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 年我们将见证冷兵器时代的智慧。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，提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?

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：有哪些对象并且在哪里？

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

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“一般目标检测”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]为目标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。

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.

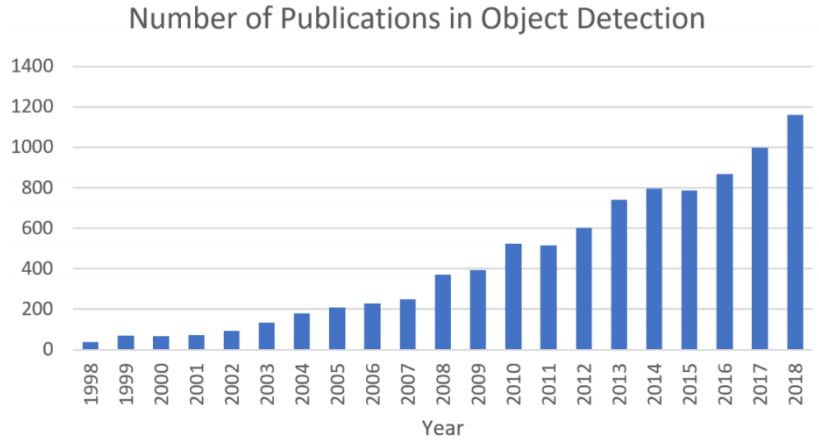


Figure 1. The increasing number of publications in object detection from 1998 to 2018. (Data from Google scholar advanced search: all-in-title: "object detection" AND "detecting objects".)

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：all-in-title："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

- **与其他综述文章的区别**

近年来已发表了许多关于一般目标检测的综述[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。

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.

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。

- **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.

- **目标检测的难点和挑战**

尽管人们总是问“目标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、光照和类别的物体，目标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，密集和遮挡目标检测，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。

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.

本文的其余部分安排如下。在第 2 节中，我们回顾了 20 年来目标检测的进化历史。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。

2. OBJECT DETECTION IN 20 YEARS

2. 目标检测的 20 年

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.

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，度量标准和关键技术的演变。

2.1 A Road Map of Object Detection

2.1 目标检测的路线图

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.

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："传统目标检测期（2014 年之前）"和"深度学习检测期（2014 年之后）"，如图 2 所示。

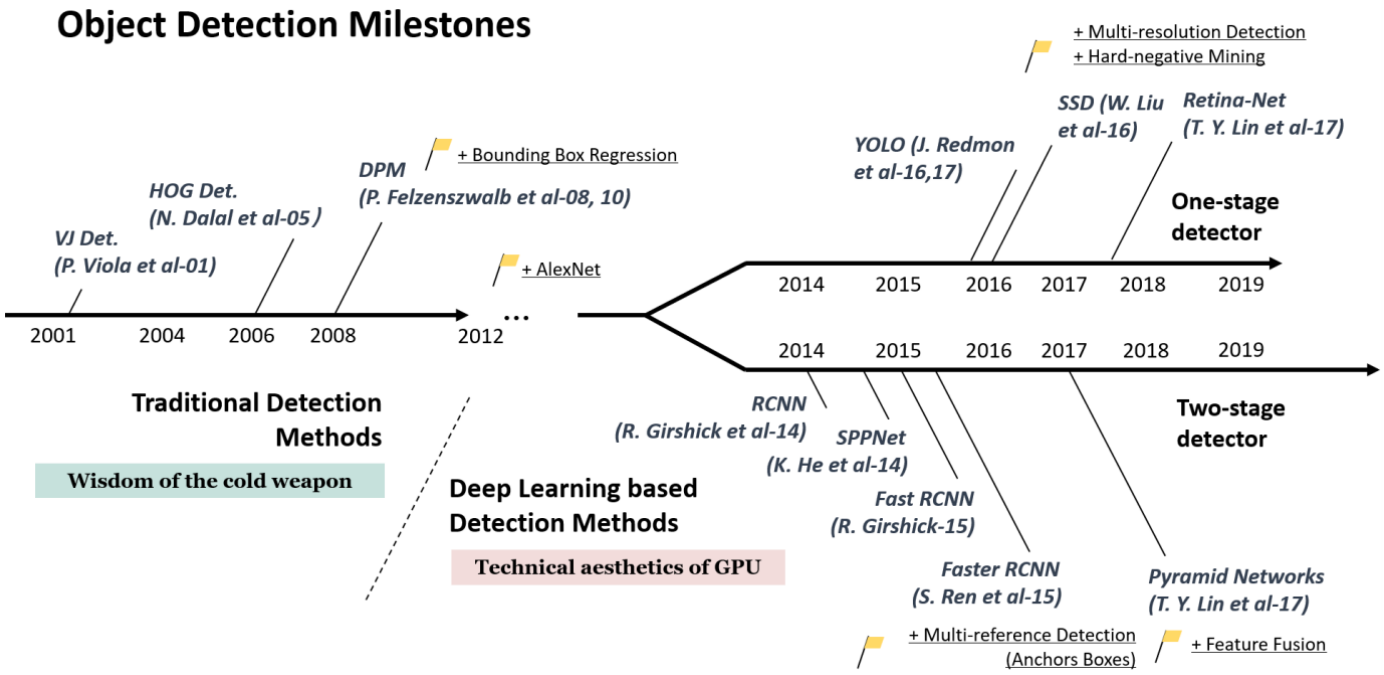


Figure 2. A road map of object detection. Milestone detectors in this figure: VJ Det. [10, 11], HOG Det. [12], DPM [13–15], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], Pyramid Networks [22], Retina-Net [23].

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.

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"冷兵器时代的智慧"。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 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)

- **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。

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".

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当时计算机的算力。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) **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.

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。与其他同时期的目标检测算法[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 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

- **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。

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-

- **可变形部件模型 (DPM)**

DPM 作为 VOC-07，-08 和-09 检测挑战的赢家，是传统目标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"star-model"[13]。后来，R. Girshick 进一步将 star-model 扩展为"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"上下文启发"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，检测速度超过 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

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.

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目标检测？R. Girshick 等于 2014 年率先通过提取具有 CNN 特征的区域（RCNN）进行目标检测来打破僵局[16,41]。从那时起，目标检测开始以前所未有的速度发展。

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".

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"单阶段检测"，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**

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%.

- **RCNN**

RCNN 背后的想法很简单：它首先通过选择性搜索[42]提取一组候选区域（目标候选框）。然后将每个候选区域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存在目标的可能性并识别目标的类别。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％。

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.

尽管 RCNN 取得了很大的进步，但它的缺点是显而易见的：大量重叠候选区域（一个图像超过 2000 个候选框）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**

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

- **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 进行目标检测时，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，

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%).

SPPNet 比 R-CNN 快 20 倍（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.

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**

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.

- **Fast RCNN**

2015 年，R. Girshick 提出了 Fast 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。

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.

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候选区域检测的限制（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 模型生成候选目标区域吗？”后来，Faster R-CNN [19]回答了这个问题。

- **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.

- **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候选区域网络（RPN），该网络提供了几乎无成本的候选区域。从 R-CNN 到 Faster RCNN，目标检测系统的大多数单独的块（例如，候选区域检测，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。

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.)

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）

- **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.

- **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目标的巨大进步。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 Milestones: CNN based One-stage Detectors

2.1.3 里程碑：基于 CNN 的单阶段检测器

- **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.

- **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 "候选区域检测+验证" 的检测范例。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。

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.

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] 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-

- **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

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.

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，而后者仅在其网络顶层上进行检测。

- **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。在目标检测中，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。

Dataset	train		validation		trainval		test	
	images	objects	images	objects	images	objects	images	objects
VOC-2007	2,501	6,301	2,510	6,307	5,011	12,608	4,952	14,976
VOC-2012	5,717	13,609	5,823	13,841	11,540	27,450	10,991	-
ILSVRC-2014	456,567	478,807	20,121	55,502	476,688	534,309	40,152	-
ILSVRC-2017	456,567	478,807	20,121	55,502	476,688	534,309	65,500	-
MS-COCO-2015	82,783	604,907	40,504	291,875	123,287	896,782	81,434	-
MS-COCO-2018	118,287	860,001	5,000	36,781	123,287	896,782	40,670	-
OID-2018	1,743,042	14,610,229	41,620	204,621	1,784,662	14,814,850	125,436	625,282

TABLE 1
Some well-known object detection datasets and their statistics.

