016	79.8	57.9
RON384	VGG-16	2017	75.4	73.0
YOLOv2	Darknet	2017	78.6	73.5
DSSD513	ResNet-101	2017	81.5	80.0
RefineDet512	VGG-16	2018	81.8	80.1
RFBNet512	VGG-16	2018	82.2	-
CenterNet	ResNet-101	2019	78.7	-

Table 4: Some of state-of-the-art detectors result on MS COCO

Detector	Backbone Network	Year	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Region Based Detectors:								
Fast RCNN	VGG-16	2015	19.7	35.9	-	-	-	-
Faster RCNN	VGG-16	2015	21.9	42.7	-	-	-	-
OHEM	VGG-16	2016	22.6	42.5	22.2	5.0	23.7	37.9
ION	VGG-16	2016	23.6	43.2	23.6	6.4	24.1	38.3
R-FCN	ResNet-101	2016	29.9	51.9	-	10.8	32.8	45.0
DeNet-101	ResNet-101	2017	33.8	53.4	36.1	12.3	36.1	50.8
CoupleNet	ResNet-101	2017	34.4	54.8	37.2	13.4	38.1	50.8
Mask RCNN	ResNeXt-101	2017	39.8	62.3	43.4	22.1	43.2	51.2
DCN+R-CNN	ResNeXt-101	2018	42.6	65.3	46.5	26.4	46.1	56.4
Cascade R-CNN	ResNeXt-101	2018	42.8	62.1	46.3	23.7	45.5	55.2
SNIP++	DPN-98	2018	45.7	67.3	51.1	29.3	48.8	57.1
SNIPER++	ResNet-101	2018	46.1	67.0	51.6	29.6	48.9	58.1
Grid R-CNN	ResNeXt-101	2019	43.2	63.0	46.6	25.1	46.5	55.2
TridentNet	ResNet-101	2019	42.7	63.6	46.5	23.9	46.6	56.6
Unified Detectors:								
SSD512	VGG-16	2016	28.8	48.5	30.3	10.9	31.8	43.5
YOLOv2	DarkNet-19	2017	21.6	44.0	19.2	5.0	22.4	35.5
DSSD513	ResNet-101	2017	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet800+	ResNet-101	2018	39.1	59.1	42.3	21.8	42.7	50.2
RefineDet512	ResNet-101	2018	36.4	57.5	39.5	16.6	39.9	51.4
CornerNet511	Hourglass-104	2018	40.5	56.5	43.1	19.4	42.7	53.9
FCOS	ResNeXt-101	2019	42.1	62.1	45.2	25.6	44.9	52.0
CenterNet511++	Hourglass-104	2019	47.0	64.5	50.7	28.9	49.9	58.9

6 Conclusion and Future researches

In the past two decades, universal object detectors based on deep learning have flourished and achieved remarkable achievements. This article not only reviews some commonly used detection datasets, evaluation metrics, objects detectors based on deep learning methods, but also summarizes some of innovative technologies and inadequacies of model for some significant detectors to provide direction for future improvement. Concretely speaking, one of the hottest research topics in the future is to combine Auto Machine Learning to find the optimal detector architecture and optimization strategies. Then, when the number of pictures is rich and picture information is insufficient, excellent results can also be obtained by applying weakly supervised detection to detectors. In addition, recent researches have shown that efficient combination of contextual information can greatly improve detection performance, as objects in the pictures have strong relationships. Therefore, one of promising directions of future research on image object recognition is how to effectively and correctly incorporate image context information. In conclusion, the above three points would be promising directions to further improve detectors performance.

Funding Statement: This work is supported in part by the National Natural Science Foundation of China (Grant No. 61802058); in part by the International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 61911530397); in part by the Equipment Advance Research Foundation Project of China (Grant No. 61403120106); in part by the Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology (Grant No. 2018r057); in part by the Open Project Program of the State Key Lab of CAD&CG (Grant No. A1919), Zhejiang University, and the PAPD fund.

Conflicts of Interest: Compliance with ethical standards, the authors declare that they have no conflict of interest.

References

Alexe, B.; Deselaers, T.; Ferrari, V. (2012): Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189-2202.

Bay, H.; Tuytelaars, T.; Van Gool, L. (2006): Surf: Speeded up robust features. *European Conference on Computer Vision*, Springer, Berlin, Heidelberg. pp. 404-417.

Bell, S.; Lawrence Zitnick, C.; Bala, K.; Girshick, R. (2016): Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Bucher, M.; Herbin, S.; Jurie, F. (2016): Hard negative mining for metric learning based zero-shot classification. *European Conference on Computer Vision*, Springer, Cham, pp. 524-531.

Cai, Z.; Vasconcelos, N. (2018): Cascade R-CNN: Delving into high quality object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Dai, J.; Li, Y.; He, K.; Sun, J. (2016): R-FCN: Object detection via region-based fully convolutional networks. *Advances in Neural Information Processing Systems*, pp. 379-387.

