Computer Systems Science & Engineering
DOI:10.32604/csse.2021.017016
Article

Tech Science Press

Deep Learning for Object Detection: A Survey

Jun Wang[1], Tingjuan Zhang[2,*], Yong Cheng[3] and Najla Al-Nabhan[4]

[1]Science and Technology Industry Department, Nanjing University of Information Science & Technology, 210044, China
[2]School of Computer & Software, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, 210044, China
[3]Science and Technology Industry Department, Nanjing University of Information Science & Technology, 210044, China
[4]Department Computer Science, King Saud University, Riyadh, KSA
[*]Corresponding Author: Tingjuan Zhang. Email: 20191221035@nuist.edu.cn
Received: 18 January 2021; Accepted: 19 February 2021

Abstract: Object detection is one of the most important and challenging branches of computer vision, which has been widely applied in people s life, such as monitoring security, autonomous driving and so on, with the purpose of locating instances of semantic objects of a certain class. With the rapid development of deep learning algorithms for detection tasks, the performance of object detectors has been greatly improved. In order to understand the main development status of target detection, a comprehensive literature review of target detection and an overall discussion of the works closely related to it are presented in this paper. This paper various object detection methods, including one-stage and two-stage detectors, are systematically summarized, and the datasets and evaluation criteria used in object detection are introduced. In addition, the development of object detection technology is reviewed. Finally, based on the understanding of the current development of target detection, we discuss the main research directions in the future.

Keywords: Object detection; convolutional neural network; computer vision

1 Introduction

In current years, more and more people pay attention to object detection, because of its wide application and technological innovation. This task has been studied in a wide range of academic and practical applications, such as surveillance safety, autonomous driving, traffic monitoring, and robot vision. With the enhancement of deep convolutional neural networks and the improvement of GPU computing speed, image object detection technology has been emerging rapidly [1]. In this day and age, deep learning models are applied to every area of computer vision. The backbone of deep learning networks is basically state-of-the-art object detectors, object detection is a kind of computer vision technology, and associated with image processing, it is used to detect images and videos, such as the face, trees, lights and so on. Multi-type detection, posture detection, face detection, and pedestrian detection is in the realm of image object detection research. We will also encounter devices that apply image object detection in our daily life. To date, benchmarks such as Caltech, ImageNet, PASCAL VOC [2], and MS COCO have played an important role in the area of object detection.

The category that can be identified from a picture is called object recognition, while object detection [3] needs to determine the location of the target in addition to the category. Image target detection has always been a challenging problem in computer vision, and it has always been a research field that people pay close attention to. The purpose of target detection is to determine whether an instance of a target exists in the image (e.g., car, bicycle, dog, cat, etc.). If so, the spatial location and scope of each target instance is returned. Since target detection is the foundation of image understanding and computer vision, it can also solve more problematical and higher-level visual tasks. Target detection is also applied to many areas of artificial intelligence, for example, driverless cars can drive safely on highways, and intelligent video surveillance can detect anomalies in video content. Different from the previous target detection surveys, this paper systematically and comprehensively reviews the object detection methods based on deep learning, and most importantly, studies the latest detection solutions and research trends.

Other parts of this paper are organized as follows: the second part summarizes the related background of target detection, including problems, the main challenges, and progress. The third part introduces the structure of the target detection. In the fourth part, the datasets and evaluation indexes for target detection are described. The fifth part describes the development of target detection technology. The sixth part reviews the typical target detection fields. In the seventh part, this paper is summarized, and further research directions are analyzed.

2 Background of Target Detection

General target detection (general target class detection), also known as target class detection or target class detection, focuses more on the detection of a wide range of natural categories, rather than a specific target category. Only narrow predefined categories, such as faces, pedestrians, or cars, may exist. Although thousands of objects dominate the visual world in which we live, the current research community is primarily interested in the location of highly structured objects, such as cars, faces, bicycles, and airplanes.

In general, the spatial location and scope of an object can be roughly defined through a bounding box. The four tasks of target detection are object classification, object localization, semantic segmentation, and object instance segmentation. As far as we know, boundary boxes are widely used in the current literature to evaluate common target detection algorithms. The future challenge, however, is at the pixel level, as the community moves toward a more detailed understanding of the scene (from image-level object classification to single-object positioning, common object detection, and pixelated object segmentation). Therefore, it is expected that there will be great challenges at the pixel level.

3 Structure of Target Detection

This chapter mainly describes the structure of target detection from two aspects: Two-stage framework and One-stage pipeline. One-stage is directly given by the trunk network category and location information, don't use the RPN (region proposals network). Two-stage is the extraction features of the convolution of the CNN during the target detection process through the convolutional neural network.

3.1 Two-Stage Framework

- RCNN

Regions with CNN Features (RCNN) is a region-based detector that can be described as a pioneer in deep learning for target detection. The process of the RCNN algorithm can be summarized into four steps. The first step is to determine 1,000 to 2,000 candidate boxes in the image by using the selective search method. The second step is to input each candidate box into CNN to extract features. The third

step is to classify the extracted features using a classifier to determine whether they belong to a specific class. Finally, the regression is used to adjust the position of the candidate boxes belonging to a particular feature.

There are obvious problems in RCNN: images corresponding to multiple candidate regions need to be extracted in advance, which takes up a large amount of disk space; Traditional CNN requires input images with fixed sizes, and the crop/ warp (normalized) operation will result in truncation or stretch of objects, which will lead to loss of information input to CNN. Each candidate region needs to be calculated in the CNN, and there is a large amount of scope overlap in thousands of regions, and repeated feature extraction brings huge calculation waste.

- SPP-Net

Since the feature extraction process of CNN is time-cosuming due to a large amount of convolution computation, why is it necessary to calculate each candidate region independently instead of extracting the overall feature, and only make a region interception before classification? SPP-Net was born.

This is shown in Fig. 1. SPP-Net has made two monumental refinements based on RCNN, one is the removal of crop/ warp image normalization process, which solves the information disappearance and storage problems caused by image malformation. Second, spatial pyramid pooling is adopted to replace the last pooling layer before the full connection layers. SPP is a pooling of multiple scales to obtain multi-scale information in the image [4]. After adding SPP into CNN, CNN can process the input of any size, which makes the model more flexible. Although SPP-Net makes great contributions, it still has many problems. First, just like RCNN, the training process is still independent, and the extraction of candidate boxes, calculation of CNN features, SVM classification, and bounding box regression are all independent training, with a lot of intermediate results needing to be saved and whole training parameters cannot be achieved. Second, SPP- Net cannot synchronously correct the Convolutional Layer and the full connection Layer on both sides of the SPP-layer, which limits the effect of DCNN (Deep Convolutional Neural Network) to a great extent.