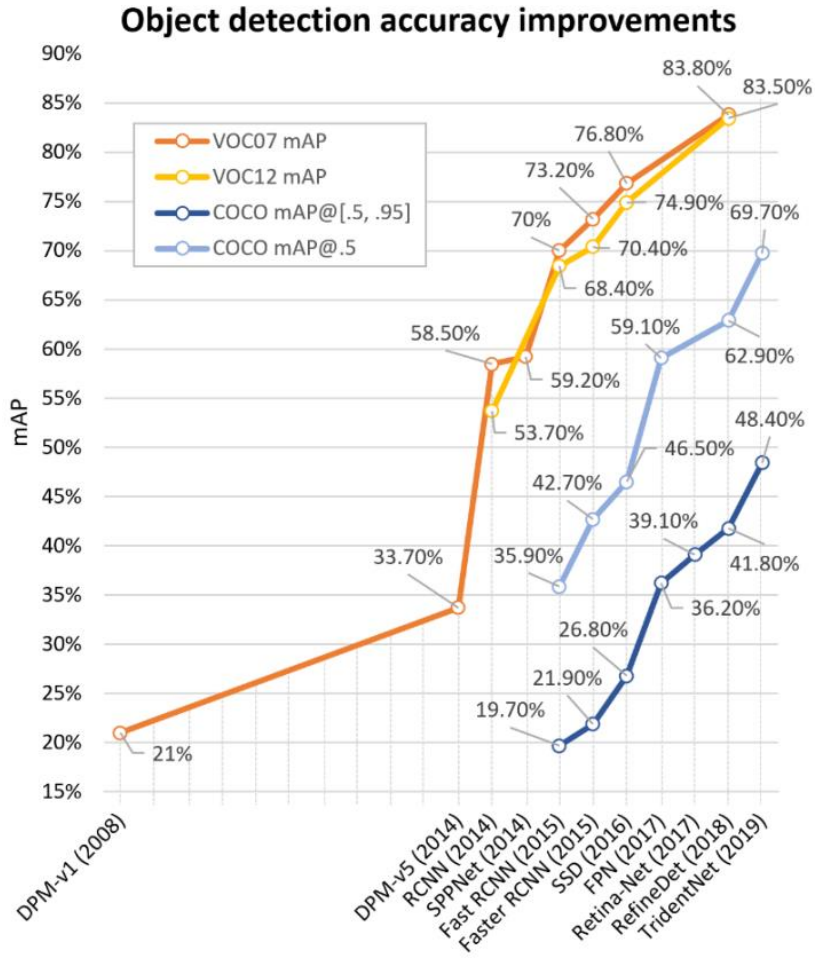


Figure 3. The accuracy improvements of object detection on VOC07, VOC12 and MS-COCO datasets. Detectors in this figure: DPM-v1 [13], DPM-v5 [54], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], SSD [21], FPN [22], Retina-Net [23], RefineDet [55], TridentNet[56].

图 3. VOC07，VOC12 和 MS-COCO 数据集上目标检测的准确性改进。此图中的检测器：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 (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

- **Pascal VOC**

PASCAL 视觉目标类别（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 训练图像+ 12k 标注目标组成，后者由 11k 训练图像+27k 标注目标组成。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：

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.

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。

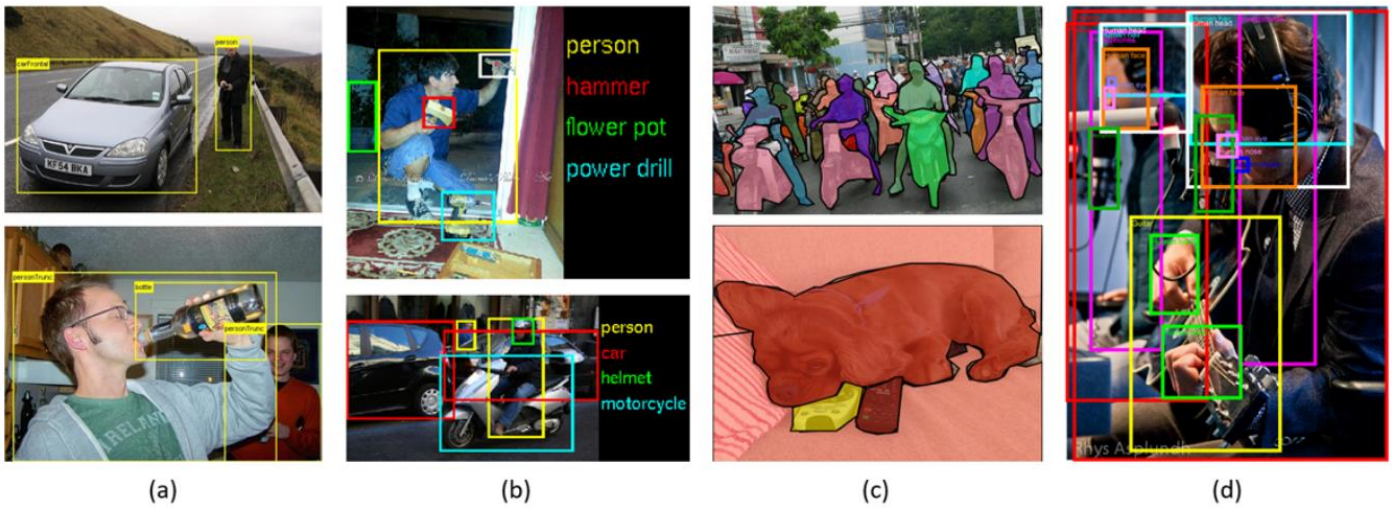


Figure 4. Some example images and annotations in (a) PASCAL-VOC07, (b) ILSVRC, (c) MS-COCO, and (d) Open Images.

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) [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.

- **ILSVRC**

ImageNet 大规模视觉感知挑战赛（ILSVRC）[52]推动了通用目标检测的最新技术水平。ILSVRC 从 2010 年到 2017 年每年组织一次。它包含使用 ImageNet 图像的检测挑战[57]。ILSVRC 检测数据集包含 200 类可视目标。其图像/目标实例的数量比 VOC 大两个数量级。例如，ILSVRC-14 包含 517k 图像和 534k 标注目标。

- **MS-COCO**

MS-COCO [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

- **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，但有更多的目标实例。例如，MS-COCO-17 包含来自 80 个类别的 164k 图像和 897k 个标注目标。与 VOC 和 ILSVRC 相比，MS-COCO 的最大进步是除了边界框标注之外，每个目标还使用实例分割进行标注，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 中的目标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

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.

事实标准。

- **Open Images**

The year of 2018 sees the introduction of the Open Images Detection (OID) challenge [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.

- **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 Open Images 中有两个任务：1）标准目标检测，以及 2）在特定关系中检测成对对象的视觉关系检测。对于目标检测任务，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。

- **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.

- **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/红绿灯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。

Dataset	Year	Description	#Cites
MIT Ped.[30]	2000	One of the first pedestrian detection datasets. Consists of ~500 training and ~200 testing images (built based on the LabelMe database). url: http://cbcl.mit.edu/software-datasets/PedestrianData.html	1515
INRIA [12]	2005	One of the most famous and important pedestrian detection datasets at early time. Introduced by the HOG paper [12]. url: http://pascal.inrialpes.fr/data/human/	24705
Caltech [59, 60]	2009	One of the most famous pedestrian detection datasets and benchmarks. Consists of ~190,000 pedestrians in training set and ~160,000 in testing set. The metric is Pascal-VOC @ 0.5 IoU. url: http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/	2026
KITTI [61]	2012	One of the most famous datasets for traffic scene analysis. Captured in Karlsruhe, Germany. Consists of ~100,000 pedestrians (~6,000 individuals). url: http://www.cvlibs.net/datasets/kitti/index.php	2620
CityPersons [62]	2017	Built based on CityScapes dataset [63]. Consists of ~19,000 pedestrians in training set and ~11,000 in testing set. Same metric with CalTech. url: https://bitbucket.org/shanshanzhang/citypersons	50
EuroCity [64]	2018	The largest pedestrian detection dataset so far. Captured from 31 cities in 12 European countries. Consists of ~238,000 instances in ~47,000 images. Same metric with CalTech.	1

TABLE 2
An overview of some popular pedestrian detection datasets.

Dataset	Year	Description	#Cites
FDDB [65]	2010	Consists of ∼2,800 images and ∼5,000 faces from Yahoo! With occlusions, pose changes, out-of-focus, etc. url: http://vis-www.cs.umass.edu/fddb/index.html	531
AFLW [66]	2011	Consists of ∼26,000 faces and 22,000 images from Flickr with rich facial landmark annotations. url: https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/aflw/	414
IJB [67]	2015	IJB-A/B/C consists of over 50,000 images and videos frames, for both recognition and detection tasks. url: https://www.nist.gov/programs-projects/face-challenges	279
WiderFace [68]	2016	One of the largest face detection dataset. Consists of ∼32,000 images and 394,000 faces with rich annotations i.e., scale, occlusion, pose, etc. url: http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/	193
UFDD [69]	2018	Consists of ∼6,000 images and ∼11,000 faces. Variations include weather-based degradation, motion blur, focus blur, etc. url: http://www.ufdd.info/	1
WildestFaces [70]	2018	With ∼68,000 video frames and ∼2,200 shots of 64 fighting celebrities in unconstrained scenarios. The dataset hasn't been released yet.	2

TABLE 3
An overview of some popular face detection datasets.

