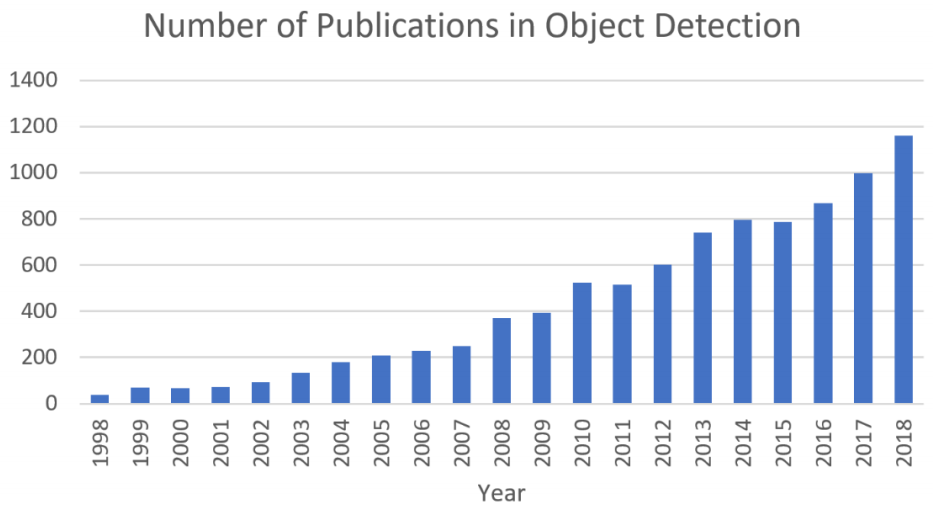
Object Detection in 20 Years: A Survey

目标检测的20年综述

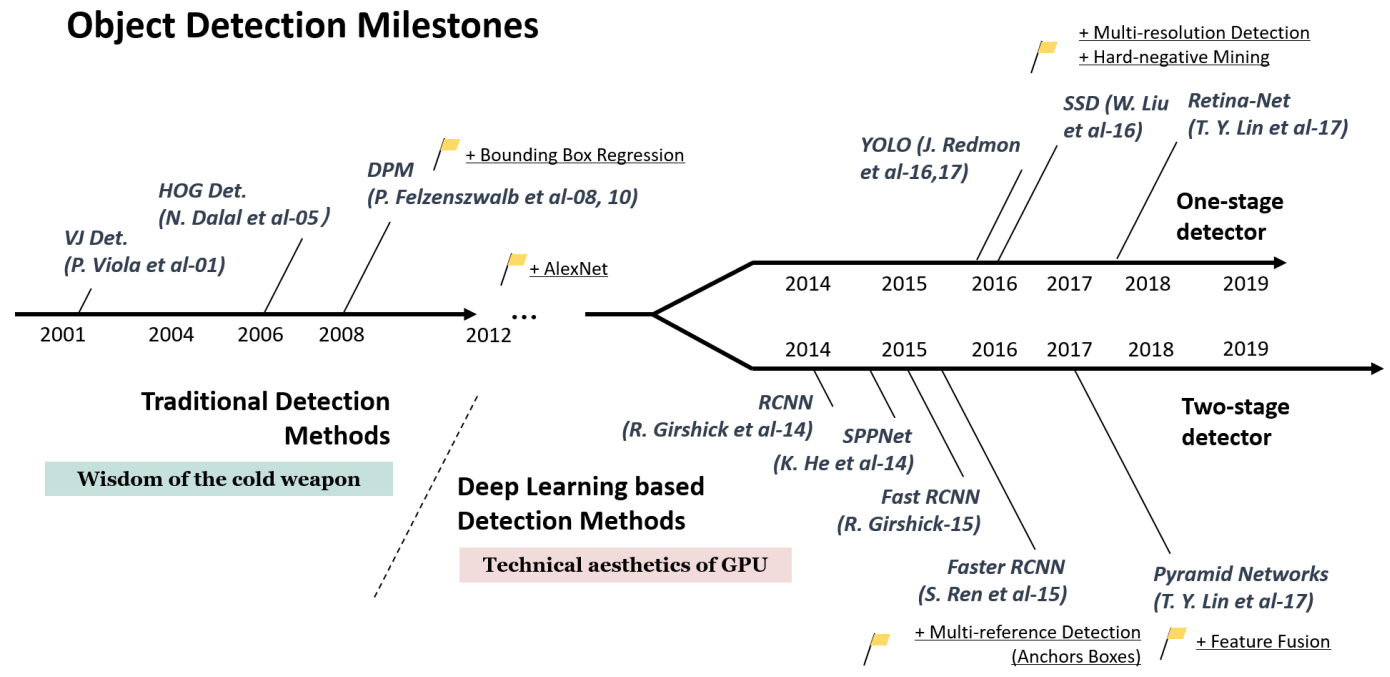
论文引用：

Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. (May. 2019). “Object Detection in 20 Years: A Survey.” [Online]. Available: https://arxiv.org/abs/1905.05055

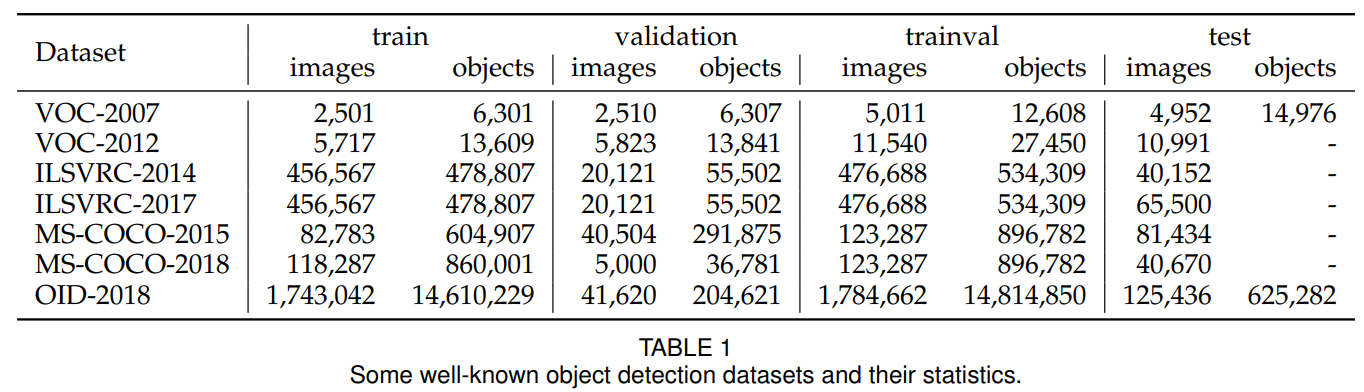
|  |  |
| --- | --- |
| **ABSTRACT**  Object detection, as of one the most fundamental and challenging problems in computer vision, has received great attention in recent years. Its development in the past two decades can be regarded as an epitome of computer vision history. If we think of today’s object detection as a technical aesthetics under the power of deep learning, then turning back the clock 20 years we would witness the wisdom of cold weapon era. This paper extensively reviews 400+ papers of object detection in the light of its technical evolution, spanning over a quarter-century’s time (from the 1990s to 2019). A number of topics have been covered in this paper, including the milestone detectors in history, detection datasets, metrics, fundamental building blocks of the detection system, speed up techniques, and the recent state of the art detection methods. This paper also reviews some important detection applications, such as pedestrian detection, face detection, text detection, etc, and makes an in-deep analysis of their challenges as well as technical improvements in recent years. | **摘要**  目标检测作为计算机视觉中最基本和最具挑战性的问题之一，近年来受到了极大的关注。它在过去二十年的发展可以被视为计算机视觉历史的缩影。如果我们将今天的目标检测视为深度学习的力量下的技术美学，那么将时钟倒退20年我们将见证冷武器时代的智慧。本文广泛回顾了400多篇关于目标检测的论文，结合其技术发展，跨越了四分之一世纪的时间（从20世纪90年代到2019年）。本文涵盖了许多主题，包括历史里程碑检测器，检测数据集，度量，检测系统的基本构建模块，加速技术以及最新的检测方法。本文还回顾了一些重要的检测应用，如行人检测，人脸检测，文本检测等，并对近年来的挑战和技术改进进行了深入的分析。 |
| 1. **Introduction**   Object detection is an important computer vision task that deals with detecting instances of visual objects of a certain class (such as humans, animals, or cars) in digital images. The objective of object detection is to develop computational models and techniques that provide one of the most basic pieces of information needed by computer vision applications: What objects are where?  As one of the fundamental problems of computer vision, object detection forms the basis of many other computer vision tasks, such as instance segmentation [1–4], image captioning [5–7], object tracking [8], etc. From the application point of view, object detection can be grouped into two research topics “general object detection” and “detection applications”, where the former one aims to explore the methods of detecting different types of objects under a unified framework to simulate the human vision and cognition, and the later one refers to the detection under specific application scenarios, such as pedestrian detection, face detection, text detection, etc. In recent years, the rapid development of deep learning techniques [9] has brought new blood into object detection, leading to remarkable breakthroughs and pushing it forward to a research hot-spot with unprecedented attention. Object detection has now been widely used in many real-world applications, such as autonomous driving, robot vision, video surveillance, etc. Fig. 1 shows the growing number of publications that are associated with “object detection” over the past two decades. | 1. **引文**   目标检测是一种重要的计算机视觉任务，其涉及在数字图像中检测特定类（例如人，动物或汽车）的视觉对象的实例。目标检测的目标是开发计算模型和技术，提供计算机视觉应用所需的最基本信息之一：哪些对象在哪里？  作为计算机视觉的基本问题之一，目标检测构成了许多其他计算机视觉任务的基础，如实例分割[1-4]，图像描述生成[5-7]，目标跟踪[8]等。从应用的角度来看，目标检测可以分为两个研究课题“一般对象检测”和“检测应用”，前者旨在探索在统一框架下检测不同类型对象的方法，以模拟人类视觉和认知，后者是指特定应用场景下的检测，如行人检测，人脸检测，文本检测等。近年来，深度学习技术的快速发展[9]为物体检测带来了新的血液，取得了显着的突破，并将其推向了前所未有的关注研究热点。目标检测现已广泛用于许多实际应用中，例如自动驾驶，机器人视觉，视频监视等。图1显示了过去二十年中与“目标检测”相关的越来越多的出版物。 |