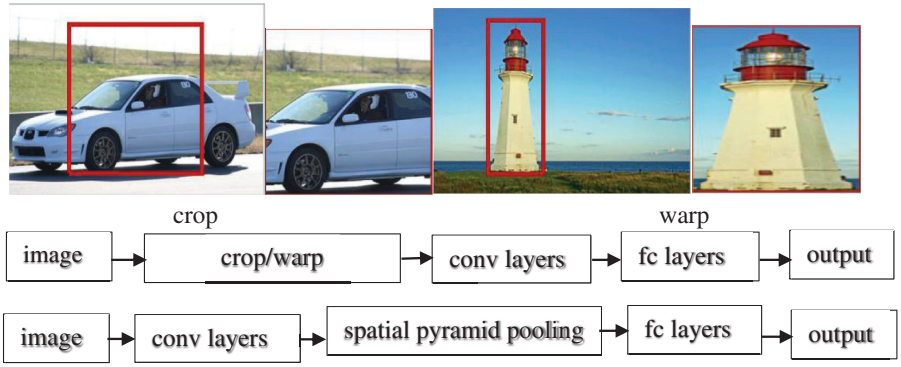

Figure 1: Upper: choosing the right size. Central: deep convolutional network structure. Bottom: structure after joining the space pyramid pooling

Third, the selection of the candidate box is still time-consuming during the whole process.

- Fast RCNN

Fast RCNN [5] can solve the disadvantages of RCNN and SPP-Net and improve its speed and accuracy at the same time. Compared with RCNN, SPP-Net and Fast RCNN have higher MAP (mean average precision), their training process is one-stage [6], and they use multi-task loss, all network layers share parameters, and they don't need disk space to temporarily serve as the feature cache.

The input of Fast R-CNN is composed of the whole image to be processed and the candidate region. The first step of Fast RCNN processing is to conduct multiple convolution kernel pooling for the image to obtain the convolution feature map. Since there are multiple candidate regions, the system will determine the Region of Interest. The role of the layer can be on any size of feature mapping for each input ROI area, fixed dimension feature extracting ROI pooling is SPP layer of special circumstances, it can be extracted from the characteristic figure a fixed-length feature vector, every feature vector can be transported to full connection layer in sequence, the FC branch into two output layers at the same level [7]. The function of one layer is to classify the target about K object classes (including all "background" classes) and output the probability distribution of each ROI, which is to generate softmax probability estimation. The other layer is to output four real values for each class of K objects, and each of the four values encodes the exact bounding box position of each class of K. The entire structure is end-to-end training using multitasking losses (excluding the Region Proposal extraction stage).

Although the speed and accuracy of Fast RCNN have been greatly improved, there are still many shortcomings, because of the Fast RCNN using selective search, this process is very time-consuming, the candidate area to extract the time it takes about 2 to 3 seconds, and extract the feature classification need only 0.32 seconds, which can cause cannot meet the demand of real-time applications, and because the use of selective search to extract candidate area in advance, Fast RCNN did not realize the true sense of the end-to-end training mode. Faster RCNN emerges as the times requires.

- Faster RCNN

Faster RCNN can be said to be composed of two modules: the RPN candidate box extraction module for the regional generation network and the Fast RCNN detection module. RPN is a full convolutive neural network, and its internal difference from the ordinary convolutive neural network is that the full connection layer in CNN is turned into a convolutive layer. Faster RCNN is to detect and identify the target in the proposal based on the extraction of RPN. The specific process can be roughly summarized in five steps. The first step is to input the image, the second step is to generate candidate regions through RPN, the third step is to extract the features, the fourth step is to classify the classifier, and the last step is to regression and adjusts the position of the regressor.

- Mask RCNN

The idea of Mask RCNN is also very concise, since Faster RCNN target detection effect is very good, each candidate region can output type label and the location information, then the Faster - RCNN add a branch to increase on the basis of an output, namely the object Mask (object mask), which is added on the original two tasks a segmentation task into three tasks. Mask RCNN combines binary Mask with the classification and boundary box from Faster RCNN to produce an amazing and accurate image separation, as shown in Fig. 2 below, Mask RCNN is a malleable and universal object instance segmentation framework, it not only detects the target in the image, but also outputs a high-quality segmentation result for each target. In addition, Mask R-5CNN is also easy to generalize to other tasks, such as character key point detection.

3.2 Unified Pipeline (One-Stage Pipeline)

- YOLO

YOLO is an innovation in the field of object detection in recent years. RCNN series framework also has many disadvantages, such as the whole network cannot do end-to-end training, the intermediate training process needs a lot of memory to store some features, the calculation speed is slow, etc. YOLO algorithm puts forward a new idea, which transforms the object detection problem into a regression problem. Given

an input image, it directly returns the target bounding box and its classification categories in multiple locations of images.

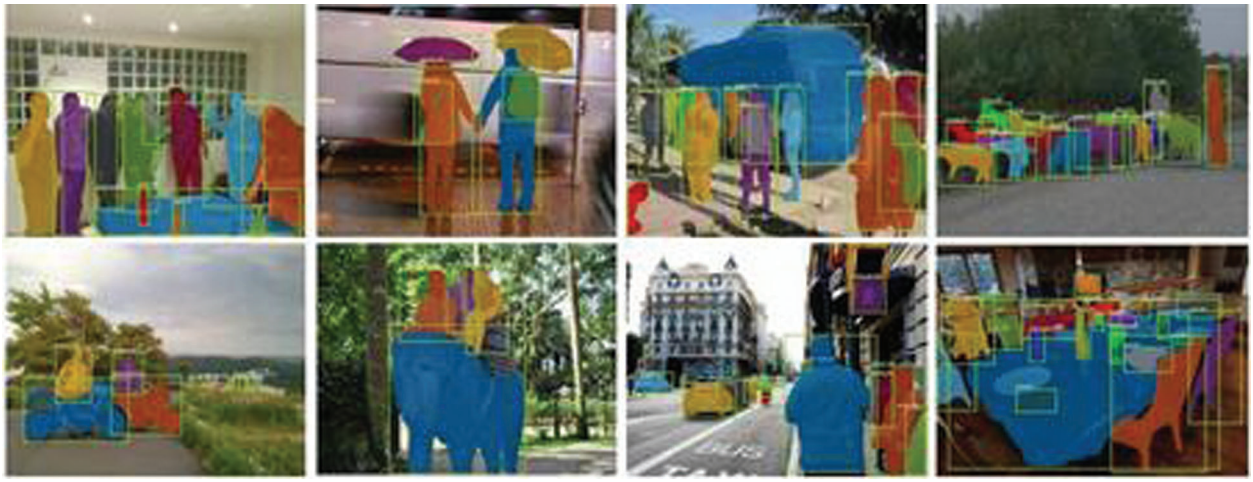

Figure 2: Mask RCNN results on the COCO test set

Compared with other object detection algorithms, YOLO detection speed is very fast. The faster YOLO detection speed can reach 155 FPS. Different from other object detection algorithms, the input of YOLO is a whole picture, which makes good use of the overall information during detection, and the low probability will predict the wrong object information on the background. YOLO can learn highly generalized characteristics and have better mobility. However, the accuracy of object detection is not optimal, and it is easy to produce positioning errors. And because a Grid cell can only predict two objects, it does a poor job of detecting small ones.