Dataset	Year	Description	#Cites
ICDAR [71]	2003	ICDAR2003 is one of the first public datasets for text detection. ICDAR 2015 and 2017 are other popular iterations of the ICDAR challenge [72, 73]. url: http://rrc.cvc.uab.es/	530
STV [74]	2010	Consists of ∼350 images and ∼720 text instances taken from Google StreetView. url: http://tc11.cvc.uab.es/datasets/SVT_1	339
MSRA-TD500 [75]	2012	Consists of ∼500 indoor/outdoor images with Chinese and English texts. url: http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)	413
IIIT5k [76]	2012	Consists of ∼1,100 images and ∼5,000 words from both streets and born-digital images. url: http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/IIIT5K.html	165
Syn90k [77]	2014	A synthetic dataset with 9 million images generated from a 90,000 vocabulary of multiple fonts. url: http://www.robots.ox.ac.uk/~vgg/data/text/	246
COCOText [78]	2016	The largest text detection dataset so far. Built based on MS-COCO, Consists of ∼63,000 images and ∼173,000 text annotations. https://bgshih.github.io/cocotext/.	69

TABLE 4
An overview of some popular scene text detection datasets.

Dataset	Year	Description	#Cites
TLR [79]	2009	Captured by a moving vehicle in Paris. Consists of ∼11,000 video frames and ∼9,200 traffic light instances. url: http://www.lara.prd.fr/benchmarks/trafficlightsrecognition	164
LISA [80]	2012	One of the first traffic sign detection dataset. Consists of ∼6,600 video frames, ∼7,800 instances of 47 US signs. url: http://cvrr.ucsd.edu/LISA/lisa-traffic-sign-dataset.html	325
GTSDB [81]	2013	One of the most popular traffic signs detection dataset. Consists of ∼900 images with ∼1,200 traffic signs capture with various weather conditions during different time of a day. url: http://benchmark.ini.rub.de/?section=gtsdb&subsection=news	259
BelgianTSD [82]	2012	Consists of ∼7,300 static images, ∼120,000 video frames, and ∼11,000 traffic sign annotations of 269 types. The 3D location of each sign has been annotated. url: https://btsd.ethz.ch/shareddata/	224
TT100K [83]	2016	The largest traffic sign detection dataset so far, with ∼100,000 images (2048 x 2048) and ∼30,000 traffic sign instances of 128 classes. Each instance is annotated with class label, bounding box and pixel mask. url: http://cg.cs.tsinghua.edu.cn/traffic%2Dsign/	111
BSTL [84]	2017	The largest traffic light detection dataset. Consists of ∼5000 static images, ∼8300 video frames, and ∼24000 traffic light instances. https://hci.iwr.uni-heidelberg.de/node/6132	21

TABLE 5
An overview of some popular traffic light detection and traffic sign detection datasets.

Dataset	Year	Description	#Cites
TAS [85]	2008	Consists of 30 images of 729x636 pixels from Google Earth and ∼1,300 vehicles. url: http://ai.stanford.edu/~gaheitz/Research/TAS/	419
OIRDS [86]	2009	Consists for 900 images (0.08-0.3m/pixel) captured by aircraft-mounted camera and 1,800 annotated vehicle targets. url: https://sourceforge.net/projects/oirds/	32
DLR3K [87]	2013	The most frequently used datasets for small vehicle detection. Consists of 9,300 cars and 160 trucks. url: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-5431/9230_read-42467/	68
UCAS-AOD [88]	2015	Consists of ∼900 Google Earth images, ∼2,800 vehicles and ∼3,200 airplanes. url: http://www.ucassdl.cn/resource.asp	19
VeDAI [89]	2016	Consists of ∼1,200 images (0.1-0.25m/pixel), ∼3,600 targets of 9 classes. Designed for detecting small target in remote sensing images. url: https://downloads.greyc.fr/vedai/	65
NWPU-VHR10 [90]	2016	The most frequently used remote sensing detection dataset in recent years. Consists of ∼800 images (0.08-2.0m/pixel) and ∼3,800 remote sensing targets of ten classes (e.g., airplanes, ships, baseball diamonds, tennis courts, etc). url: http://jiong.tea.ac.cn/people/JunweiHan/NWPUVHR10dataset.html	204
LEVIR [91]	2018	Consists of ∼22,000 Google Earth images and ∼10,000 independently labeled targets (airplane, ship, oil-pot). url: https://pan.baidu.com/s/1geTwAVD	15
DOTA [92]	2018	The first remote sensing detection dataset to incorporate rotated bounding boxes. Consists of ∼2,800 Google Earth images and ∼200,000 instances of 15 classes. url: https://captain-whu.github.io/DOTA/dataset.html	32
xView [93]	2018	The largest remote sensing detection dataset so far. Consists of ∼1,000,000 remote sensing targets of 60 classes (0.3m/pixel), covering1,415km^2 of land area. url: http://xviewdataset.org	10

TABLE 6
An overview of some remote sensing target detection datasets.

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

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，加州理工学院创建了行人检测基准[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，最常用的目标检测评估是“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 随后成为多年来目标检测问题的事实上的度量标准。

"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).

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目标定位，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）。

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.

最近，在 Open Images 数据集中存在评估标准的一些进一步发展，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"定位召回精度"[94]。尽管最近发生了变化，但基于 VOC / COCO 的 mAP 仍然是最常用的目标检测评估指标。

2.3 Technical Evolution in Object Detection

In this section, we will introduce some important building blocks of a detection system and their technical evolutions in the past 20 years.

2.3 目标检测的技术演进

在本节中，我们将介绍过去 20 年中检测系统的一些重要组成部分及其技术演变。

2.3.1 Early Time's Dark Knowledge

The early time's object detection (before 2000) did not follow a unified detection philosophy like sliding window detection. Detectors at that time were usually designed based on low-level and mid-level vision as follows.

2.3.1 早期的暗黑知识

早期的目标检测（2000 年之前）没有遵循统一的检测理念，如滑动窗口检测。当时的检测器通常基于如下的低级和中级视觉设计。

- **Components, shapes and edges**

"Recognition-by-components", as an important cognitive theory [98], has long been the core idea of image recognition and object detection [13, 99, 100]. Some early researchers framed the object detection as a measurement of similarity between the object components, shapes and contours, including Distance Transforms [101], Shape Contexts [35], and Edgelet [102], etc. Despite promising initial results, things did not work out well on more complicated detection problems. Therefore, machine learning based detection methods were beginning to prosper.

- **部件，形状和边缘**

"部件识别"作为一种重要的认知理论[98]，长期以来一直是图像识别和目标检测的核心思想[13,99,100]。一些早期的研究人员将目标检测定义为目标部件，形状和轮廓之间相似性的测量，包括距离变换[101]，形状上下文[35]和 Edgelet [102]等。尽管有希望的初步结果，事情并没有在更复杂的检测问题上做得很好。因此，基于机器学习的检测方法开始繁荣起来。

Machine learning based detection has gone through multiple periods, including the statistical models of appearance (before 1998), wavelet feature representations (1998-2005), and gradient-based representations (2005-2012).

基于机器学习的检测经历了多个时期，包括外观统计模型（1998 年之前），小波特征表示（1998-2005）和基于梯度的表示（2005-2012）。

Building statistical models of an object, like Eigenfaces [95, 106] as shown in Fig 5 (a), was the first wave of learning based approaches in object detection history. In 1991, M.Turk et al. achieved real-time face detection in a lab environment by using Eigenface decomposition [95]. Compared with the rule-based or template based approaches of its time [107, 108], a statistical model better provides holistic descriptions of an object's appearance by learning task-specific knowledge from data.

构建目标的统计模型，如图 5 (a)所示的 Eigenfaces [95,106]，是目标检测历史中第一波基于学习的方法。1991 年，M.Turk 等人通过使用脸特征分解[95]在实验室环境中实现了实时人脸检测。与当时基于规则或基于模板的方法[107,108]相比，统计模型通过从数据中学习任务特定的知识，更好地提供了对象外观的整体描述。

Wavelet feature transform started to dominate visual recognition and object detection since 2000. The essence of this group of methods is learning by transforming an image from pixels to a set of wavelet coefficients. Among these methods, the Haar wavelet, owing to its high computational efficiency, has been mostly used in many object detection tasks, such as general object detection [29], face detection [10, 11, 109], pedestrian detection [30, 31], etc. Fig 5 (d) shows a set of Haar wavelets basis learned by a VJ detector [10, 11] for human faces.

自 2000 年以来，小波特征变换开始主导视觉识别和目标检测。这组方法的本质是通过将图像从像素变换为一组小波系数来学习。在这些方法中，Haar 小波由于其高计算效率，主要用于许多目标检测任务，如一般目标检测[29]，人脸检测[10,11,109]，行人检测[30,31]。图 5（d）示出了由人脸 VJ 检测器[10,11]学习的一组 Haar 小波原理。

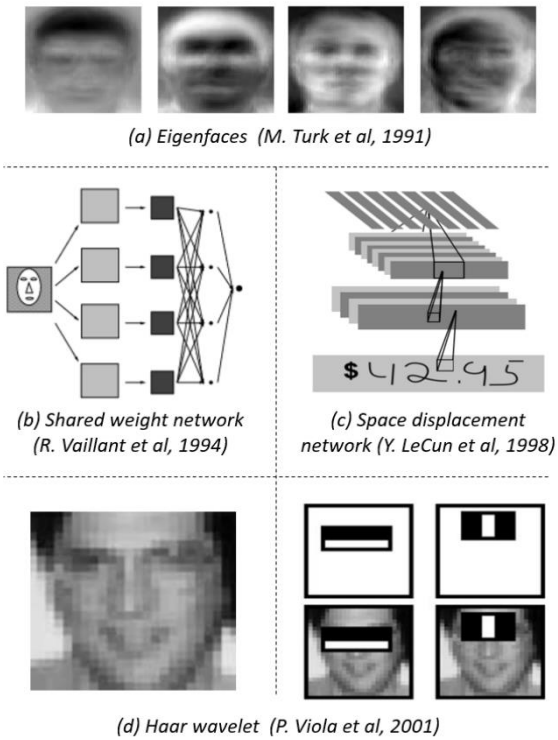


Figure 5. Some well-known detection models of the early time: (a) Eigenfaces [95], (b) Shared weight networks [96], (c) Space displacement networks (Lenet-5) [97], (d) Haar wavelets of VJ detector [10].