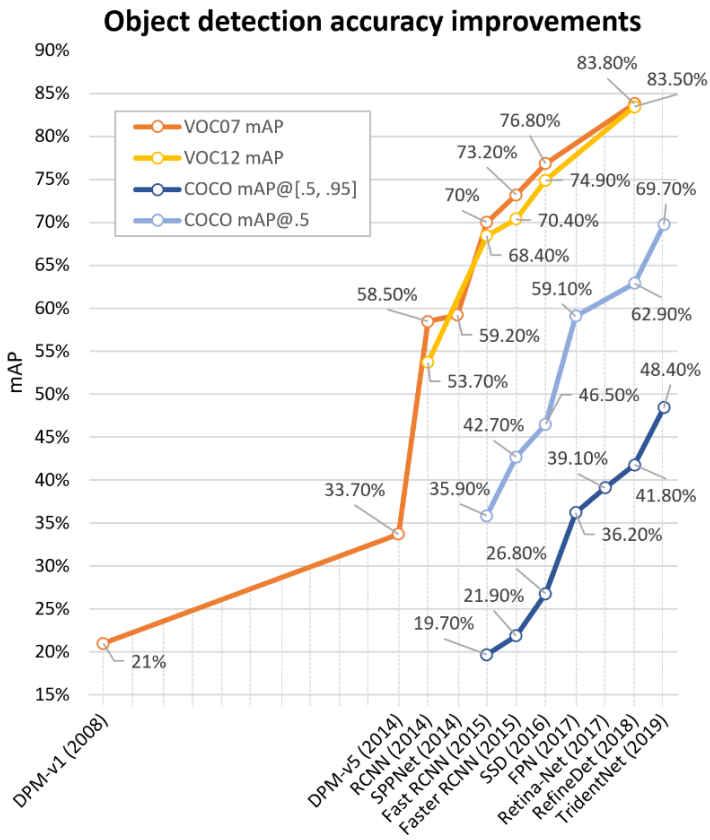


|  |  |
| --- | --- |
| *Figure 1. The increasing number of publications in object detection from 1998 to 2018. (Data from Google scholar advanced search: allintitle: “object detection” AND “detecting objects”.)* | *图1.* *从1998年到2018年，目标检测的出版物数量不断增加。(Google学术高级搜索数据：allintitle：“object detection” 和 “detecting objects”。)* |
| * **Difference from other related reviews**   A number of reviews of general object detection have been published in recent years [24–28]. The main difference between this paper and the above reviews are summarized as follows:   1. **A comprehensive review in the light of technical evolutions**: This paper extensively reviews 400+ papers in the development history of object detection, spanning over a quarter-century’s time (from the 1990s to 2019). Most of the previous reviews merely focus on a short historical period or on some specific detection tasks without considering the technical evolutions over their entire lifetime. Standing on the highway of the history not only helps readers build a complete knowledge hierarchy but also helps to find future directions of this fast developing field. 2. **An in-depth exploration of the key technologies and the recent state of the arts**: After years of development, the state of the art object detection systems have been integrated with a large number of techniques such as “multiscale detection”, “hard negative mining”, “bounding box regression”, etc. However, previous reviews lack fundamental analysis to help readers understand the nature of these sophisticated techniques, e.g., “Where did they come from and how did they evolve?” “What are the pros and cons of each group of methods?” This paper makes an in-depth analysis for readers of the above concerns. 3. **A comprehensive analysis of detection speed up techniques**: The acceleration of object detection has long been a crucial but challenging task. This paper makes an extensive review of the speed up techniques in 20 years of object detection history at multiple levels, including “detection pipeline” (e.g., cascaded detection, feature map shared computation), “detection backbone” (e.g., network compression, lightweight network design), and “numerical computation” (e.g., integral image, vector quantization). This topic is rarely covered by previous reviews.  * **Difficulties and Challenges in Object Detection**   Despite people always asking “what are the difficulties and challenges in object detection?”, actually, this question is not easy to answer and may even be over-generalized. As different detection tasks have totally different objectives and constraints, their difficulties may vary from each other. In addition to some common challenges in other computer vision tasks such as objects under different viewpoints, illuminations, and intraclass variations, the challenges in object detection include but not limited to the following aspects: object rotation and scale changes (e.g., small objects), accurate object localization, dense and occluded object detection, speed up of detection, etc. In Sections 4 and 5, we will give a more detailed analysis of these topics.  The rest of this paper is organized as follows. In Section 2, we review the 20 years’ evolutionary history of object detection. Some speed up techniques in object detection will be introduced in Section 3. Some state of the art detection methods in the recent three years are summarized in Section 4. Some important detection applications will be reviewed in Section 5. In Section 6, we conclude this paper and make an analysis of the further research directions. | * **与其他综述文章的区别**   近年来已发表了许多关于一般物体检测的综述[24-28]。本文与上述综述文章的主要区别概括如下：   1. **基于技术演进的全面回顾**：本文广泛回顾了目标检测发展史上的400多篇论文，跨越了四分之一世纪的时间（从20世纪90年代到2019年）。以前的大多数综述文章仅关注短暂的历史时期或某些特定的检测任务，而不考虑其整个生命周期中的技术演变。站在历史的高速公路上不仅有助于读者建立一个完整的知识层次，而且有助于找到这个快速发展的领域的未来方向。 2. **深入探索关键技术及最新技术状态**：经过多年的发展，最先进的目标检测系统已经与“多尺度检测”，“难例挖掘”，“边界框回归”等大量技术相结合。但是，以前的回顾缺乏基本的分析，以帮助读者理解这些复杂技术的本质，例如， “它们来自哪里以及它们是如何演变的？” “每组方法有哪些优点和缺点？”本文对上述问题向读者进行了深入分析。 3. **对检测提速技术的全面分析**：目标检测的提速一直是一项至关重要但具有挑战性的任务。本文从多个层面对20年来目标检测历史上的提速技术进行了广泛的回顾，包括“检测流水线”（例如，级联检测，特征图共享计算），“检测骨干”（例如，网络压缩，轻量级网络设计）和“数值计算”（例如，积分图像，矢量量化）。以前的综述很少涉及这个主题。  * **目标检测的难点和挑战**   尽管人们总是问“物体检测中的困难和挑战是什么？”，实际上，这个问题并不容易回答，甚至可能过于笼统。由于不同的检测任务具有完全不同的目标和约束，因此它们的困难可能彼此不同。除了在其他计算机视觉任务中的一些常见挑战，例如不同视角下的物体，光照和类内变化，物体检测中的挑战包括但不限于以下方面：物体旋转和尺度变化（例如，小物体），精确的物体定位，密集和遮挡物体检测，检测速度等。在第4节和第5节中，我们将对这些主题进行更详细的分析。  本文的其余部分安排如下。在第2节中，我们回顾了20年来物体检测的进化历史。第3节将介绍目标检测中的一些提速技术。第4节总结最近三年中的一些最先进的检测方法。第5节回顾一些重要的检测应用。在第6节中，我们得出论文的结论并对进一步的研究方向进行分析。 |
| 1. **OBJECT DETECTION IN 20 YEARS**   In this section, we will review the history of object detection in multiple aspects, including milestone detectors, object detection datasets, metrics, and the evolution of key techniques. | 1. **目标检测的20年**   在本节中，我们将从多个方面回顾目标检测的历史，包括里程碑检测器，目标检测数据集，度量和关键技术的演变。 |
| **2.1 A Road Map of Object Detection**  In the past two decades, it is widely accepted that the progress of object detection has generally gone through two historical periods: “traditional object detection period (before 2014)” and “deep learning based detection period (after 2014)”, as shown in Fig. 2. | **2.1 目标检测的路线图**  在过去的二十年中，人们普遍认为目标检测的进展一般经历了两个历史时期：“传统物体检测期（2014年之前）”和“深度学习检测期（2014年之后）”，如图2所示。 |