- SSD

SSD is an end-to-end, step by step operation mode, which is similar to YOLO but different from Faster RCNN, so the speed is similar to that of YOLO, but higher than that of Faster RCNN. This relatively high speed has strong real-time performance and can be applied to many occasions. In terms of accuracy, SSD uses different feature layers for detection (multi-scale), so it can accommodate many profiles of problems of different status and different sizes. Therefore, the effect of SSD is much better than that of YOLO in the inspection of small objects. The overall accuracy is close to or even better than Faster RCNN. So, it's a combination of speed and precision.

4 Datasets and Indicators of Evaluation

Using challenging datasets as benchmarks is important in many areas of research because they allow constant comparisons between different algorithms to select an algorithm suitable for the current solution. Early on, they preferred to use specific datasets about face detection, and later they slowly created more datasets containing pedestrians.

4.1 Figures Introduction to Datasets

VOC dataset is a commonly used dataset for object detection. It includes VOC2007 and V0C2012 [8] version. The former consists of 9,963 labeled images, consisting of training set, validation set, and test set, with a total of 24,640 objects labeled. The latter is an updated version of the former dataset, with

11,530 images. For detection tasks, VOC2012's training set, validation set, and test set contain all corresponding images from 2008 to 2011. The training set and verification set have 11540 images and 27450 objects. For the split task, VOC2012's training set and validation set contain all corresponding images from 2007 to 2011, while the test set contains only images from 2008 to 2011. The training set and verification set have 2913 images with 6929 objects. These two datasets mainly annotate many common objects in life, such as pedestrians, animals, vehicles, and so on.

MS-COCO [9] is by far the most challenging object detection dataset. COCO dataset is a tremendous and affluent object detection dataset. This dataset is primarily used to understand scenarios that are captured from complex daily scenarios. The object is adjusted by precise segmentation in the image. The image contains 91 different targets, 328,000 images, and 2.5 million tags. We have by far the largest set of semantically split data. There are 80 categories and more than 330,000 images, 200,000 of which have been tagged. The total number of individuals is more than 1.5 million in the dataset.

ImageNet is a computer vision system recognition project, and is the world's largest image recognition database. ImageNet is a computer scientist at Stanford University in the United States. He has simulated the human cognitive system. Objects can be distinguished from photographs. The ImageNet dataset is well documented, managed by a dedicated team, and is easy to use and widely used in research papers in the field of computer vision. It is becoming the "standard" dataset for performance testing algorithms in the current deep learning image domain. The ImageNet dataset has over 14 million photos, covering over 20,000 categories. More than a million of these images have clear category tags and object location tags in the image.

DOTA is a commonly used aerial image detection dataset in remote sensing. It contains 2,806 aerial images of approximately 4K $\times$ 4K in size. It contains 15 categories and 188,282 instances, of which 14 are the main categories. Both small and large vehicles are subclasses of vehicles. The marking method is a quadrilateral with arbitrary shape and direction determined by four points. Different from the traditional datasets, aerial photography has the characteristics of large-scale variability, small target intensive detection, and uncertainty. The data is divided into 1/6 validation set, 1/3 test set, and 1/2 training set. Training sets and validation sets are currently available, with image sizes ranging from 800 $\times$ 800 to 4000 $\times$ 4000.

4.2 Evaluation Index

4.2.1 Accuracy

Accuracy is the ratio of the correct samples to all samples, which is generally used to evaluate the accuracy of the detection model. The information it contains is limited, so it is impossible to comprehensively evaluate the performance of the model.

4.2.2 Confusion Matrix

The obfuscation matrix is a matrix drawn with the predicted number of categories on the horizontal axis and the actual number of labels on the vertical axis. Since the diagonals represent the number of consistent model predictions and data labels, it is also possible to calculate the sum of accuracy divided by the number of images in the diagonal test set of confounding matrices. The higher the number on the diagonal, the darker the color confuses the matrix visualization results and the better the model predicts the results of this class. Other places, of course, are those that have been mispredicted. The smaller the natural value, the lighter the color, the better the prediction.

4.2.3 Precision, Recall and PR curves

A classic example is the existence of a test set consisting only of basketball and football images, assuming that the terminal goal of the classification system is to extract all football images in the test set, rather than basketball images. Then you can define:

True Positives: TP for short, that is, the positive sample is correctly identified as the positive sample, and the picture of the football is correctly identified as the football.

True negatives: TN for short, negative samples are correctly identified as negative samples, basketball pictures are not identified, the system correctly thinks they are basketball.

False Positives: FP for short, a negative sample is misidentified as a positive sample, and a picture of a basketball is misidentified as a football.

False negatives: FN for short, the positive samples are wrongly identified as the negative samples, the pictures of the football are not identified, and the system mistakenly thinks they are basketballs.

Precision is the percentage of the images that you identify, where True can flourish. That is, the proportion of all the recognized footballs in this hypothesis, that are real football.

The recall rate is the proportion of all positive samples in the test set that are correctly identified as positive samples. That is the ratio of the number of correctly identified balls in this hypothesis to the number of real balls in the test set.

In the PR curve, P stands for precision and R stands for recall, which represents the relationship between precision and recall. In general, recall is set as the x-coordinate, and precision is set as the y-coordinate.

4.2.4 Average Precision (AP) and Mean Average Precision (MAP)

In object detection, a curve can be drawn for each category by using recall rate and accuracy, where average Precision AP is the area of the curve, and MAP is the average value of each category obtained by AP. MAP is used to judge the detection accuracy. The higher the MAP value is, the better the performance of the detector will be.

5 Development of Object Detection Technology

In this chapter, we will introduce the development of object detection technology over the years.

5.1 Technical Evolution of Multi-Scale Detection

The “different size” and “different aspect ratio” of the object of multi-scale detection is one of the preeminent technical problems of target detection. As shown in Fig. 3, it has gone through many indispensable historical periods.

First, the method of feature pyramid and sliding window is introduced. HOG detector is facing with a fixed length-width ratio of objects, such as face and stand of pedestrian, by building the characteristics of the pyramid, on which the sliding window fixed size detection, then to test more complex view object (e.g., PASCAL VOC), considered the detection of various aspect ratio, this paper proposes the methods of mixture model, it is through the training multiple models to detect objects with different aspect ratio. When mixed models and exemplar-based approaches are introduced, more complex detection models are proposed and we wonder if they are going to be a method for detecting objects with different aspect ratios, and the emergence of “Object Proposals” answers that question.

Object proposals are applied to target detection for the first time in 2010 and refer to a set of candidate boxes that may contain any object unrelated to a class. Using object proposals to detect the model can avoid a thorough sliding window search of the image. The detection algorithm for object proposals needs to meet three points, namely high recall rate, high positioning accuracy and improved accuracy while reducing processing time. The current regional detection methods can be divided into three categories, namely, segmentation and grouping method [10–13], window scoring method [14–16], and neural network-based method [17–21]. Later, we began to think about the role of object proposal in the detection of models. Was it to improve the accuracy or speed up the detection? To settle this problem, researchers started from

two aspects, one was to weaken the effect of the proposal, the other was to conduct sliding window detection on CNN features. However, the final results were not satisfactory. Then, with the rise of unipolar detectors and deep regression techniques, regional detection gradually faded out of sight.