图 5. 一些众所周知的早期检测模型：（a）Eigenfaces [95]，（b）共享权重网络[96]，（c）空间位移网络（Lenet-5）[97]，（d）Haar 小波 VJ 检测器[10]。

- **Early time's CNN for object detection**

The history of using CNN to detecting objects can be traced back to the 1990s [96], where Y. LeCun et al. have made great contributions at that time. Due to limitations in computing resources, CNN models at the time were much

- **早期的基于 CNN 的目标检测**

使用 CNN 检测物体的历史可以追溯到 20 世纪 90 年代[96]，其中 Y. LeCun 等人当时做出了巨大的贡献。由于计算资源的限制，当时的 CNN 模型比现在更小更浅。尽管如此，计算效率仍被认为是早期基于 CNN 的检测

smaller and shallower than those of today. Despite this, the computational efficiency was still considered as one of the tough nuts to crack in early times's CNN based detection models. Y. LeCun et al. have made a series of improvements like "shared-weight replicated neural network" [96] and "space displacement network" [97] to reduce the computations by extending each layer of the convolutional network so as to cover the entire input image, as shown in Fig. 5 (b)-(c). In this way, the feature of any location of the entire image can be extracted by taking only one time of forward propagation of the network. This can be considered as the prototype of today's fully convolutional networks (FCN) [110, 111], which was proposed almost 20 years later. CNN also has been applied to other tasks such as face detection [112, 113] and hand tracking [114] of its time.

模型中难以破解的难题之一。Y. LeCun 等已经做了一系列改进，如"共享权重神经网络"[96]和"空间位移网络"[97]，通过扩展卷积网络的每一层来减少计算，以覆盖整个输入图像，如图 5(b)-(c) 所示。以这种方式，可以通过仅采用网络的前向传播一次来提取整个图像的任何位置的特征。这可以被认为是今天完全卷积网络（FCN）[110,111]的原型，这是近 20 年后提出的。CNN 还应用于其他任务，如人脸检测[112,113]和手部跟踪[114]。

2.3.2 Technical Evolution of Multi-Scale Detection

Multi-scale detection of objects with "different sizes" and "different aspect ratios" is one of the main technical challenges in object detection. In the past 20 years, multiscale detection has gone through multiple historical periods: "feature pyramids and sliding windows (before 2014)", "detection with object proposals (2010-2015)", "deep regression (2013-2016)", "multi-reference detection (after 2015)", and "multi-resolution detection (after 2016)", as shown in Fig. 6.

2.3.2 多尺度检测的技术演进

具有"不同尺寸"和"不同纵横比"的多尺度目标检测是目标检测中的主要技术挑战之一。在过去的 20 年中，多尺度检测经历了多个历史时期："特征金字塔和滑动窗口（2014 年之前）"，"候选目标区域检测（2010-2015）"，"深度回归（2013-2016）"，"多参考检测（2015 年后）"和"多分辨率检测（2016 年后）"，如图 6 所示。

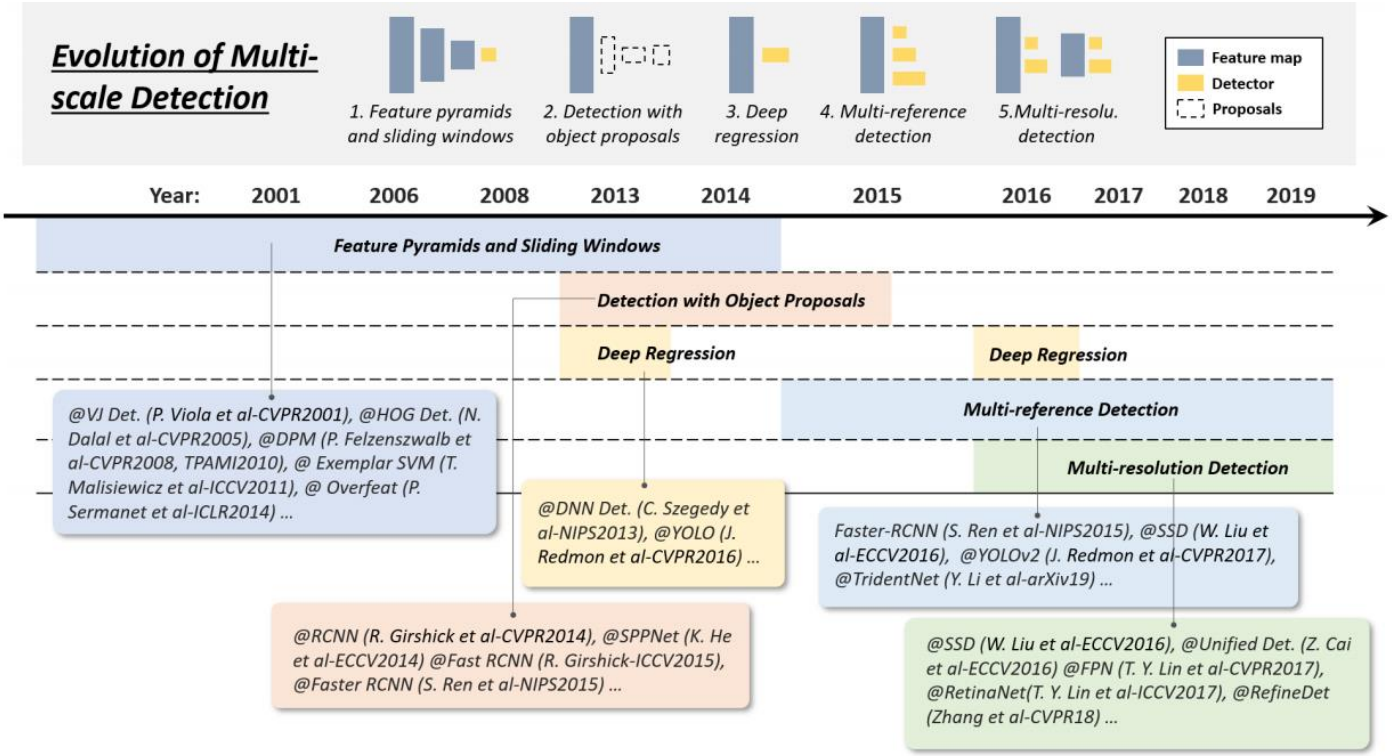


Figure 6. Evolution of multi-scale detection techniques in object detection from 2001 to 2019: 1) feature pyramids and sliding windows, 2) detection with object proposals, 3) deep regression, 4) multi-reference detection, and 5) multi-

图 6. 2001 年至 2019 年目标检测中多尺度检测技术的演变：1）特征金字塔和滑动窗口，2）具有对象建议的检测，3）深度回归，4）多参考检测，以及 5）多分辨率检测。此图中的检测器：VJ Det. [10], HOG Det.

resolution detection. Detectors in this figure: VJ Det. [10], HOG Det. [12], DPM [13, 15], Exemplar SVM [36], Overfeat [103], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], DNN Det. [104], YOLO [20], YOLO-v2 [48], SSD [21], Unified Det. [105], FPN [22], RetinaNet [23], RefineDet [55], TridentNet [56].

[12], DPM [13, 15], Exemplar SVM [36], Overfeat [103], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], DNN Det. [104], YOLO [20], YOLO-v2 [48], SSD [21], Unified Det. [105], FPN [22], RetinaNet [23], RefineDet [55], TridentNet [56]。

- **Feature pyramids + sliding windows (before 2014)**

With the increase of computing power after the VJ detector, researchers started to pay more attention to an intuitive way of detection by building "feature pyramid + sliding windows". From 2004 to 2014, a number of milestone detectors were built based on this detection paradigm, including the HOG detector, DPM, and even the Overfeat detector [103] of the deep learning era (winner of ILSVRC13 localization task).

- **特征金字塔+滑动窗口（2014 年之前）**

随着 VJ 检测器之后计算能力的提高，研究人员开始通过构建"特征金字塔+滑动窗口"来更加注重直观的检测方式。从 2004 年到 2014 年，基于这种检测范例构建了许多里程碑检测器，包括 HOG 检测器，DPM，甚至深度学习时代的 Overfeat 检测器[103]（ILSVRC13 定位任务的获胜者）。

Early detection models like VJ detector and HOG detector were specifically designed to detect objects with a "fixed aspect ratio" (e.g., faces and upright pedestrians) by simply building the feature pyramid and sliding fixed size detection window on it. The detection of "various aspect ratios" was not considered at that time. To detect objects with a more complex appearance like those in PASCAL VOC, R. Girshick et al. began to seek better solutions outside the feature pyramid. The "mixture model" [15] was one of the best solutions at that time, by training multiple models to detect objects with different aspect ratios. Apart from this, exemplar-based detection [36, 115] provided another solution by training individual models for every object instance (exemplar) of the training set.