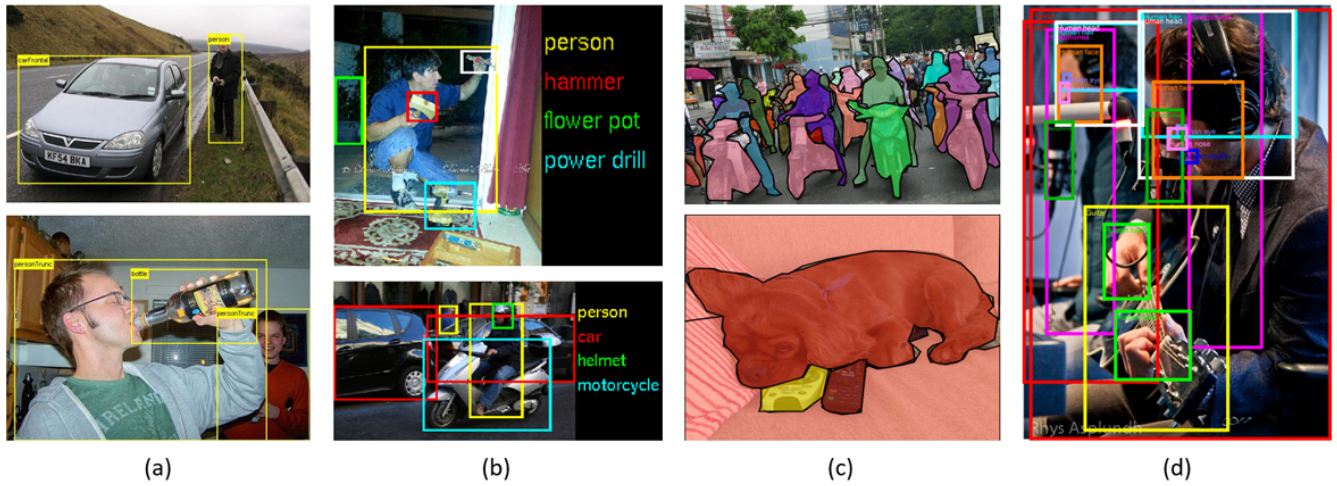
|  |  |
| --- | --- |
| *Figure 2. A road map of object detection. Milestone detectors in this figure: VJ Det. [10, 11], HOG Det. [12], DPM [13–15], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], Pyramid Networks [22], Retina-Net [23].* | *图2.* *目标检测的路线图。图中的里程碑检测器：VJ Det. [10, 11], HOG Det. [12], DPM [13–15], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], Pyramid Networks [22], Retina-Net [23]。* |
| **2.1.1 Milestones: Traditional Detectors**  If we think of today’s object detection as a technical aesthetics under the power of deep learning, then turning back the clock 20 years we would witness “the wisdom of cold weapon era”. Most of the early object detection algorithms were built based on handcrafted features. Due to the lack of effective image representation at that time, people have no choice but to design sophisticated feature representations, and a variety of speed up skills to exhaust the usage of limited computing resources.   * **Viola Jones Detectors**   18 years ago, P. Viola and M. Jones achieved real-time detection of human faces for the first time without any constraints (e.g., skin color segmentation) [10, 11]. Running on a 700MHz Pentium III CPU, the detector was tens or even hundreds of times faster than any other algorithms in its time under comparable detection accuracy. The detection algorithm, which was later referred to the “Viola-Jones (VJ) detector”, was herein given by the authors’ names in memory of their significant contributions.  The VJ detector follows a most straight forward way of detection, i.e., sliding windows: to go through all possible locations and scales in an image to see if any window contains a human face. Although it seems to be a very simple process, the calculation behind it was far beyond the computer’s power of its time. The VJ detector has dramatically improved its detection speed by incorporating three important techniques: “integral image”, “feature selection”, and “detection cascades”.   1. **Integral image**: The integral image is a computational method to speed up box filtering or convolution process. Like other object detection algorithms in its time [29–31], the Haar wavelet is used in VJ detector as the feature representation of an image. The integral image makes the computational complexity of each window in VJ detector independent of its window size. 2. **Feature selection**: Instead of using a set of manually selected Haar basis filters, the authors used Adaboost algorithm [32] to select a small set of features that are mostly helpful for face detection from a huge set of random features pools (about 180k-dimensional). 3. **Detection cascades**: A multi-stage detection paradigm (a.k.a. the “detection cascades”) was introduced in VJ detector to reduce its computational overhead by spending less computations on background windows but more on face targets.  * **HOG Detector**   Histogram of Oriented Gradients (HOG) feature descriptor was originally proposed in 2005 by N.Dalal and B.Triggs [12]. HOG can be considered as an important improvement of the scale-invariant feature transform [33, 34] and shape contexts [35] of its time. To balance the feature invariance (including translation, scale, illumination, etc) and the nonlinearity (on discriminating different objects categories), the HOG descriptor is designed to be computed on a dense grid of uniformly spaced cells and use overlapping local contrast normalization (on “blocks”) for improving accuracy. Although HOG can be used to detect a variety of object classes, it was motivated primarily by the problem of pedestrian detection. To detect objects of different sizes, the HOG detector rescales the input image for multiple times while keeping the size of a detection window unchanged. The HOG detector has long been an important foundation of many object detectors [13, 14, 36] and a large variety of computer vision applications for many years.   * **Deformable Part-based Model (DPM)**   DPM, as the winners of VOC-07, -08, and -09 detection challenges, was the peak of the traditional object detection methods. DPM was originally proposed by P. Felzenszwalb [13] in 2008 as an extension of the HOG detector, and then a variety of improvements have been made by R. Girshick [14, 15, 37, 38].  The DPM follows the detection philosophy of “divide and conquer”, where the training can be simply considered as the learning of a proper way of decomposing an object, and the inference can be considered as an ensemble of detections on different object parts. For example, the problem of detecting a “car” can be considered as the detection of its window, body, and wheels. This part of the work, a.k.a. “star-model”, was completed by P. Felzenszwalb et al. [13]. Later on, R. Girshick has further extended the star-model to the “mixture models” [14, 15, 37, 38] to deal with the objects in the real world under more significant variations.  A typical DPM detector consists of a root-filter and a number of part-filters. Instead of manually specifying the configurations of the part filters (e.g., size and location), a weakly supervised learning method is developed in DPM where all configurations of part filters can be learned automatically as latent variables. R. Girshick has further formulated this process as a special case of Multi-Instance learning [39], and some other important techniques such as “hard negative mining”, “bounding box regression”, and “context priming” are also applied for improving detection accuracy (to be introduced in Section 2.3). To speed up the detection, Girshick developed a technique for “compiling” detection models into a much faster one that implements a cascade architecture, which has achieved over 10 times acceleration without sacrificing any accuracy [14, 38].  Although today’s object detectors have far surpassed DPM in terms of the detection accuracy, many of them are still deeply influenced by its valuable insights, e.g., mixture models, hard negative mining, bounding box regression, etc. In 2010, P. Felzenszwalb and R. Girshick were awarded the “lifetime achievement” by PASCAL VOC.  **2.1.2 Milestones: CNN based Two-stage Detectors**  As the performance of hand-crafted features became saturated, object detection has reached a plateau after 2010. R.Girshick says: “... progress has been slow during 2010-2012, with small gains obtained by building ensemble systems and employing minor variants of successful methods”[38]. In 2012, the world saw the rebirth of convolutional neural networks [40]. As a deep convolutional network is able to learn robust and high-level feature representations of an image, a natural question is whether we can bring it to object detection? R. Girshick et al. took the lead to break the deadlocks in 2014 by proposing the Regions with CNN features (RCNN) for object detection [16, 41]. Since then, object detection started to evolve at an unprecedented speed.  