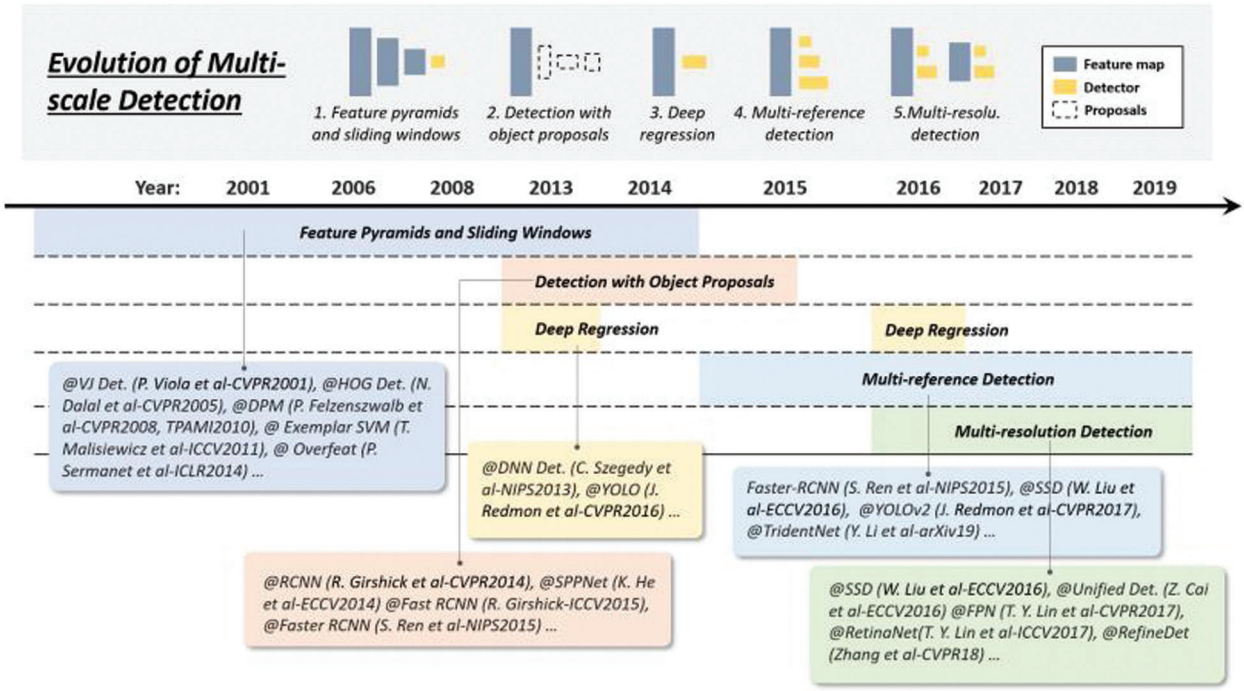

Figure 3: Development of multi-scale detection techniques in object detection

People direct prediction of coordinates of bounding boxes based on deep learning features is a method of using deep regression to deal with multi-scale problems. This method has both advantages and disadvantages. On the one hand, it is simple and easy to operate, while on the other hand, the positioning is not accurate enough, especially for some small objects followed by multi-reference detection to solve this problem. Soon after, multi-reference testing solved this problem.

At present, the commonly used detection frameworks are multi-reference detection and multi- resolution detection. The former is mainly to select a set of anchor frames of different sizes and aspect ratios at different locations of images in advance, and then to predict detection frames based on these anchor frames. The latter is mainly to detect instances of different scales at different layers of the network. CNN naturally formed a characteristic pyramid in the process of forwarding propagation, making it easy to find large objects in the deep layer and small objects in the shallow layer. Multi-reference and multi-resolution detection are two indispensable components of the most advanced object detection system.

5.2 Technical Evolution of Bounding Box Regression

Boundary box regression is an indispensable target detection technique. Its purpose is to adjust the position of the predicted bounding box according to the initial proposal or anchor box. Fig. 4 shows the evolution process of bounding box regression.

The HOG detector mentioned above does not use BB regression and generally takes the sliding window as the result of the detection. If you want to get the precise target position, you can only build a very dense feature pyramid and slide the detector at each position. Then BB regression was introduced into the object

detection system as a post-processing block. When Faster RCNN is introduced, BB regression is no longer treated as a separate processing block but conducts training with end-to-end integration with the detector.

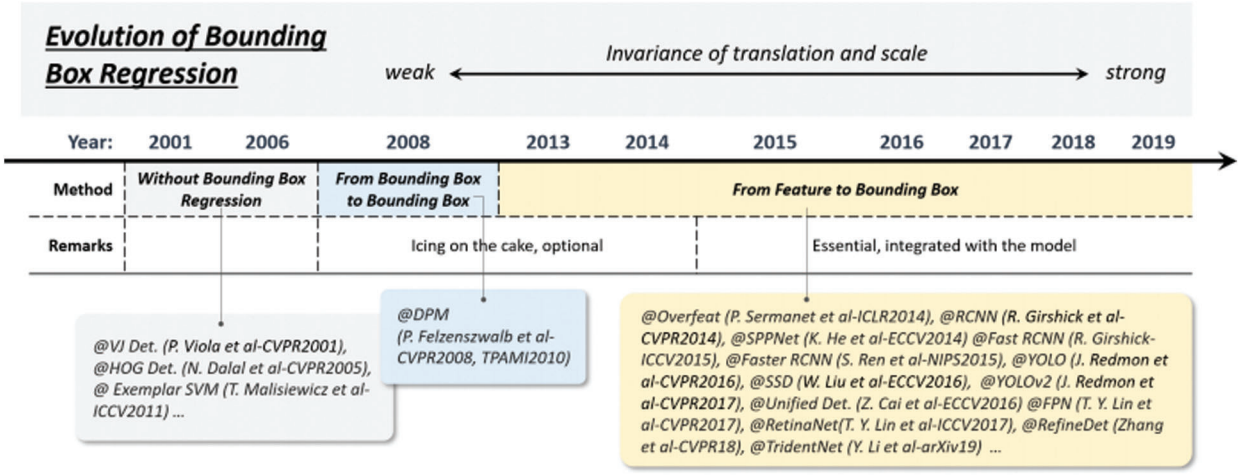

Figure 4: Evolution of bounding box regression techniques in object detection

5.3 Technical Evolution of Context Priming

Methods to improve detection have always had context priming. During its evolution, there are three commonly used methods: 1) local context detection, 2) global context detection, and 3) context interaction, as shown in Fig. 5.