像 VJ 检测器和 HOG 检测器这样的早期检测模型专门设计用于通过简单地在图像上构建特征金字塔和滑动固定尺寸检测窗口来检测具有"固定纵横比"（例如，人脸和直立行人）的物体。当时没有考虑"各种纵横比"的检测。检测具有更复杂外观的物体，如 PASCAL VOC，R.Girshick 等人开始在特征金字塔之外寻求更好的解决方案。"混合模型"[15]是当时最好的解决方案之一，通过训练多个模型来检测具有不同纵横比的物体。除此之外，基于样本的检测[36,115]通过训练集的每个对象实例（范例）的单个模型训练提供了另一种解决方案。

As objects in the modern datasets (e.g., MS-COCO) become more diversified, the mixture model or exemplarbased methods inevitably lead to more miscellaneous detection models. A question then naturally arises: is there a unified multi-scale approach to detect objects of different aspect ratios? The introduction of "object proposals" (to be introduced) has answered this question.

随着现代数据集（例如，MS-COCO）中的对象变得更加多样化，混合模型或基于示例的方法不可避免地导致更多杂项检测模型。那么自然会出现一个问题：是否存在统一的多尺度方法来检测不同宽高比的物体？ 引入"目标候选区域"（将介绍）已经回答了这个问题。

- **Detection with object proposals (2010-2015)**

Object proposals refer to a group of class-agnostic candidate boxes that likely to contain any objects. It was first time applied in object detection in 2010 [116]. Detection with object proposals helps to avoid the exhaustive sliding window search across an image.

- **采用候选区域检测(2010-2015)**

目标候选区域是指一组可能包含任何目标的类别不可知候选框。它于 2010 年首次应用于目标检测[116]。使用目标候选区域进行检测有助于避免在图像上进行暴力的滑动窗口搜索。

An object proposal detection algorithm should meet the following three requirements: 1) high recall rate, 2) high

目标候选区域检测算法应满足以下三个要求：1)高召回率，2)高定位精度，3)在前两个要求的基础上，提高精

localization accuracy, and 3) on basis of the first two requirements, to improve precision and reduce processing time. Modern proposal detection methods can be divided into three categories: 1) segmentation grouping approaches [42, 117–119], 2) window scoring approaches [116, 120–122], and 3) neural network based approaches [123–128]. We refer readers to the following papers for a comprehensive review of these methods [129, 130].

度，减少处理时间。现代候选区域检测方法可以分为三类：1)分割分组方法[42,117-119]，2)窗口评分方法[116,120-122]，和 3)基于神经网络的方法[123-128]。我们向读者推荐以下论文，以全面审查这些方法[129,130]。

Early time's proposal detection methods followed a bottom-up detection philosophy [116, 120] and were deeply affected by visual saliency detection. Later, researchers started to move to low-level vision (e.g., edge detection) and more careful handcrafted skills to improve the localization of candidate boxes [42, 117–119, 122, 131]. After 2014, with the popularity of deep CNN in visual recognition, the topdown, learning-based approaches began to show more advantages in this problem [19, 121, 123, 124]. Since then, the object proposal detection has evolved from the bottom-up vision to "overfitting to a specific set of object classes", and the distinction between detectors and proposal generators is becoming blurred [132].

早期的候选区域检测方法遵循自下而上的检测理念[116,120]，并且受到视觉显着性检测的深刻影响。后来，研究人员开始转向低级视觉（例如边缘检测）和更精细的手工技巧，以改善候选盒的定位[42,117-119,122,131]。2014 年之后，随着深度 CNN 在视觉识别中的普及，自上而下，基于学习的方法开始在这个问题上显示出更多的优势[19,121,123,124]。从那时起，目标候选区域检测已经从自下而上的视觉演变为"过拟合到一组特定的对象类"，并且检测器和候选区域生成器之间的区别变得模糊[132]。

As "object proposal" has revolutionized the sliding window detection and has quickly dominated the deep learning based detectors, in 2014-2015, many researchers began to ask the following questions: what is the main role of the object proposals in detection? Is it for improving accuracy, or simply for detection speed up? To answer this question, some researchers have tried to weaken the role of the proposals [133] or simply perform sliding window detection on CNN features [134–138], but none of them obtained satisfactory results. The proposal detection has soon slipped out of sight after the rise of one-stage detectors and "deep regression" techniques (to be introduced).

由于"目标候选区域"彻底改变了滑动窗口检测并迅速占据了基于深度学习的检测器，在 2014-2015 年，许多研究人员开始提出以下问题：候选区域在检测中的主要作用是什么？它是为了提高准确度，还是仅仅为了提高检测速度？为了回答这个问题，一些研究人员试图削弱候选区域的作用[133]或简单地对 CNN 特征进行滑动窗口检测[134-138]，但没有一个获得满意的结果。在单阶段检测器和"深度回归"技术（将要介绍）的兴起之后，候选区域检测很快就消失了。

- **Deep regression (2013-2016)**

In recent years, as the increase of GPU's computing power, the way people deal with multi-scale detection has become more and more straight forward and brute-force. The idea of using the deep regression to solve multi-scale problems is very simple, i.e., to directly predict the coordinates of a bounding box based on the deep learning features [20, 104]. The advantage of this approach is that it is simple and easy to implement while the disadvantage is the localization may not be accurate enough especially for some small objects. "Multi-reference detection" (to be introduced) has latter solved this problem.

- **深度回归 (2013-2016)**

近年来，随着 GPU 计算能力的提高，人们处理多尺度检测的方式变得越来越直接和蛮力。使用深度回归来解决多尺度问题的想法非常简单，即，基于深度学习特征直接预测边界框的坐标[20,104]。这种方法的优点是它简单易行，而缺点是定位可能不够准确，特别是对于一些小物体。"多参考检测"（待介绍）随后解决了这个问题。

- **Multi-reference/-resolution detection (after 2015)**

Multi-reference detection is the most popular framework for multi-scale object detection [19, 21, 44, 48]. Its main idea is to pre-define a set of reference boxes (a.k.a. anchor boxes) with different sizes and aspect-ratios at different locations of an image, and then predict the detection box based on these references.

A typical loss of each predefined anchor box consists of two parts: 1) a cross-entropy loss for category recognition and 2) an L1/L2 regression loss for object localization. A general form of the loss function can be written as follows:

$$L(p, p^*, t, t^*) = L_{cls.}(p, p^*) + \beta I(t) L_{loc.}(t, t^*)$$
$$I(t) = \begin{cases} 1 & \text{IOU}\{a, a^*\} > \eta \\ 0 & \text{else} \end{cases} \quad (1)$$

where t and t^* are the locations of predicted and ground-truth bounding box, p and p^* are their category probabilities. IOU{a, a^*} is the IOU between the anchor a and its ground-truth a^* . η is an IOU threshold, say, 0.5. If an anchor that does not cover any objects, its localization loss does not count in the final loss.

Another popular technique in the last two years is multiresolution detection [21, 22, 55, 105], i.e. by detecting objects of different scales at different layers of the network. Since a CNN naturally forms a feature pyramid during its forward propagation, it is easier to detect larger objects in deeper layers and smaller ones in shallower layers. Multi-reference and multi-resolution detection have now become two basic building blocks in the state of the art object detection systems.

2.3.3 Technical Evolution of Bounding Box Regression

The Bounding Box (BB) regression is an important technique in object detection. It aims to refine the location of a predicted bounding box based on the initial proposal or the anchor box. In the past 20 years, the evolution of BB regression has gone through three historical periods: "without BB regression (before 2008)", "from BB to BB (2008-2013)", and "from feature to BB (after 2013)". Fig. 7 shows the evolutions of bounding box regression.

- **Without BB regression (before 2008)**

Most of the early detection methods such as VJ detector and HOG detector do not use BB regression, and usually directly consider the sliding window as the detection result. To obtain accurate locations of an object, researchers have no choice but to build very dense pyramid and slide the detector densely on each location.