In deep learning era, object detection can be grouped into two genres: “two-stage detection” and “one-stage detection”, where the former frames the detection as a “coarse-to-fine” process while the later frames it as to “complete in one step”.   * **RCNN**   The idea behind RCNN is simple: It starts with the extraction of a set of object proposals (object candidate boxes) by selective search [42]. Then each proposal is rescaled to a fixed size image and fed into a CNN model trained on ImageNet (say, AlexNet [40]) to extract features. Finally, linear SVM classifiers are used to predict the presence of an object within each region and to recognize object categories. RCNN yields a signicant performance boost on VOC07, with a large improvement of mean Average Precision (mAP) from 33.7% (DPM-v5 [43]) to 58.5%.  Although RCNN has made great progress, its drawbacks are obvious: the redundant feature computations on a large number of overlapped proposals (over 2000 boxes from one image) leads to an extremely slow detection speed (14s per image with GPU). Later in the same year, SPPNet [17] was proposed and has overcome this problem.   * **SPPNet**   In 2014, K. He et al. proposed Spatial Pyramid Pooling Networks (SPPNet) [17]. Previous CNN models require a fixed-size input, e.g., a 224x224 image for AlexNet [40]. The main contribution of SPPNet is the introduction of a Spatial Pyramid Pooling (SPP) layer, which enables a CNN to generate a fixed-length representation regardless of the size of image/region of interest without rescaling it. When using SPPNet for object detection, the feature maps can be computed from the entire image only once, and then fixed length representations of arbitrary regions can be generated for training the detectors, which avoids repeatedly computing the convolutional features. SPPNet is more than 20 times faster than R-CNN without sacrificing any detection accuracy (VOC07 mAP=59.2%).  Although SPPNet has effectively improved the detection speed, there are still some drawbacks: first, the training is still multi-stage, second, SPPNet only fine-tunes its fully connected layers while simply ignores all previous layers. Later in the next year, Fast RCNN [18] was proposed and solved these problems.   * **Fast RCNN**   In 2015, R. Girshick proposed Fast RCNN detector [18], which is a further improvement of R-CNN and SPPNet [16, 17]. Fast RCNN enables us to simultaneously train a detector and a bounding box regressor under the same network configurations. On VOC07 dataset, Fast RCNN increased the mAP from 58.5% (RCNN) to 70.0% while with a detection speed over 200 times faster than R-CNN.  Although Fast RCNN successfully integrates the advantages of RCNN and SPPNet, its detection speed is still limited by the proposal detection (see Section 2.3.2 for more details). Then, a question naturally arises: “can we generate object proposals with a CNN model?” Later, Faster R-CNN [19] has answered this question.   * **Faster RCNN**   In 2015, S. Ren et al. proposed Faster RCNN detector [19, 44] shortly after the Fast RCNN. Faster RCNN is the first end-to-end, and the first near-realtime deep learning detector (COCO mAP@.5=42.7%, COCO mAP@[.5,.95]=21.9%, VOC07 mAP=73.2%, VOC12 mAP=70.4%, 17fps with ZFNet [45]). The main contribution of Faster-RCNN is the introduction of Region Proposal Network (RPN) that enables nearly cost-free region proposals. From R-CNN to Faster RCNN, most individual blocks of an object detection system, e.g., proposal detection, feature extraction, bounding box regression, etc, have been gradually integrated into a unified, end-to-end learning framework.  Although Faster RCNN breaks through the speed bottleneck of Fast RCNN, there is still computation redundancy at subsequent detection stage. Later, a variety of improvements have been proposed, including RFCN [46] and Light head RCNN [47]. (See more details in Section 3.)   * **Feature Pyramid Networks**   In 2017, T.-Y.Lin et al. proposed Feature Pyramid Networks (FPN) [22] on basis of Faster RCNN. Before FPN, most of the deep learning based detectors run detection only on a network’s top layer. Although the features in deeper layers of a CNN are beneficial for category recognition, it is not conducive to localizing objects. To this end, a top-down architecture with lateral connections is developed in FPN for building high-level semantics at all scales. Since a CNN naturally forms a feature pyramid through its forward propagation, the FPN shows great advances for detecting objects with a wide variety of scales. Using FPN in a basic Faster R-CNN system, it achieves state-of-the-art single model detection results on the MSCOCO dataset without bells and whistles (COCO mAP@.5=59.1%, COCO mAP@[.5, .95]=36.2%). FPN has now become a basic building block of many latest detectors.  **2.1.3 Milestones: CNN based One-stage Detectors**   * **You Only Look Once (YOLO)**   YOLO was proposed by R. Joseph et al. in 2015. It was the first one-stage detector in deep learning era [20]. YOLO is extremely fast: a fast version of YOLO runs at 155fps with VOC07 mAP=52.7%, while its enhanced version runs at 45fps with VOC07 mAP=63.4% and VOC12 mAP=57.9%. YOLO is the abbreviation of “You Only Look Once”. It can be seen from its name that the authors have completely abandoned the previous detection paradigm of “proposal detection + verification”. Instead, it follows a totally different philosophy: to apply a single neural network to the full image. This network divides the image into regions and predicts bounding boxes and probabilities for each region simultaneously. Later, R. Joseph has made a series of improvements on basis of YOLO and has proposed its v2 and v3 editions [48, 49], which further improve the detection accuracy while keeps a very high detection speed.  In spite of its great improvement of detection speed, YOLO suffers from a drop of the localization accuracy compared with two-stage detectors, especially for some small objects. YOLO’s subsequent versions [48, 49] and the latter proposed SSD [21] has paid more attention to this problem.   * **Single Shot MultiBox Detector (SSD)**   SSD [21] was proposed by W. Liu et al. in 2015. It was the second one-stage detector in deep learning era. The main contribution of SSD is the introduction of the multi-reference and multi-resolution detection techniques (to be introduce in Section 2.3.2), which significantly improves the detection accuracy of a one-stage detector, especially for some small objects. SSD has advantages in terms of both detection speed and accuracy (VOC07 mAP=76.8%, VOC12 mAP=74.9%, COCO mAP@.5=46.5%, mAP@[.5,.95]=26.8%, a fast version runs at 59fps). The main difference between SSD and any previous detectors is that the former one detects objects of different scales on different layers of the network, while the latter ones only run detection on their top layers.   * **RetinaNet**   In despite of its high speed and simplicity, the one-stage detectors have trailed the accuracy of two-stage detectors for years. T.-Y. Lin et al. have discovered the reasons behind and proposed RetinaNet in 2017 [23]. They claimed that the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause. To this end, a new loss function named “focal loss” has been introduced in RetinaNet by reshaping the standard cross entropy loss so that detector will put more focus on hard, misclassified examples during training. Focal Loss enables the one-stage detectors to achieve comparable accuracy of two-stage detectors while maintaining very high detection speed. (COCO mAP@.5=59.1%, mAP@[.5, .95] =39.1%).  **2.2 Object Detection Datasets and Metrics**  Building larger datasets with less bias is critical for developing advanced computer vision algorithms. In object detection, a number of well-known datasets and benchmarks have been released in the past 10 years, including the datasets of PASCAL VOC Challenges [50, 51] (e.g., VOC2007, VOC2012), ImageNet Large Scale Visual Recognition Challenge (e.g., ILSVRC2014) [52], MS-COCO Detection Challenge [53], etc. The statistics of these datasets are given in Table 1. Fig. 4 shows some image examples of these datasets. Fig. 3 shows the improvements of detection accuracy on VOC07, VOC12 and MS-COCO datasets from 2008 to 2018. | **2.1.1里程碑：传统检测器**  如果我们将今天的目标检测视为深度学习的力量下的技术美学，那么将时钟倒退20年我们将见证“冷武器时代的智慧”。大多数早期目标检测算法都是基于手工制作的特征构建的。由于当时缺乏有效的图像表示，人们别无选择，只能设计复杂的特征表示，以及各种提速技能来耗尽有限的计算资源。   * **Viola Jones 检测器**   18年前，P. Viola和M. Jones首次实现了人脸的实时检测，没有任何限制（例如，肤色分割）[10,11]。在700MHz奔腾III CPU上运行，在同等检测精度下，检测器的速度比其他任何算法快几十甚至几百倍。后来被称为“Viola-Jones（VJ）检测器”的检测算法在此由作者的名字给出以记忆它们的重要贡献。  VJ检测器遵循最直接的检测方式，即滑动窗口：遍历所有可能的位置并在图像中缩放以查看是否有任何窗口包含人脸。虽然这似乎是一个非常简单的过程，但它背后的计算远远超出了计算机的时代力量。VJ检测器通过结合三种重要技术（“积分图像”，“特征选择”和“检测级联”）显着提高了检测速度。   1. **积分图像**：积分图像是加速盒滤波或卷积过程的计算方法。与其他同时期的物体检测算法[29-31]一样，Haar小波在VJ检测器中用作图像的特征表示。积分图像使VJ检测器中每个窗口的计算复杂度与其窗口大小无关。 2. **特征选择**：作者不是使用一组手动选择的Haar基础滤波器，而是使用Adaboost算法[32]来选择一小组特征，这些特征最有助于从大量随机特征池（约180k维）中进行人脸检测。 3. **检测级联**：在VJ检测器中引入了多阶段检测范例（也就是“检测级联”），通过在背景窗口上花费更少的计算但在面部目标上花费更多来减少其计算开销。  * **HOG检测器**   定向梯度直方图（HOG）特征描述符最初由N.Dalal和B.Triggs在2005年提出[12]。HOG可以被认为是当时尺度不变特征变换[33,34]和形状上下文[35]的重要改进。为了平衡特征不变性（包括平移，缩放，照明等）和非线性（在区分不同对象类别上），HOG描述符被设计为在均匀间隔的单元的密集网格上计算并使用重叠的局部对比度归一化（在“块”）用于提高准确性。尽管HOG可用于检测各种对象类别，但主要是用于行人检测问题。为了检测不同尺寸的物体，HOG检测器多次重新调整输入图像，同时保持检测窗口的大小不变。长期以来，HOG检测器一直是许多目标检测器[13,14,36]和多种计算机视觉应用的重要基础。   * **可变形部件模型 (DPM)**   DPM作为VOC-07，-08和-09检测挑战的赢家，是传统物体检测方法的巅峰之作。DPM最初是由P. Felzenszwalb [13]于2008年提出的，作为HOG检测器的扩展，然后R. Girshick [14,15,37,38]进行了各种改进。  DPM遵循“分而治之”的检测理念，其中训练可以简单地被认为是对分解对象的适当方式的学习，并且推理可以被认为是对不同对象部分的检测的集合。例如，检测“汽车”的问题可以被认为是其窗户，车身和车轮的检测。P. Felzenszwalb等人完成了这项工作，即a.k.a.“star-model”[13]。后来，R. Girshick进一步将star-model扩展到“mixture models”[14,15,37,38]，以便在更显着的变化下处理现实世界中的物体。  典型的DPM检测器由根滤波器和多个部分滤波器组成。不是手动指定部分滤波器的配置（例如，大小和位置），而是在DPM中开发弱监督学习方法，其中部分滤波器的所有配置可以作为潜在变量自动学习。R. Girshick进一步将此过程作为多实例学习的一个特例进行了阐述[39]，其他一些重要的技术，如“难例挖掘”，“边界框回归”和“背景启动”也被应用于改进检测精度（将在2.3节中介绍）。为了加快检测速度，Girshick开发了一种技术，用于将检测模型“编译”成更快的检测模型，实现级联结构，加速度超过10倍而不牺牲任何精度[14,38]。  尽管今天的目标检测器在检测精度方面远远超过DPM，但其中许多仍深受其宝贵见解的影响，例如混合模型，难例挖掘，边界框回归等。2010年，P. Felzenszwalb和R.Girshick被PASCAL VOC授予“终身成就”。  **2.1.2里程碑：基于CNN的两阶段检测器**  随着手工制作特征的表现逐渐饱和，目标检测在2010年后达到了一个平台。R.Girshick说：“...... 2010 - 2012年进展缓慢，通过建立整体系统和采用微小变体获得了小幅增长成功的方法”[38]。2012年，世界看到了卷积神经网络的重生[40]。由于深度卷积网络能够学习图像的强大和高级特征表示，一个自然的问题是我们是否可以将其用于对象检测？R. Girshick等于2014年率先通过提取具有CNN特征的区域（RCNN）进行物体检测来打破僵局[16,41]。从那时起，物体检测开始以前所未有的速度发展。  在深度学习时代，目标检测可以分为两种类型：“两阶段检测”和“一阶段检测”，前者将检测框架化为“粗略转精细”过程，而后者将其框架化为“一步完成”。   * **RCNN**   RCNN背后的想法很简单：它首先通过选择性搜索[42]提取一组对象提议（对象候选框）。然后将每个提案重新调整为固定大小的图像，并将其输入到在ImageNet上训练的CNN模型（例如，AlexNet [40]）以提取特征。最后，线性SVM分类器用于预测每个区域内对象的存在并识别对象类别。RCNN对VOC07有显着的性能提升，平均精度（mAP）从33.7％（DPM-v5 [43]）大幅提升至58.5％。  尽管RCNN取得了很大的进步，但它的缺点是显而易见的：大量重叠提议（一个图像超过2000个盒子）的冗余特征计算导致检测速度极慢（使用GPU每个图像14s）。同年晚些时候，提出了SPPNet [17]并克服了这个问题。   * **SPPNet**   2014年，K. He等人提出的空间金字塔池化网络（SPPNet）[17]。以前的CNN模型需要固定大小的输入，例如AlexNet [44]的224x224图像。SPPNet的主要贡献是引入了空间金字塔池化（SPP）层，该层使得CNN能够生成固定长度的表示，而不管感兴趣的图像/区域的大小从而不用重新缩放它。当使用SPPNet进行物体检测时，可以仅从整个图像计算一次特征图，然后可以生成任意区域的固定长度表示以训练检测器，这避免了重复计算卷积特征。在不牺牲任何检测精度的情况下，SPPNet比R-CNN快20倍（VOC07 ​​mAP = 59.2％）。  尽管SPPNet有效地提高了检测速度，但仍然存在一些缺点：首先，训练仍然是多阶段的，其次，SPPNet仅对其完全连接的层进行微调，而忽略了所有先前的层。次年晚些时候，Fast RCNN [18]被提出并解决了这些问题。   * **Fast RCNN**   2015年，R. Girshick提出了快速RCNN检测器[18]，这是对R-CNN和SPPNet的进一步改进[16,17]。Fast RCNN使我们能够在相同的网络配置下同时训练检测器和边界框回归器。在VOC07数据集上，Fast RCNN将mAP从58.5％（RCNN）增加到70.0％，而检测速度比R-CNN快200倍。  尽管Fast RCNN成功地集成了RCNN和SPPNet的优势，但其检测速度仍然受到提议检测的限制（更多详细信息，请参见第2.3.2节）。然后，一个问题自然而然地出现了：“我们能用CNN模型生成对象建议吗？”后来，Faster R-CNN [19]回答了这个问题。   * **Faster RCNN**   2015年，S．Ren等人在Fast RCNN之后不久，提出了Faster RCNN检测器[19,44]。Faster RCNN是第一个端到端，近实时深度学习检测器(COCO mAP@.5=42.7%, COCO mAP@[.5,.95]=21.9%, VOC07 mAP=73.2%, VOC12 mAP=70.4%, 17fps with ZFNet [45])。Faster-RCNN的主要贡献是引入了区域提案网络（RPN），该网络提供了几乎无成本的区域提案。从R-CNN到Faster RCNN，目标检测系统的大多数单独的块（例如，提议检测，特征提取，边界框回归等）已经逐渐集成到统一的端到端学习框架中。  虽然Faster RCNN突破了Fast RCNN的速度瓶颈，但在后续检测阶段仍然存在计算冗余。后来，提出了各种改进，包括RFCN [46]和Light head RCNN [47]。（详见第3节）   * **特征金字塔网络**   2017年，T.-Y.Lin等人提出基于Faster RCNN的特征金字塔网络（FPN）[22]。在FPN之前，大多数基于深度学习的检测器仅在网络的顶层运行检测。虽然CNN的更深层中的特征有益于类别识别，但是它不利于对象的定位。为此，在FPN中开发了具有横向连接的自顶向下架构，用于在所有尺度上构建高级语义。由于CNN通过其向前传播自然地形成特征金字塔，因此FPN显示出用于检测具有各种尺度的物体的巨大进步。在基本的Faster R-CNN系统中使用FPN，它在MSCOCO数据集上实现了最先进的单一模型检测结果(COCO mAP@.5=59.1%, COCO mAP@[.5, .95]=36.2%)。FPN现已成为许多最新检测器的基本构建模块。  **2.1.3里程碑：基于CNN的单阶段检测器**   * **You Only Look Once (YOLO)**   YOLO由R. Joseph等人提出。在2015年，它是深度学习时代的第一个单阶段检测器[20]。YOLO非常快：YOLO的快速版本以155fps运行，VOC07 ​​mAP = 52.7％，而其增强版本运行速度为45fps，VOC07 ​​mAP = 63.4％，VOC12 mAP = 57.9％。YOLO是“You Only Look Once”的缩写。从其名称可以看出，作者完全放弃了先前的“提案检测+验证”检测范例。相反，它遵循完全不同的理念：将单个神经网络应用于完整图像。该网络将图像划分为区域，并同时预测每个区域的边界框和概率。后来，R. Joseph在YOLO的基础上进行了一系列的改进，并提出了v2和v3版本[48,49]，这进一步提高了检测精度，同时保持了非常高的检测速度。  尽管检测速度有了很大提高，但与两阶段检测器相比，YOLO的定位精度有所下降，特别是对于某些小物体。YOLO的后续版本[48,49]和后面提出的SSD [21]更加关注这个问题。   * **Single Shot MultiBox Detector (SSD)**   SSD [21]由W. Liu等人提出。在2015年，它是深度学习时代的第二个单阶段检测器。SSD的主要贡献在于引入了多参考和多分辨率检测技术（将在2.3.2节中介绍），这显着提高了单阶段检测器的检测精度，特别是对于某些小物体。SSD在检测速度和准确度方面具有优势(VOC07 mAP=76.8%, VOC12 mAP=74.9%, COCO mAP@.5=46.5%, mAP@[.5,.95]=26.8%, a fast version runs at 59fps)。SSD与任何先前的检测器之间的主要区别在于，前者在网络的不同层上检测到不同比例的物体，而后者仅在其顶层上进行检测。   * **RetinaNet**   尽管其快速和简单，但单阶段检测器已经落后于两阶段检测器的精确度多年。2017年T.-Y. Lin等人的RetinaNet已经发现了背后的原因[23]。他们声称在密集检测器训练过程中遇到的极端前景-背景类不平衡是其主要原因。为此，通过重塑标准交叉熵损失，RetinaNet引入了一种名为“focal loss”的新损失函数，以便检测器在训练期间更加关注难的，错误分类的例子。Focal Loss使单阶段检测器能够实现两级检测器的精确度，同时保持极高的检测速度(COCO mAP @.5=59.1%, mAP@ [.5, .95] = 39.1%)。  **2.2 目标检测数据集和度量标准**  以较少的偏差构建较大的数据集对于开发高级计算机视觉算法至关重要。在物体检测中，过去10年发布了许多众所周知的数据集和基准，包括PASCAL VOC Challenges [50,51]的数据集（例如，VOC2007，VOC2012），ImageNet大规模视觉识别挑战（例如，ILSVRC2014）[52]，MS-COCO检测挑战[53]等。这些数据集的统计数据在表1中给出。图4显示了这些数据集的一些图像示例。图3显示了2008年至2018年VOC07，VOC12和MS-COCO数据集的检测精度的提高。 |