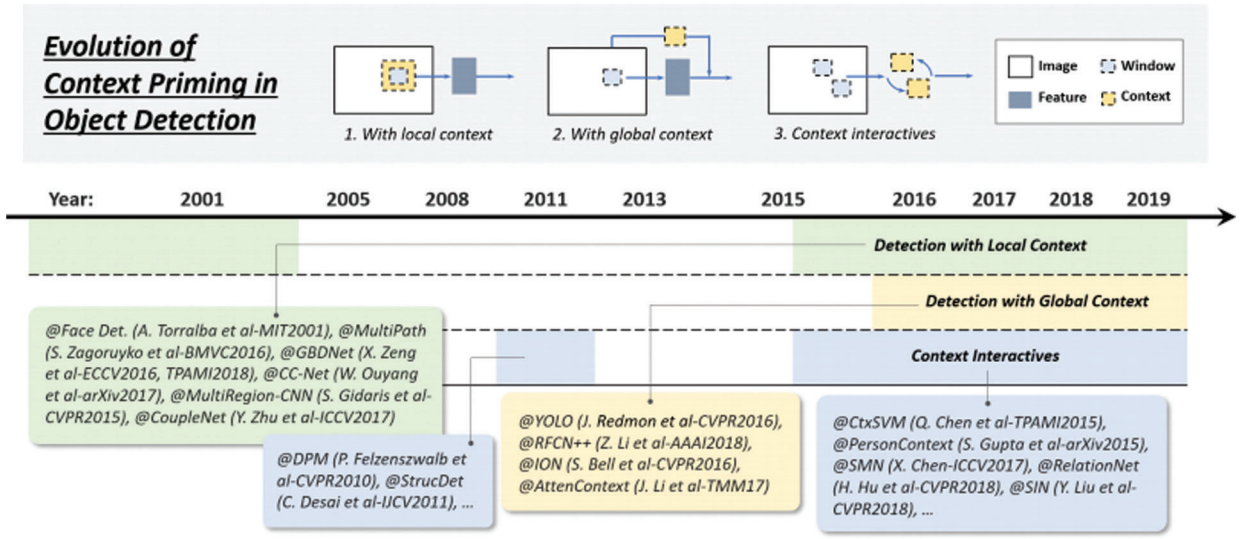

Figure 5: Evolution of context priming in object detection: 1) detection with local context, 2) detection with global context, 3) detection with context interactives

Local context refers to the visual information near the target to be detected. It has always been thought that local context can improve the performance of target detection. It has been found that local context including facial boundary contours can significantly improve the performance of face detection. It is also found that adding some background information can improve the accuracy of pedestrian detection, and

the detector based on deep learning is also improved according to the local context. A global context is additional information that detects scenarios as objects. In existing detectors, there are two ways to integrate the global context. The first is to utilize the global pooling operation of the large acceptance domain or the CNN feature [22]. The other is to treat the global context as a sequence of information and use recursive neural networks to learn it [23,24]. Context interaction refers to the transfer of information through constraints and dependencies between visual elements. Research shows that context interaction can improve the performance of the target detector. Exploring relationships between individual objects [25,26] and exploring dependencies between modeled objects and scenarios are two of the most recently improved categories.

5.4 Technical Evolution of Non-Maximum Suppression

Non-maximum suppression (NMS) mainly removes some candidate boxes whose IOU value is greater than a certain threshold. Since the detection scores of adjacent windows are generally similar, NMS is adopted as a post-processing step in this paper to remove repeated boundary boxes and get the final detection result. Over the past 20 years, NMS has evolved into three approaches: 1) greedy selection, 2) bounding box aggregation, and 3) Learning NMS, as shown in Fig. 6.

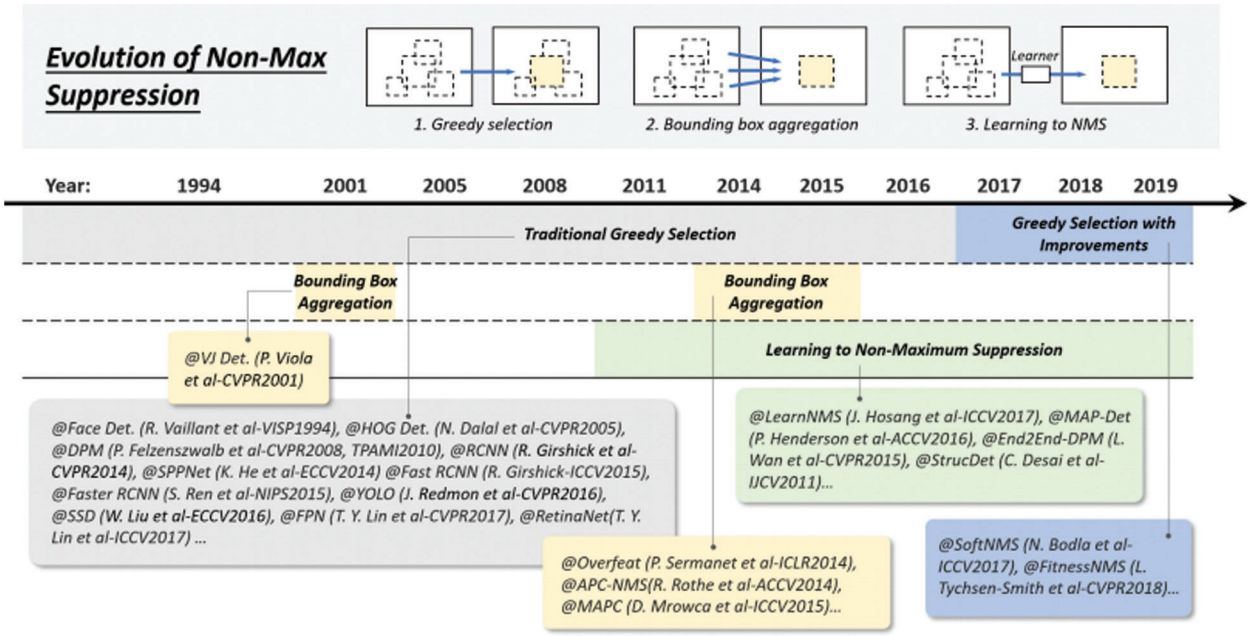

Figure 6: Evolution of non-max suppression (NMS) techniques in object detection from 1994 to 2019: 1) Greedy selection, 2) Bounding box aggregation, and 3) Learn to NMS

Greedy selection is the most famous object detection method nowadays. Its idea is as follows: There is a set of overlapping checks, the largest bounding box is selected in the check score, and then some adjacent boxes that are greater than a certain threshold is removed and executed iteratively. There is still room for improvement in the greedy selection, first, the box with the highest score is not necessarily the most appropriate, second, it may inhibit the surrounding objects, and third, it does not inhibit false positives. But the greedy choice is by far the strongest baseline for object detection. BB aggregation is another technique for NMS, which combines or clusters several overlapping bounding boxes into a final detection, with the advantage of fully considering the object relationship and its spatial layout. Finally,

the NMS technology is studied. The main idea is to treat the NMS as a filter, re-test, and score all the tested prediction boxes, and then train the NMS as a part of the network in an end-to-end manner. Compared with the traditional NMS method, learning NMS is more beneficial to improve occlusion and dense targets.