- **多参考/多分辨率检测(2013-2016)**

多参考检测是最常用的多尺度目标检测框架[19,21,44,48]。其主要思想是在图像的不同位置预先定义一组具有不同尺寸和纵横比的参考框（称为锚盒），然后基于这些参考预测检测框。

每个预定义锚盒的典型损失由两部分组成：1）类别识别的交叉熵损失和 2）对象定位的 L1/L2 回归损失。损失函数的一般形式可以写成如下：

其中 t 和 t^*是预测和真值边界框的位置，p 和 p^*是它们的类别概率。IOU {a，a^*}是锚 a 和它的地面实例 a^*之间的 IOU。η 是 IOU 阈值，比如 0.5。如果锚不覆盖任何物体，其定位损失不计入最终损失。

过去两年中另一种流行的技术是多分辨率检测[21,22,55,105]，即通过在网络的不同层检测不同尺度的物体。由于 CNN 在其向前传播期间自然地形成特征金字塔，因此更容易检测较深层中的较大对象和较浅层中较小的对象。多参考和多分辨率检测现在已成为现有技术对象检测系统中的两个基本构建块。

2.3.3 边界框回归的技术演进

边界框(BB)回归是对象检测中的重要技术。它旨在根据初始获选区域或锚框优化预测边界框的位置。在过去的 20 年中，BB 回归的演变经历了三个历史时期："没有 BB 回归（2008 年之前）"，"从 BB 到 BB（2008-2013）"，以及"从特征到 BB（2013 年之后）"。图 7 显示了边界框回归的演变。

- **没有 BB 回归(2008 年之前)**

大多数早期检测方法如 VJ 检测器和 HOG 检测器都不使用 BB 回归，通常直接将滑动窗口视为检测结果。为了获得物体的准确位置，研究人员别无选择，只能建造非常密集的金字塔，并在每个位置密集地滑动检测器。

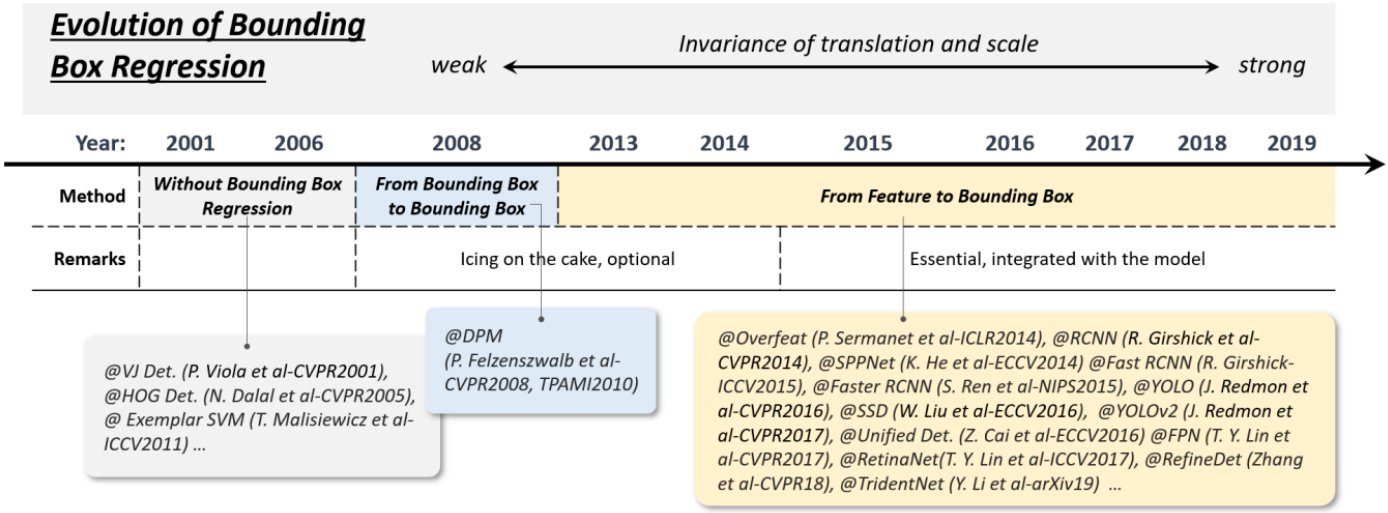


Figure 7. Evolution of bounding box regression techniques in object detection from 2001 to 2019. Detectors in this figure: VJ Det. [10], HOG Det. [12], Exemplar SVM [36], DPM [13, 15], Overfeat [103], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], YOLO-v2 [48], Unified Det. [105], FPN [22], RetinaNet [23], RefineDet [55], TridentNet [56].

图 7. 2001 年至 2019 年目标检测中边界框回归技术的演变。图中的检测器：VJ Det. [10], HOG Det. [12], Exemplar SVM [36], DPM [13, 15], Overfeat [103], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], YOLO-v2 [48], Unified Det. [105], FPN [22], RetinaNet [23], RefineDet [55], TridentNet [56]。

- **From BB to BB (2008-2013)**

The first time that BB regression was introduced to an object detection system was in DPM [15]. The BB regression at that time usually acted as a post-processing block, thus it is optional. As the goal in the PASCAL VOC is to predict single bounding box for each object, the simplest way for a DPM to generate final detection should be directly using its root filter locations. Later, R. Girshick et al. introduced a more complex way to predict a bounding box based on the complete configuration of an object hypothesis and formulate this process as a linear least-squares regression problem [15]. This method yields noticeable improvements of the detection under PASCAL criteria.

- **从 BB 到 BB (2008-2013)**

第一次将 BB 回归引入物体检测系统是在 DPM [15]。当时的 BB 回归通常充当后处理块，因此它是可选的。由于 PASCAL VOC 的目标是预测每个对象的单个边界框，因此 DPM 生成最终检测的最简单方法应该是直接使用其根过滤器位置。后来，R.Girshick 等人引入了一种更复杂的方法来预测基于对象假设的完整配置的边界框，并将此过程表示为线性最小二乘回归问题[15]。该方法在 PASCAL 标准下产生显着的检测改善。

- **From features to BB (after 2013)**

After the introduction of Faster RCNN in 2015, BB regression no longer serves as an individual post-processing block but has been integrated with the detector and trained in an end-to-end fashion. At the same time, BB regression has evolved to predicting BB directly based on CNN features. In order to get more robust prediction, the smooth-L1 function [19] is commonly used,

$$L(t) = \begin{cases} 5t^2 & |t| \leqslant 0.1 \\ |t| - 0.05 & \text{else,} \end{cases} \quad (2)$$

or the root-square function [20],

$$L(x, x^*) = (\sqrt{x} - \sqrt{x^*})^2, \quad (3)$$

as their regression loss, which are more robust to the outliers

- **从特征到 BB (2013 年之后)**

在 2015 年推出 Faster RCNN 之后，BB 回归不再作为单独的后处理模块，而是与检测器集成并以端到端的方式进行训练。同时，BB 回归已演变为直接基于 CNN 特征预测 BB。为了获得更强大的预测，通常使用 smooth-L1 函数[19]，

或平方根函数[20]，

作为他们的回归损失，对于异常值比 DPM 中使用的最

than the least square loss used in DPM. Some researchers also choose to normalize the coordinates to get more robust results [18, 19, 21, 23].

小平方损失更稳健。一些研究人员还选择标准化坐标以获得更稳健的结果[18,19,21,23]。

2.3.4 Technical Evolution of Context Priming

Visual objects are usually embedded in a typical context with the surrounding environments. Our brain takes advantage of the associations among objects and environments to facilitate visual perception and cognition [160]. Context priming has long been used to improve detection. There are three common approaches in its evolutionary history: 1) detection with local context, 2) detection with global context, and 3) context interactives, as shown in Fig. 8.

2.3.4 上下文启发的演进

视觉对象通常嵌入在周围的典型环境中。我们的大脑利用物体和环境之间的联系来促进视觉感知和认知[160]。长期以来，上下文启发一直被用于改进检测。在其进化历史中有三种常见的方法：1）用局部上下文检测，2）用全局上下文检测，和 3）上下文交互，如图 8 所示。

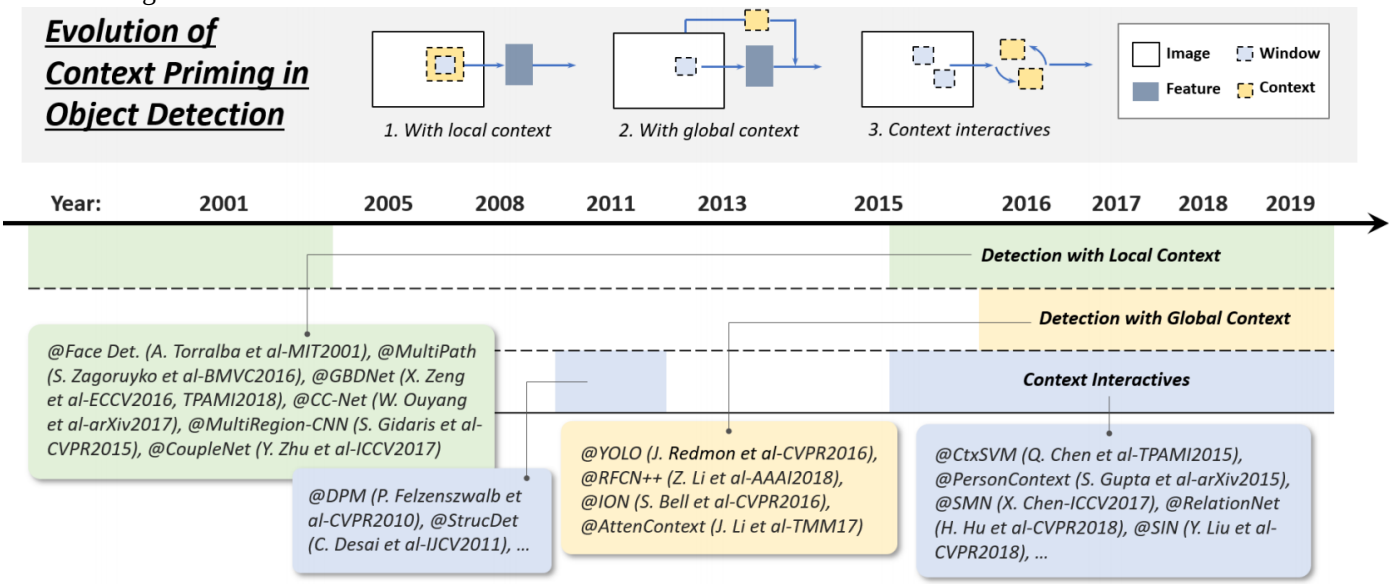


Figure 8. Evolution of context priming in object detection from 2001 to 2019: 1) detection with local context, 2) detection with global context, 3) detection with context interactives. Detectors in this figure: Face Det. [139], MultiPath [140], GBDNet [141, 142], CC-Net [143], MultiRegion-CNN [144], CoupleNet [145], DPM [14, 15], StructDet [146], YOLO [20], RFCN++ [147], ION [148], AttenContext [149], CtxSVM [150], PersonContext [151], SMN [152], RetinaNet [23], SIN [153].