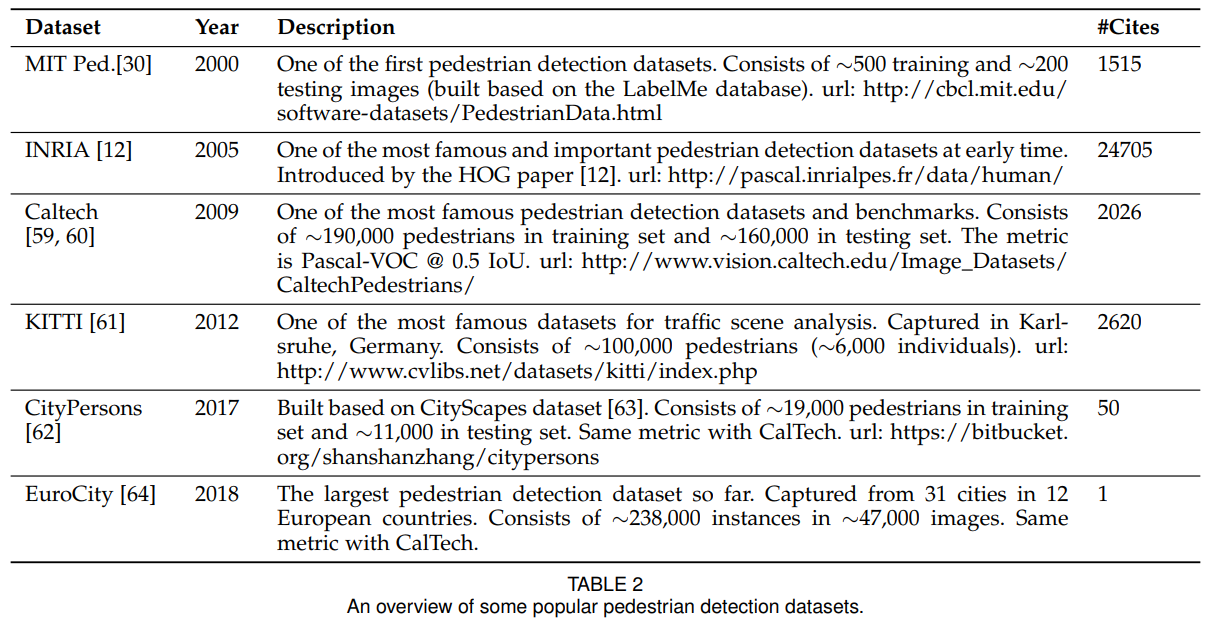


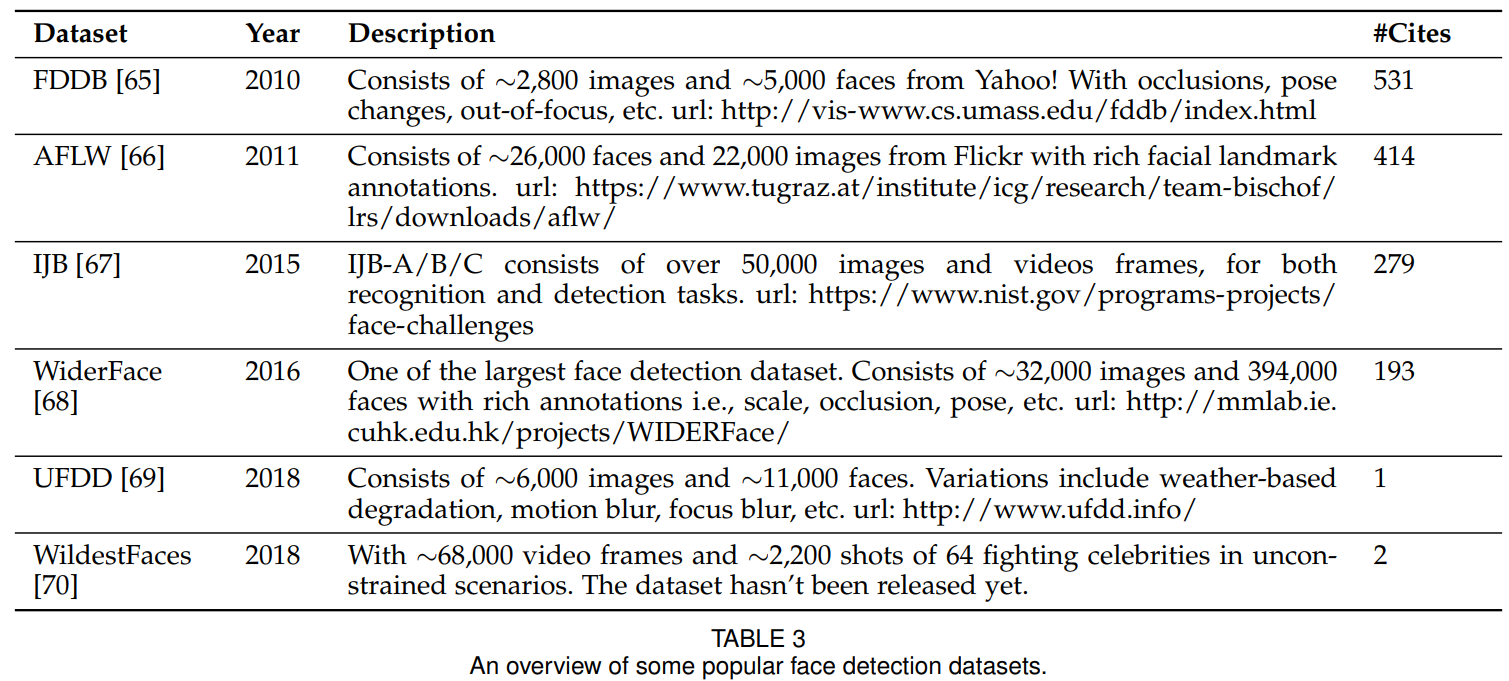
|  |  |
| --- | --- |
| *Figure 3. The accuracy improvements of object detection on VOC07, VOC12 and MS-COCO datasets. Detectors in this figure: DPM-v1 [13], DPM-v5 [54], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], SSD [21], FPN [22], Retina-Net [23], RefineDet [55], TridentNet[56].* | *图3.* *VOC07，VOC12和MS-COCO数据集上物体检测的准确性改进。此图中的探测器：DPM-v1 [13]，DPM-v5 [54]，RCNN [16]，SPPNet [17]，Fast RCNN [18]，Faster RCNN [19]，SSD [21]，FPN [22]，Retina-Net [23]，RefineDet [55]，TridentNet [56]。* |