6 Applications

In this section, we will review some important detection applications in the past, including pedestrian detection, face detection, text detection, traffic sign and traffic light detection, and remote sensing object detection.

6.1 Pedestrian Detection

Pedestrian detection is a crucial problem in computer vision. It has many applications in life, which can improve the quality of people's life [27–29]. Early pedestrian detection method had HOG detector [30]. For example, the object detection algorithm of Faster RCNN has been successfully applied to pedestrian detection, which is very beneficial to the development of this field.

The challenges and difficulties encountered in pedestrian detection [31] can be summarized into three major parts, namely, Small pedestrian, Hard negatives, Dense and occluded pedestrian. Fig. 7 is an example of pedestrian detection in Caltech [32,33] dataset. Fig. 7a is a pedestrian image taken in the far lens with small pixels and difficult to recognize. Fig. 7b shows some of the background scenes of the street scene which are indistinguishable from pedestrians. Fig. 7c shows that when the image is crowded with pedestrians, pedestrians close to the lens will block other pedestrians. In the Caltech dataset, unblocked pedestrians only account for 29% of the total number of pedestrians.

Figure 7: Some hard examples of pedestrian detection from Caltech dataset: (a) small pedestrians, (b) hard negatives, and (c) dense and occluded pedestrians

In Fig. 7a, the detection effect of small pedestrians is general. Recent approaches to solving this problem include feature fusion, the introduction of additional high-resolution manual features, and integrated detection results based on multiple resolutions. For Hard Negative detection in Fig. 7, some recent developments include enhanced decision tree integration and semantic segmentation [34]. In addition, the idea of "cross-modal learning" is also introduced, and RGB images and infrared images are used to enrich the features of difficult negative samples [35]. As we know, the convolutional neural network (CNN) has richer semantics for deep features but has a poor effect on dense detection objects. Moreover, target occlusion is also one of the problems frequently encountered in dense pedestrian detection. So far, methods to improve target occlusion have been proposed, such as Integration of local detectors [36,37] and Attention Mechanism [38].

6.2 Face Detection

Face detection is a prototypical problem in the field of machine vision. It has important application value in security monitoring, witness comparison and human-computer interaction. Face detection technology has been widely used in digital cameras, smartphones, and other devices to perform functions.

The obstructions in face detection can be outlined into three points, namely, intra-class variation, occlusion, and multi-scale detection. As can be seen from Fig. 8a, people's faces can have many kinds of expressions and actions, as well as different skin tones. Face detection and pedestrian detection mentioned above meet the same occlusion challenge. In a dense scene, the face is likely to be blocked by other objects, as shown in Fig. 8b. In another case, face with multiple proportions should be detected, especially some remote and small faces, as shown in Fig. 8c.

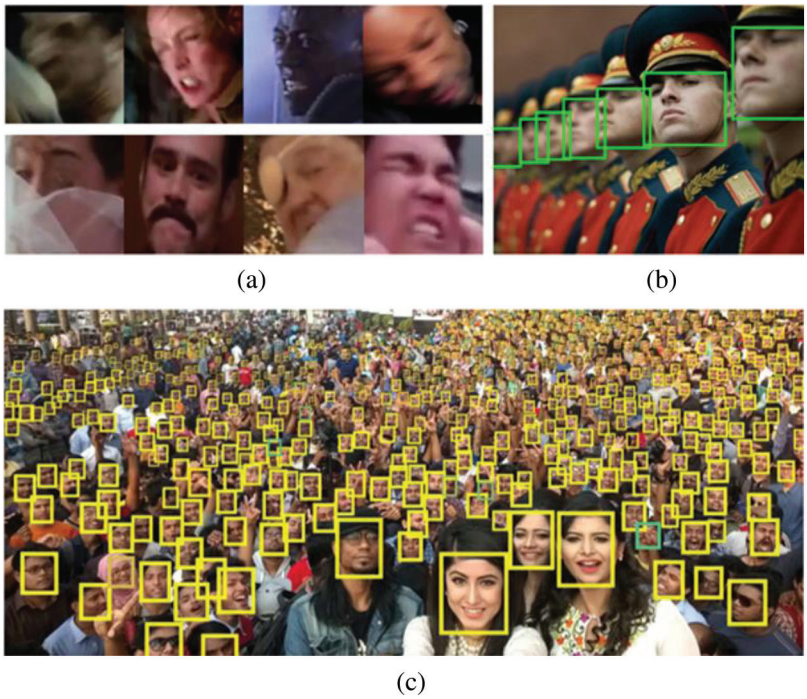

Figure 8: Challenges for face detection

Cascaded Detection is the most commonly used method to speed up face detection in the era of deep learning [39,40]. Another method of acceleration is to predict the proportional distribution of faces [41] in the image and then detect them at a selected ratio. "Face calibration" improves a multi-gesture face detection by estimating the calibration parameter [42], or by using progressive calibration over multiple detection stages [43]. Recently, two methods are proposed to improve the detection of occluded faces. The first combines "attention mechanisms" [44] to highlight the characteristics of potential facial targets. The second type is "partial based detection" [45], which inherits the concept of DPM. In recent years, research on multi-scale face detection [46–51] adopts detection strategies similar to general object detection, such as multi-scale feature fusion and multi-resolution detection.

6.3 Text Detection

The text has always been the main information carrier for human beings. The basic goal of text detection is to determine whether a particular image contains text and if so, to locate and recognize it. Text detection

has a wide range of applications. Help visually impaired people "read" street signs and currency [52,53]. Geographic information systems facilitate the creation of digital maps by detecting and identifying house numbers and road signs [54,55].

The difficulties and challenges of Text detection can be summarized as four points,namely, different fonts and languages, text rotation and perspective distortion, dense Text localization, and Dense Text. Font sizes and colors in the text may vary, and multiple languages may appear. The direction of the text may be different, and it may be distorted by perspective. Text lines with aspect ratios and dense layouts are difficult to position accurately. It is also common for text in street view images to be corrupted or blurred.

For text rotation and perspective, the most common solution is to introduce additional parameters related to rotation and perspective changes to the anchor frame and ROI pooling layer. The segmented approach is more advantageous when it comes to detecting densely arranged text. Two sets of solutions were recently proposed to distinguish between adjacent lines of text. The first group is "Fragments and links". Where "segment" refers to a character heat map, and "link" refers to a connection between two adjacent line segments, indicating that they belong to the same word or text line [56,57]. The second group introduces additional boundary detection tasks to help separate densely arranged text, where a set of angles or closed boundaries corresponds to a line of text. A recent approach to dealing with incomplete and fuzzy text is to use word-level recognition and sentient-level recognition [58–60]. The most effective way to deal with text with different fonts is to train it with a composite sample.