图 8. 从 2001 年到 2019 年的目标检测中的上下文启发的演进：1）使用局部上下文检测，2）使用全局上下文检测，3）使用上下文交互检测。
此图中的检测器：Face Det. [139], MultiPath [140], GBDNet [141, 142], CC-Net [143], MultiRegion-CNN [144], CoupleNet [145], DPM [14, 15], StructDet [146], YOLO [20], RFCN++ [147], ION [148], AttenContext [149], CtxSVM [150], PersonContext [151], SMN [152], RetinaNet [23], SIN [153]。

- **Detection with local context**

Local context refers to the visual information in the area that surrounds the object to detect. It has long been acknowledged that local context helps improve object detection. At early 2000s, Sinha and Torralba [139] found that inclusion of local contextual regions such as the facial bounding contour substantially improves face detection performance. Dalal and Triggs also found that incorporating a small amount of background information improves the accuracy of pedestrian detection [12]. Recent deep learning based detectors can also be improved with local context by

- **局部上下文检测**

局部上下文是指围绕要检测目标区域中的视觉信息。人们早就认识到，局部环境有助于改善目标检测。在 2000 年早期，Sinha 和 Torralba [139]发现包含诸如面部边界轮廓的局部上下文区域实质上改善了面部检测性能。Dalal 和 Triggs 还发现，加入少量背景信息可以提高行人检测的准确性[12]。通过简单地扩大网络的感知领域或目标候选区域的大小[140-145,161]，最近的基于深度学习的检测器也可以通过局部环境得到改善。

simply enlarging the networks' receptive field or the size of object proposals [140–145, 161].

- **Detection with global context**

Global context exploits scene configuration as an additional source of information for object detection. For early time's object detectors, a common way of integrating global context is to integrate a statistical summary of the elements that comprise the scene, like Gist [160]. For modern deep learning based detectors, there are two methods to integrate global context. The first way is to take advantage of large receptive field (even larger than the input image) [20] or global pooling operation of a CNN feature [147]. The second way is to think of the global context as a kind of sequential information and to learn it with the recurrent neural networks [148, 149].

- 全局上下文检测

全局上下文利用场景配置作为目标检测的附加信息源。对于早期的目标检测器，整合全局上下文的一种常用方法是整合构成场景元素的统计汇总，如 Gist [160]。对于基于现代深度学习的检测器，有两种方法可以整合全局背景。第一种方法是利用大的感受野（甚至大于输入图像）[20]或 CNN 特征的全局合并操作[147]。第二种方法是将全局背景视为一种顺序信息，并用递归神经网络来学习[148,149]。

- Context interactive

Context interactive refers to the piece of information that conveys by the interactions of visual elements, such as the constraints and dependencies. For most object detectors, object instances are detected and recognized individually without exploiting their relations. Some recent researches have suggested that modern object detectors can be improved by considering context interactives. Some recent improvements can be grouped into two categories, where the first one is to explore the relationship between individual objects [15, 146, 150, 152, 162], and the second one is to explore modeling the dependencies between objects and scenes [151, 151, 153].

- 上下文交互

上下文交互是指通过视觉元素的交互传达的信息片段，例如约束和依赖关系。对于大多数目标检测器，可以单独检测和识别目标实例，而不会利用它们的关系。最近的一些研究表明，通过考虑上下文交互作用可以改进现代目标检测器。最近的一些改进可以分为两类，第一类是探索各个目标之间的关系[15,146,150,152,162]，第二类是探索目标和场景之间的依赖关系建模[151]，151,153]。

2.3.5 Technical Evolution of Non-Maximum Suppression

Non-maximum suppression (NMS) is an important group of techniques in object detection. As the neighboring windows usually have similar detection scores, the non-maximum suppression is herein used as a post-processing step to remove the replicated bounding boxes and obtain the final detection result. At early times of object detection, NMS was not always integrated [30]. This is because the desired output of an object detection system was not entirely clear at that time. During the past 20 years, NMS has been gradually developed into the following three groups of methods: 1) greedy selection, 2) bounding box aggregation, and 3) learning to NMS, as shown in Fig. 9.

2.3.5 非极大抑制的演进

非最大抑制（NMS）是目标检测中的一组重要技术。由于相邻窗口通常具有相似的检测分数，因此非极大抑制在此用作后处理步骤以移除重复的边界框并获得最终的检测结果。在目标检测的早期，NMS 并不总是被整合[30]。这是因为此时目标检测系统的期望输出并不完全清楚。在过去的 20 年中，NMS 逐渐发展成以下三组方法：1）贪婪选择，2）边界框聚合，3）NMS 学习，如图 9 所示。

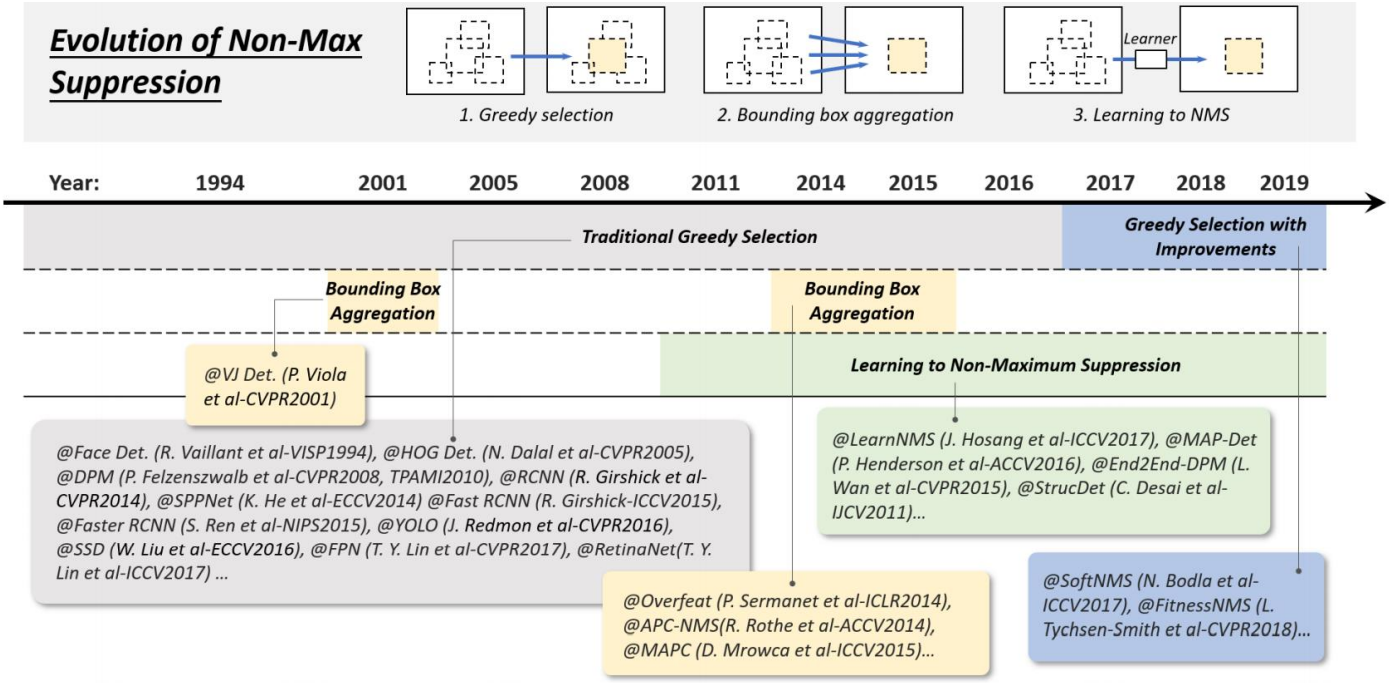


Figure 9. Evolution of non-max suppression (NMS) techniques in object detection from 1994 to 2019: 1) Greedy selection, 2) Bounding box aggregation, and 3) Learn to NMS. Detectors in this figure: VJ Det. [10], Face Det. [96], HOG Det. [12], DPM [13, 15], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], FPN [22], RetinaNet [23], LearnNMS [154], MAP-Det [155], End2End-DPM [136], StrucDet [146], Overfeat [103], APC-NMS [156], MAPC [157], SoftNMS [158], FitnessNMS [159].