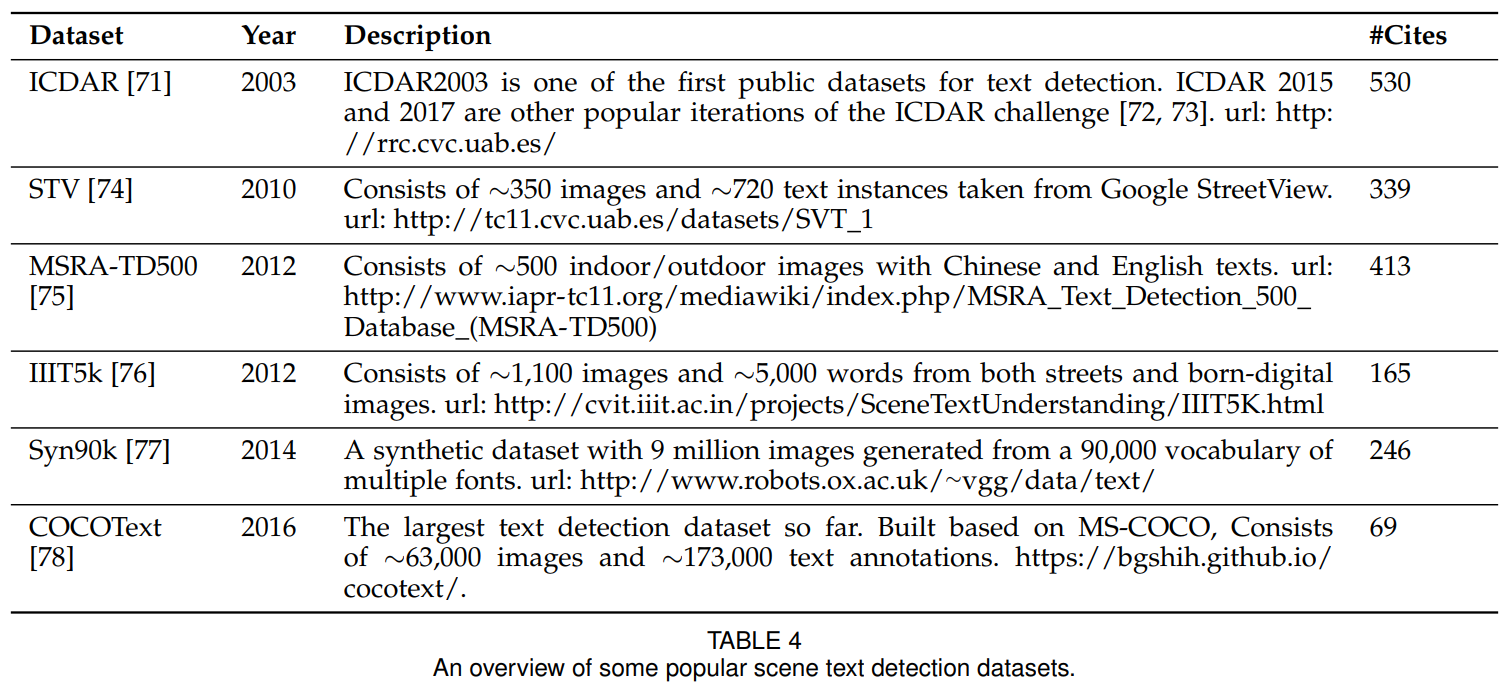
|  |  |
| --- | --- |
| * **Pascal VOC**   [The PASCAL Visual Object Classes (VOC) Challenges](http://host.robots.ox.ac.uk/pascal/VOC/) (from 2005 to 2012) [50, 51] was one of the most important competition in early computer vision community. There are multiple tasks in PASCAL VOC, including image classification, object detection, semantic segmentation and action detection. Two versions of Pascal-VOC are mostly used in object detection: VOC07 and VOC12, where the former consists of 5k tr. images + 12k annotated objects, and the latter consists of 11k tr. images + 27k annotated objects. 20 classes of objects that are common in life are annotated in these two datasets (Person: person; Animal: bird, cat, cow, dog, horse, sheep; Vehicle: aeroplane, bicycle, boat, bus, car, motor-bike, train; Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor). In recent years, as some larger datasets like ILSVRC and MS-COCO (to be introduced) has been released, the VOC has gradually fallen out of fashion and has now become a test-bed for most new detectors. | * **Pascal VOC**   PASCAL视觉对象类（VOC）挑战（2005年至2012年）[50,51]是早期计算机视觉社区中最重要的竞赛之一。PASCAL VOC有多项任务，包括图像分类，目标检测，语义分割和动作检测。两种版本的Pascal-VOC主要用于目标检测：VOC07和VOC12，前者由5k图像+ 12k标注对象，后者由11k图像+ 27k标注对象组成。生活中常见的20类对象在这两个数据集中标注（人：人; 动物：鸟，猫，牛，狗，马，羊; 车辆：飞机，自行车，船，公共汽车，汽车，摩托车，火车; 室内：瓶子，椅子，餐桌，盆栽，沙发，电视/显示器）。近年来，随着一些较大的数据集如ILSVRC和MS-COCO（即将推出）已经发布，VOC已逐渐脱离时尚，现在已成为大多数新检测器的试验台。 |