6.4 Traffic Sign and Traffic Light Detection

The detection of traffic signs and traffic lights is of great significance to the safe driving of autonomous vehicles. In the complex urban environment, the detection and identification of traffic lights is always a difficult problem. With the help of deep learning technology, the recognition effect of traffic lights has been greatly improved. However, in the complex urban environment, the detection of road traffic signals is still not very accurate. The challenges and difficulties in the detection of traffic signs and lights can be summed up in three aspects, namely, illumination changes, motion blur, and bad weather. Detection can be difficult when vehicles are driven in bright lights or at night. As shown in Fig. 9a. When the vehicle is moving rapidly, the photos captured by the on-board camera will become blurred, as shown in Fig. 9b. In the case of rain and snow, the photos taken will become unclear due to the shelter of rain and snow, as shown in Fig. 9c. In the era of deep learning, there are some famous detectors used in traffic sign and light detection tasks [61,62], such as Faster RCNN and SSD. Based on these detectors, new techniques such as attention mechanisms and confrontational training have been used to improve detection in complex traffic environments.

6.5 Remote Sensing Object Detection

Automatic detection of remote sensing targets is not only an intelligent data analysis method to realize automatic classification and location of remote sensing targets, but also an important research direction in the field of remote sensing image interpretation. The traditional remote sensing image target detection method is based on artificial experience design. In certain application scenarios, better detection results can be obtained, but this method relies on prior knowledge, resulting in poor adaptability and generalization of the detection model. Multi-scale deep convolutional neural network (MSCNN), on the other hand, uses a deep convolutional neural network that can actively learn features from data without recourse to the human experience.

The challenges and difficulties of remote sensing target Detection can be summarized as three points, namely, detection in "big data", occluded targets, and domain adaptation. It is still a big problem of how to quickly and accurately detect remote sensing targets in the case of a huge amount of remote sensing image data. Fig. 10a shows the amount of remote sensing images and natural image data. As shown in

Fig. 10b, the surface of the earth will be covered by a lot of clouds every day, so there will be many remote sensing images blocked by the target. The remote sensing images taken by different sensors will vary greatly.

Figure 9: Challenges in traffic sign detection and traffic light detection

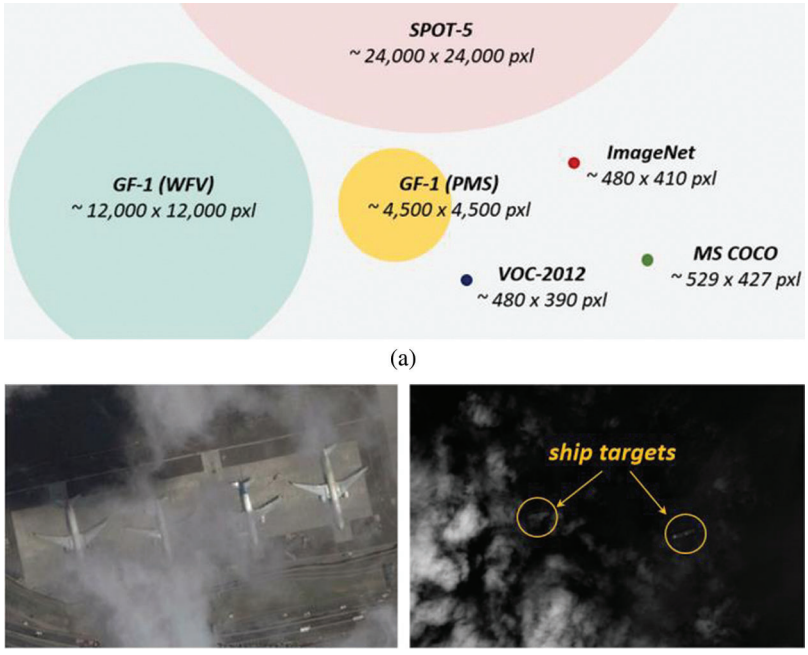

Figure 10: The challenge of remote sensing target detection

7 Summary and Analysis of Further Research Directions

Object detection is an important and challenging problem in computer vision, which has been widely concerned by people. With the advancement of deep learning technology, great changes have taken place in the field of target detection. This paper gives a systematic overview of various target detection methods, including one-stage and two-stage detectors, and introduces the dataset and evaluation criteria used in target detection. In addition, the development of target detection technology is reviewed, and the traditional and new application fields are listed. Future research on target detection may focus on the following aspects:

Lightweight object detection: Not only does it work stably on mobile devices, but it also significantly shortens working hours. It has applications in smart cameras and facial recognition. However, when detecting targets, the speed between the machine and human eyes is still very different, especially when detecting relatively small objects.

Video object detection: In the video target detection, many situations make the detection task obtain high precision, such as fast motion makes the target fuzzy, the video is out of focus, the target is small, occlusion, and so on. Future research will focus on sports goals and more complex data.

Weak supervised detection: The training of detectors based on deep learning usually relies on a large number of annotated images. The annotation process is time-consuming, expensive, and inefficient. Weak supervised detection technology only uses image level annotation or part of the boundary box annotation to train the detector, which can not only reduce the cost but also get a more accurate model.

Small-object detection: Detecting small objects in images has always been a challenge. Future applications may include the integration of visual attention mechanisms and the design of high-resolution lightweight networks in this direction.

Acknowledgement: The author would like to thank the researchers in the field of object detection and other related fields. This paper cites the research literature of several scholars. It would be difficult for me to complete this paper without being inspired by their research results. Thank you for all the help we have received in writing this article.

Funding Statement: This work was supported National Natural Science Foundation of China (Grant No.41875184) and innovation team of "Six Talent Peaks" in Jiangsu Province (Grant No.TD-XYDXX- 004).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. Malisiewicz, A. Gupta and A. A. Efros, "Ensemble of exemplar-svms for object detection aeyond," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Barcelona, Spain, pp. 89–96, 2011.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[3] Y. Lee, H. Ahn, H. B. Ahn and S. Y. Lee, "Visual object detection and tracking using analytical learning approach of validity level," *Intelligent Automation & Soft Computing*, vol. 25, no. 1, pp. 205–215, 2019.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] R. Girshick, "Fast r-cnn," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1440–1448, 2015.

[6] Q. Zhu, M. C. Yeh, K. T. Cheng and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, New York, NY, USA, pp. 1491–1498, 2006.

[7] S. Maji, A. C. Berg and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, pp. 1–8, 2008.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[9] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona *et al.,* "Microsoft coco: Common objects in context," in *Proc. of the European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 740–755, 2014.

[10] K. E. A. Van de Sande, J. R. R. Uijlings, T. Gevers and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Colorado, Springs, USA, pp. 1879–1886, 2011.

[11] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[12] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 3241–3248, 2010.

[13] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques and J. Malik, "Multiscale combinatorial grouping," in *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, pp. 328–335, 2014.

[14] B. Alexe, T. Deselaers and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[15] M. M. Cheng, Z. Zhang, W. Y. Lin and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, pp. 3286–3293, 2014.

[16] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. of the European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 391–405, 2014.