Dalal, N.; Triggs, B. (2005): Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; Li, F.F. (2009): Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition.*

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. (2019): Centernet: Object detection with keypoint triplets. *arXiv preprint arXiv:1904.08189.*

Endres, I.; Hoiem, D. (2010): Category independent object proposals. *European Conference on Computer Vision*, Springer, Berlin, Heidelberg. pp. 575-588.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision,* vol. 88, no .2, pp. 303-338.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; Ramanan, D. (2009): Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D. (2010): Cascade object detection with deformable part models. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

Ferrari, V. (2018): The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982.*

Fischler, M. A.; Elschlager, R. A. (1973): The representation and matching of pictorial structures. *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 67-92.

Freund, Y.; Schapire, R. E. (1996): Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*, vol. 96, pp. 148-156.

Fu, C. Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A. C. (2017): Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659.*

Ghiasi, G.; Lin, T. Y.; Le, Q. V. (2019): Nas-fpn: Learning scalable feature pyramid architecture for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Girshick, R. B.; Felzenszwalb, P. F.; Mcallester, D. A. (2011): Object detection with grammar models. *Advances in Neural Information Processing Systems*, pp. 442-450.

Girshick, R. B. (2012): From rigid templates to grammars: Object detection with structured models. *University of Chicago, Division of the Physical Sciences, Department of Computer Science.*

Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. (2014): Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Girshick, R. (2015): Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision.*

Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. (1998): Support vector machines. *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2016): Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

He, K.; Zhang, X.; Ren, S.; Sun, J. (2015): 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.

He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. (2017): Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision.*

Hu, J.; Shen, L.; Sun, G. (2018): Squeeze-and-excitation networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. (2017): Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Kleban, J.; Xie, X.; Ma, W. Y. (2008): Spatial pyramid mining for logo detection in natural scenes. *2008 IEEE International Conference on Multimedia and Expo.*

Kong, T.; Yao, A.; Chen, Y.; Sun, F. (2016): Hypernet: Towards accurate region proposal generation and joint object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012): Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097-1105.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I. et al. (2000): A trainable system for object detection. *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15-33.

Law, H.; Deng, J. (2018): Cornernet: Detecting objects as paired keypoints. *In Proceedings of the European Conference on Computer Vision (ECCV).*

Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. (2019): Scale-aware trident networks for object detection. *Proceedings of the IEEE International Conference on Computer Vision.*

Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. (2017): Light-head R-CNN: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264.*

Lienhart, R.; Maydt, J. (2002): An extended set of haar-like features for rapid object detection. *Proceedings. International Conference on Image Processing*, vol. 1, pp. I-I.

Lin, T. Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. (2017): Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; ... Zitnick, C. L. (2014): Microsoft coco: Common objects in context. *European Conference on Computer Vision.* Springer, Cham, pp. 740-755.

Lin, T. Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. (2017): Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision.*

Lin, T. Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. (2017): Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C. Y.; Berg, A. C. (2016): Ssd: Single shot multibox detector. *European Conference on Computer Vision.* Springer, Cham, pp. 21-37.

Lowe, D. G. (1999): Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision,* vol. 2, pp. 1150-1157.

Newell, A.; Yang, K.; Deng, J. (2016): Stacked hourglass networks for human pose estimation. *In European Conference on Computer Vision*. Springer, Cham, pp. 483-499.

Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. (2019): Grid R-CNN. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Opitz, D.; Maclin, R. (1999): Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, vol. 11, pp. 169-198.

Rahtu, E.; Kannala, J.; Blaschko, M. (2011): Learning a category independent object detection cascade. *2011 International Conference on Computer Vision.*

Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. (2016): You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Redmon, J.; Farhadi, A. (2017): YOLO9000: better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Redmon, J.; Farhadi, A. (2018): Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767.*

Ren, S.; He, K.; Girshick, R.; Sun, J. (2015): Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*. pp. 91-99.

Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. (2013): Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229.*

Shrivastava, A.; Gupta, A.; Girshick, R. (2016): Training region-based object detectors with online hard example mining. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

Singh, B.; Davis, L. S. (2018): An analysis of scale invariance in object detection snip. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Singh, B.; Najibi, M.; Davis, L. S. (2018): SNIPER: Efficient multi-scale training. *Advances in Neural Information Processing Systems*, pp. 9310-9320.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. et al. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; Smeulders, A. W. (2013): Selective search for object recognition. *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154-171.

Vedaldi, A.; Gulshan, V.; Varma, M.; Zisserman, A. (2009): Multiple kernels for object detection. *2009 IEEE 12th International Conference on Computer Vision.*

Viola, P.; Jones, M. (2001): Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 1, pp. I-I.

Viola, P.; Jones, M. (2004): Robust real-time face detection. *International Journal of Computer Vision,* vol. 57, no. 2, pp. 137-154.

Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S. Z. (2018): Single-shot refinement neural network for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Zhou, X.; Wang, D.; Krähenbühl, P. (2019): Objects as points. *arXiv preprint arXiv:1904.07850.*

Zhu, Y.; Zhao, C.; Wang, J.; Zhao, X.; Wu, Y. et al. (2017): Couplenet: Coupling global structure with local parts for object detection. *Proceedings of the IEEE International Conference on Computer Vision.*

Zitnick, C. L.; Dollár, P. (2014): Edge boxes: Locating object proposals from edges. *European Conference on Computer Vision*. Springer, Cham, pp. 391-405.