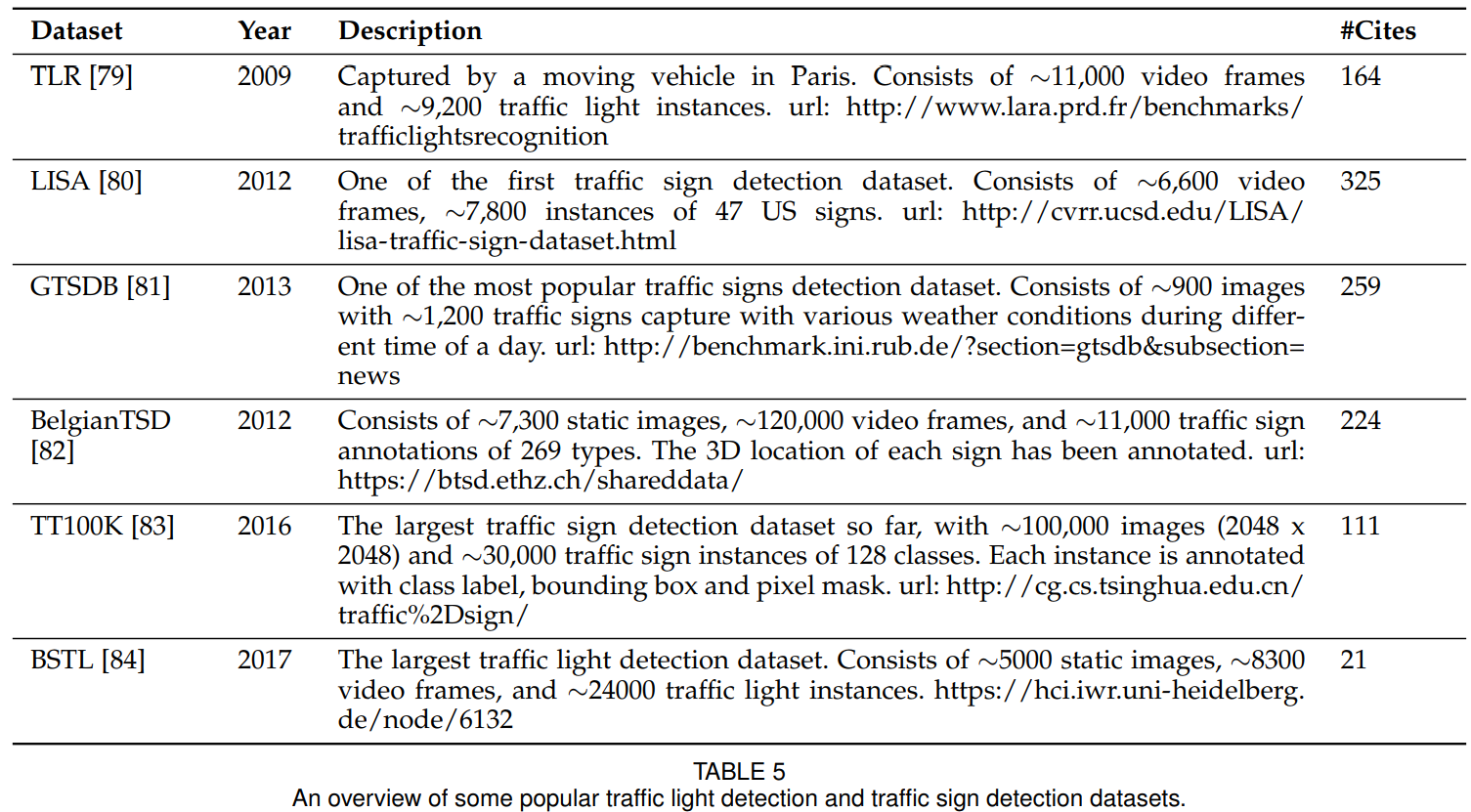


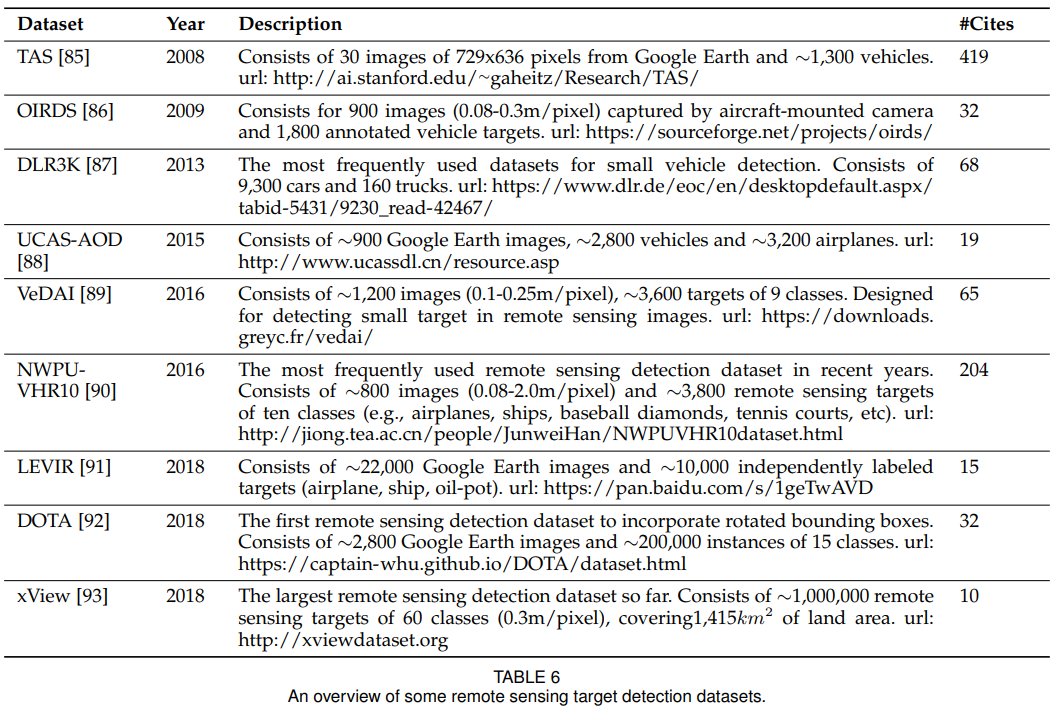
|  |  |
| --- | --- |
| *Figure 4. Some example images and annotations in (a) PASCAL-VOC07, (b) ILSVRC, (c) MS-COCO, and (d) Open Images.* | *图4.* *一些示例图像和标签（a）PASCAL-VOC07，（b）ILSVRC，（c）MS-COCO和（d）Open Images。* |
| * **ILSVRC**   [The ImageNet Large Scale Visual Recognition Challenge (ILSVRC)](http://image-net.org/challenges/LSVRC/) [52] has pushed forward the state of the art in generic object detection. ILSVRC is organized each year from 2010 to 2017. It contains a detection challenge using ImageNet images [57]. The ILSVRC detection dataset contains 200 classes of visual objects. The number of its images/object instances is two orders of magnitude larger than VOC. For example, ILSVRC-14 contains 517k images and 534k annotated objects.   * **MS-COCO**   [MS-COCO](http://cocodataset.org/) [53] is the most challenging object detection dataset available today. The annual competition based on MS-COCO dataset has been held since 2015. It has less number of object categories than ILSVRC, but more object instances. For example, MS-COCO-17 contains 164k images and 897k annotated objects from 80 categories. Compared with VOC and ILSVRC, the biggest progress of MS-COCO is that apart from the bounding box annotations, each object is further labeled using per-instance segmentation to aid in precise localization. In addition, MS-COCO contains more small objects (whose area is smaller than 1% of the image) and more densely located objects than VOC and ILSVRC. All these features make the objects distribution in MSCOCO closer to those of the real world. Just like ImageNet in its time, MS-COCO has become the de facto standard for the object detection community.   * **Open Images**   The year of 2018 sees the introduction of the [Open Images Detection (OID) challenge](https://storage.googleapis.com/openimages/web/index.html) [58], following MS-COCO but at an unprecedented scale. There are two tasks in Open Images: 1) the standard object detection, and 2) the visual relationship detection which detects paired objects in particular relations. For the object detection task, the dataset consists of 1,910k images with 15,440k annotated bounding boxes on 600 object categories.   * **Datasets of Other Detection Tasks**   In addition to general object detection, the past 20 years also witness the prosperity of detection applications in specific areas, such as pedestrian detection, face detection, text detection, traffic sign/light detection, and remote sensing target detection. Tables 2-6 list some of the popular datasets of these detection tasks (The #Cites shows statistics as of Feb. 2019.). A detailed introduction of the detection methods of these tasks can be found in Section 5. | * **ILSVRC**   ImageNet大规模视觉识别挑战赛（ILSVRC）[52]推动了通用目标检测的最新技术水平。ILSVRC每年从2010年到2017年组织。它包含使用ImageNet图像的检测挑战[57]。ILSVRC检测数据集包含200类可视对象。其图像/目标对象实例的数量比VOC大两个数量级。例如，ILSVRC-14包含517k图像和534k标注对象。   * **MS-COCO**   MS-COCO [53]是目前最具挑战性的目标检测数据集。  自2015年以来，基于MS-COCO数据集的年度竞赛已经举办。它的对象类别数量少于ILSVRC，但更多的目标实例。例如，MS-COCO-17包含来自80个类别的164k图像和897k个标注对象。与VOC和ILSVRC相比，MS-COCO的最大进步是除了边界框标注之外，每个对象还使用实例分割进行标记，以帮助精确定位。此外，MS-COCO包含更多的小物体（其面积小于图像的1％）和比VOC和ILSVRC更密集的物体。所有这些功能使MSCOCO中的对象分布更接近现实世界。就像当时的ImageNet一样，MS-COCO已成为目标检测领域的事实标准。   * **Open Images**   2018年开始引入开放式图像检测（OID）挑战 [58]，遵循MS-COCO，但规模空前。在打开图像中有两个任务：1）标准对象检测，以及2）在特定关系中检测成对对象的视觉关系检测。对于对象检测任务，数据集由1,910k图像组成，在600个对象类别上具有15,440k个带标注的边界框。   * **其他检测任务的数据集**   除了一般目标检测之外，过去20年还见证了特定领域检测应用的繁荣，例如行人检测，人脸检测，文本检测，交通标志/光检测和遥感目标检测。表2-6列出了这些检测任务的一些流行数据集(#Cites显示截止2019年2月的引用数目)。有关这些任务的检测方法的详细介绍，请参见第5节。 |











|  |  |
| --- | --- |
| **2.2.1 Metrics**  How can we evaluate the effectiveness of an object detector? This question may even have different answers at different time.  In the early time’s detection community, there is no widely accepted evaluation criteria on detection performance. For example, in the early research of pedestrian detection [12], the “miss rate vs. false positives per-window (FPPW)” was usually used as a metric. However, the perwindow measurement (FPPW) can be flawed and fails to predict full image performance in certain cases [59]. In 2009, the Caltech pedestrian detection benchmark was created [59, 60] and since then, the evaluation metric has changed from per-window (FPPW) to false positives perimage (FPPI).  In recent years, the most frequently used evaluation for object detection is “Average Precision (AP)”, which was originally introduced in VOC2007. AP is defined as the average detection precision under different recalls, and is usually evaluated in a category specific manner. To compare performance over all object categories, the mean AP (mAP) averaged over all object categories is usually used as the final metric of performance. To measure the object localization accuracy, the Intersection over Union (IoU) is used to check whether the IoU between the predicted box and the ground truth box is greater than a predefined threshold, say, 0.5. If yes, the object will be identified as “successfully detected”, otherwise will be identified as “missed”. The 0.5IoU based mAP has then become the de facto metric for object detection problems for years.  After 2014, due to the popularity of MS-COCO datasets, researchers started to pay more attention to the accuracy of the bounding box location. Instead of using a fixed IoU threshold, MS-COCO AP is averaged over multiple IoU thresholds between 0.5 (coarse localization) and 0.95 (perfect localization). This change of the metric has encouraged more accurate object localization and may be of great importance for some real-world applications (e.g., imagine there is a robot arm trying to grasp a spanner).  Recently, there are some further developments of the evaluation in the Open Images dataset, e.g., by considering the group-of boxes and the non-exhaustive image-level category hierarchies. Some researchers also have proposed some alternative metrics, e.g., “localization recall precision” [94]. Despite the recent changes, the VOC/COCO-based mAP is still the most frequently used evaluation metric for object detection. | **2.2.1 度量标准**  我们如何评估目标检测器的有效性？这个问题在不同的时间甚至可能有不同的答案。  在早期的检测社区中，没有广泛接受的检测性能评估标准。例如，在行人检测的早期研究[12]中，“未命中率与每个窗口的误报率（FPPW）”通常用作度量。然而，在某些情况下，逐窗口测量（FPPW）可能存在缺陷并且无法预测完整的图像性能[59]。2009年，创建了加州理工学院行人检测基准[59,60]，从那时起，评估指标从逐窗口（FPPW）变为误报周期（FPPI）。  近年来，最常用的物体检测评估是“平均精度（AP）”，最初是在VOC2007中引入的。AP定义为不同召回下的平均检测精度，通常以特定类别的方式进行评估。为了比较所有对象类别的性能，平均所有对象类别的平均AP（mAP）通常用作性能的最终度量。为了测量物体定位精度，使用交叉联合（IoU）来检查预测框和真值框之间的IoU是否大于预定阈值，例如0.5。如果是，则该对象将被识别为“成功检测到”，否则将被识别为“未命中”。基于0.5IoU的mAP随后成为多年来物体检测问题的事实上的度量。  2014年之后，由于MS-COCO数据集的普及，研究人员开始更加关注边界框位置的准确性。MS-COCO AP不是使用固定的IoU阈值，而是在0.5（粗略定位）和0.95（完美定位）之间的多个IoU阈值上取平均值。度量的这种改变促进了更准确的对象定位，并且对于一些现实世界的应用可能是非常重要的（例如，想象有一个机器人手臂试图抓住扳手）。  最近，在开放图像数据集中存在评估的一些进一步发展，例如，通过考虑分组框和非穷举图像级类别分层结构。一些研究人员还提出了一些替代指标，例如“本地化回忆精度”[94]。尽管最近发生了变化，但基于VOC / COCO的mAP仍然是最常用的物体检测评估指标。 |

1. **References**

[1] Beltrán J, Guindel C, Moreno F M, et al. Birdnet: a 3D object detection framework from LiDAR information[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 3517-3523.