[17] C. Szegedy, S. Reed, D. Erhan, D. Anguelov and S. Ioffe, "Scalable, high-quality object detection," in *arXiv preprint arXiv*, 1412.1441, 2014.

[18] D. Erhan, C. Szegedy, A. Toshev and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, pp. 2147–2154, 2014.

[19] W. Kuo, B. Hariharan and J. Malik, "Deepbox: Learning objectness with convolutional networks," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 2479–2487, 2015.

[20] S. Gidaris and N. Komodakis, "Attend refine repeat: Active box proposal generation via in-out localization," in *arXiv preprint arXiv, 1606.04446*, 2016.

[21] H. Li, Y. Liu, W. Ouyang and X. Wang, "Zoom out-and-in network with recursive training for object proposal," in *arXiv preprint arXiv, 1702.05711*, 2017.

[22] Z. Li, Y. Chen, G. Yu and Y. Deng, "R-fcn++: Towards accurate region-based fully convolutional networks for object detection," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018.

[23] G. Yang, J. Zeng, M. Yang, Y. Wei and X. Wang, "Ott messages modeling and classification based on recurrent neural networks," *Computers, Materials & Continua*, vol. 63, no. 2, pp. 769–785, 2020.

[24] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu *et al.,* "Attentive contexts for object detection," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 944–954, 2017.

[25] 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, 2010.

[26] C. Desai, D. Ramanan and C. C. Fowlkes, "Discriminative models for multi-class object layout," *International Journal of Computer Vision*, vol. 95, no. 1, pp. 1–12, 2011.

[27] J. Cao, Y. Pang and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3210–3220, 2017.

[28] Y. M. Wang, K. B. Jia and P. Y. Liu, "Impolite pedestrian detection by using enhanced YOLOv3-Tiny," *Journal on Artificial Intelligence*, vol. 2, no. 3, pp. 113–124, 2020.

[29] Q. Hu, P. Wang, C. Shen, A. V. D. Hengel and F. Porikli, "Pushing the limits of deep cnns for pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1358–1368, 2018.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, pp. 886–893, 2005.

[31] L. Zhang, L. Lin, X. Liang and K. He, "Is faster R-CNN doing well for pedestrian detection?," in *Proc. of the European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 443–457, 2016.

[32] P. Dollár, C. Wojek, B. Schiele and P. Perona, "Pedestrian detection: A benchmark," in *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition*, Miami, Florida, pp. 304–311, 2009.

[33] P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[34] Y. Tian, P. Luo, X. Wang and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 5079–5087, 2015.

[35] D. Xu, W. Ouyang, E. Ricci, X. Wang and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Hawaii, HI, USA, pp. 5363–5371, 2017.

[36] Y. Tian, P. Luo, X. Wang and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1904–1912, 2015.

[37] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan *et al.,* "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1874–1887, 2018.

[38] S. Zhang, J. Yang and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6995–7003, 2018.

[39] H. Li, Z. Lin, X. Shen, J. Brandt and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 5325–5334, 2015.

[40] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[41] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li *et al.,* "Scale-aware face detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Hawaii, HI, USA, pp. 6186–6195, 2017.

[42] X. Shi, S. Shan, M. Kan, S. Wu and X. Chen, "Real-time rotation-invariant face detection with progressive calibration networks," in *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 2295–2303, 2018.

[43] D. Chen, G. Hua, F. Wen and J. Sun, "Supervised transformer network for efficient face detection," in *Proc. of the European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 122–138, 2016.

[44] S. Yang, P. Luo, C. C. Loy and X. tang, "Faceness-net: Face detection through deep facial part responses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1845–1859, 2018.

[45] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Hawaii, HI, USA, pp. 951–959, 2017.

[46] S. Yang, Y. Xiong, C. C. Loy and X. Tang, "Face detection through scale-friendly deep convolutional networks," in *arXiv preprint arXiv, 1706.02863*, 2017.

[47] M. Najibi, P. Samangouei, R. Chellappa and L. S. Davis, "Ssh: Single stage headless face detector," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 4875–4884, 2017.

[48] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang *et al.,* "S3fd: Single shot scale-invariant face detector," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 192–201, 2017.

[49] H. Nada, V. A. Sindagi, H. Zhang and V. M. Patel, "Pushing the limits of unconstrained face detection: a challenge dataset and baseline results," in *Proc. of the IEEE Int. Conf. on Biometrics Theory, Applications and Systems*, Los Angeles, California, USA, pp. 1–10, 2018.

[50] M. K. Yucel, Y. C. Bilge, O. Oguz, N. Ikizler-Cinbis, P. Duygulu *et al.,* "Wildest faces: Face detection and recognition in violent settings," in *arXiv preprint arXiv, 1805.07566*, 2018.

[51] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Hawaii, HI, USA, pp. 951–959, 2017.

[52] X. Liu, "A camera phone based currency reader for the visually impaired," in *Proc. of the Int. ACM SIGACCESS Conf. on Computers and Accessibility*, New York, NY, USA, pp. 305–306, 2008.

[53] N. Ezaki, K. Kiyota, B. T. Minh, M. Bulacu and L. Schomaker, " Improved text-detection methods for a camera-based text reading system for blind persons," in *Proc. of the IEEE Int. Conf. on Document Analysis and Recognition*, Seoul, South Korea, pp. 257–261, 2005.

[54] P. Sermanet, S. Chintala and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," in *Proc. of the IEEE Int. Conf. on Pattern Recognition*, Tsukuba, Japan, pp. 3288–3291, 2012.

[55] Z. Wojna, A. N. Gorban, D. S. Lee, K. Murphy, Q. Yu *et al.,* "Attention-based extraction of structured information from street view imagery," in *Proc. of the IEEE IAPR Int. Conf. on Document Analysis and Recognition*, Kyoto, Japan, pp. 844–850, 2017.

[56] S. Niu, X. Li, M. Wang and Y. Li, "A Modified Method for Scene Text Detection by ResNet," *Computers, Materials &Continua*, vol. 65, no. 3, pp. 2233–2245, 2020.

[57] U. Yasmeen, J. H. Shah, M. A. Khan, G. J. Ansari, S. U. Rehman *et al.,* "Text detection and classification from low quality natural images," *Intelligent Automation & Soft Computing*, vol. 26, no. 4, pp. 1251–1266, 2020.

[58] Y. Wu and P. Natarajan, "Self-organized text detection with minimal post-processing via border learning," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 5000–5009, 2017.

[59] C. Xue, S. Lu and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in *Proc. of the European Conf. on Computer Vision*, Munich, Germany, pp. 355–372, 2018.

[60] P. Lyu, C. Yao, W. Wu, S. Yan and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7553–7563, 2018.

[61] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing and C. Lgel, "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark," in *Proc. of the IEEE Int. Joint Conf. on Neural Networks*, Dallas, TX, pp. 1–8, 2013.

[62] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li *et al.,* "Traffic-sign detection and classification in the wild," in *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 2110–2118, 2016.