图 9. 1994 年至 2019 年目标检测中非极大抑制（NMS）技术的演变：1）贪心选择，2）边界框聚合，3）NMS 学习。此图中的检测器：VJ Det. [10], Face Det. [96], HOG Det. [12], DPM [13, 15], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], FPN [22], RetinaNet [23], LearnNMS [154], MAP-Det [155], End2End-DPM [136], StrucDet [146], Overfeat [103], APC-NMS [156], MAPC [157], SoftNMS [158], FitnessNMS [159]。

- **Greedy selection**

Greedy selection is an old fashioned but the most popular way to perform NMS in object detection. The idea behind this process is simple and intuitive: for a set of overlapped detections, the bounding box with the maximum detection score is selected while its neighboring boxes are removed according to a predefined overlap threshold (say, 0.5). The above processing is iteratively performed in a greedy manner.

- **贪心选择**

贪心选择是一种老式但是在目标检测中执行 NMS 的最流行的方式。该过程背后的想法是简单和直观的：对于一组重叠检测，选择具有最大检测分数的边界框，同时根据预定义的重叠阈值（例如，0.5）移除其相邻框。以贪心的方式迭代地执行上述处理。

Although greedy selection has now become the de facto method for NMS, it still has some space for improvement, as shown in Fig 11. First of all, the top-scoring box may not be the best fit. Second, it may suppress nearby objects. Finally, it does not suppress false positives. In recent years, in spite of the fact that some manual modifications have been recently made to improve its performance [158, 159, 163] (see Section 4.4 for more details), to our best knowledge, the greedy selection still performs as the strongest baseline for today's object detection.

虽然贪心选择现在已经成为 NMS 的事实上的方法，但它仍然有一些改进的空间，如图 11 所示。首先，得分最高的盒子可能不是最合适的。其次，它可能会抑制附近的物体。最后，它不会抑制误报。近年来，尽管最近已经进行了一些手动修改以改善其性能[158,159,163]（更多细节见 4.4 节），但据我们所知，贪心选择仍然在今天物体检测的基线中表现最强。

- **BB aggregation**

BB aggregation is another group of techniques for NMS [10, 103, 156, 157] with the idea of combining or clustering multiple overlapped bounding boxes into one final detection. The advantage of this type of method is that it takes full consideration of object relationships and their spatial layout. There are some well-known detectors using this method, such as the VJ detector [10] and the Overfeat [103].

- **边界框聚合**

BB 聚合是 NMS [10,103,156,157]的另一组技术，其思想是将多个重叠的边界框组合或聚类成一个最终检测。这种方法的优点是它充分考虑了对象关系及其空间布局。有一些众所周知的检测器使用这种方法，如 VJ 检测器[10]和 Overfeat [103]。

- Learning to NMS

A recent group of NMS improvements that have recently received much attention is learning to NMS [136, 146, 154, 155]. The main idea of such group of methods is to think of NMS as a filter to re-score all raw detections and to train the NMS as part of a network in an end-to-end fashion. These methods have shown promising results on improving occlusion and dense object detection over traditional handcrafted NMS methods.

- NMS 学习

最近受到很多关注的最近一组 NMS 改进是 NMS 学习[136,146,154,155]。这类方法的主要思想是将 NMS 视为重新评分所有原始检测的过滤器，并以端到端的方式训练 NMS 作为网络的一部分。与传统的手工 NMS 方法相比，这些方法在改善遮挡和密集物体检测方面已经显示出有希望的结果。

2.3.6 Technical Evolution of Hard Negative Mining

The training of an object detector is essentially an imbalanced data learning problem. In the case of sliding window based detectors, the imbalance between backgrounds and objects could be as extreme as 104~105 background windows to every object. Modern detection datasets require the prediction of object aspect ratio, further increasing the imbalanced ratio to 106~107 [129]. In this case, using all background data will be harmful to training as the vast number of easy negatives will overwhelm the learning process. Hard negative mining (HNM) aims to deal with the problem of imbalanced data during training. The technical evolution of HNM in object detection is shown in Fig. 10.

2.3.6 难例挖掘的技术演进

目标检测器的训练本质上是不平衡的数据学习问题。在基于滑动窗口的检测器的情况下，背景和物体之间的不平衡可以与每个物体的 10^4 至 10^5 个背景窗口一样极端。现代检测数据集需要预测物体纵横比，进一步将不平衡比率增加到 10^6 至 10^7 [129]。在这种情况下，使用所有背景数据将对训练有害，因为大量的轻松否定将压倒学习过程。难例挖掘（HNM）旨在解决训练期间数据不平衡的问题。HNM 在目标检测中的技术演变如图 10 所示。

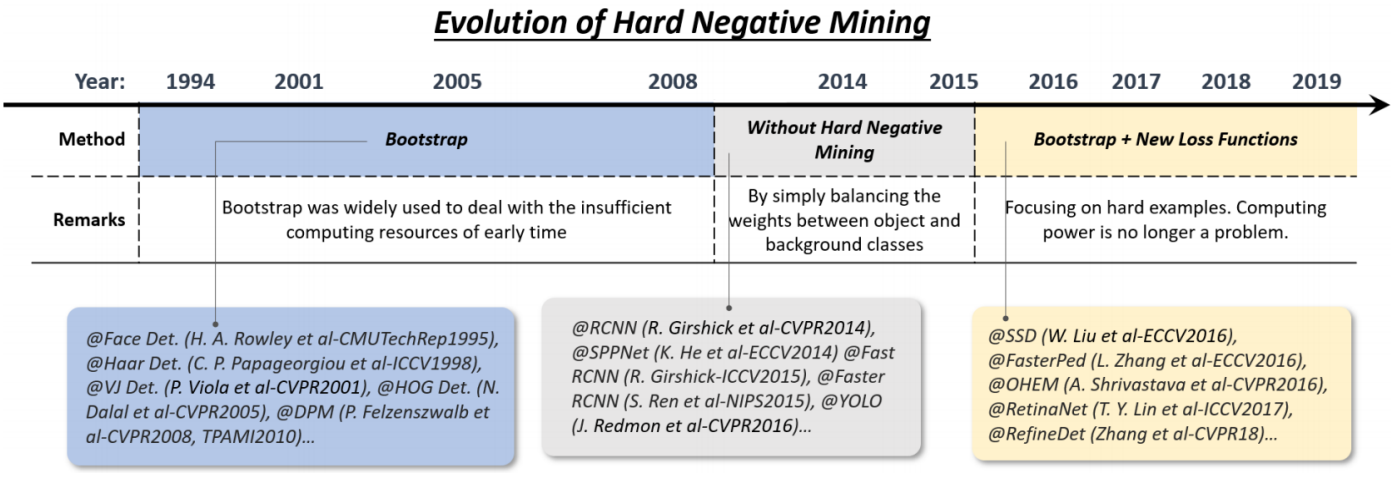


Figure 10. Evolution of hard negative mining techniques in object detection from 1994 to 2019. Detectors in this figure: Face Det. [164], Haar Det. [29], VJ Det. [10],

图 10. 1994 年至 2019 年目标检测中难例挖掘技术的演变。该图中的检测器：Face Det. [164], Haar Det. [29], VJ Det. [10], HOG Det. [12], DPM [13, 15], RCNN [16],

HOG Det. [12], DPM [13, 15], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], FasterPed [165], OHEM [166], RetinaNet [23], RefineDet [55].

SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], FasterPed [165], OHEM [166], RetinaNet [23], RefineDet [55]。

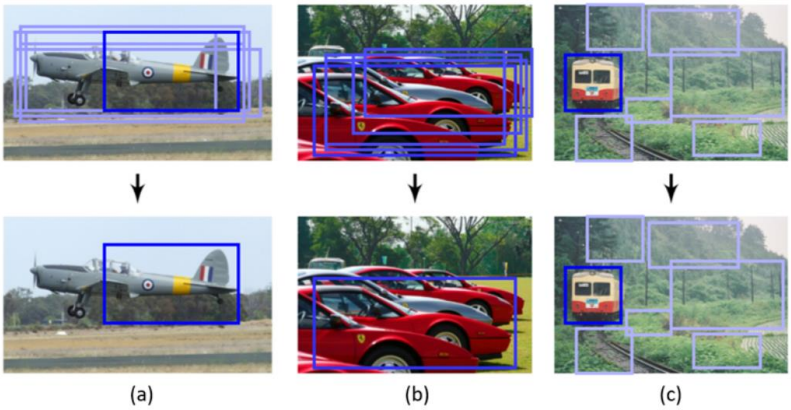


Figure 11. Examples of possible failures when using a standard greedy selection based non-max suppression: (a) the top-scoring box may not be the best fit, (b) it may suppress nearby objects, and (c) it does not suppress false positives. Images from R. Rothe et al. ACCV2014 [156].

图 11. 使用基于标准贪婪选择的非极大抑制时可能出现的失败的示例：（a）得分最高的盒子可能不是最合适的，（b）它可能抑制附近的物体，（c）它不会抑制误报。R. Rothe 等人的图片 ACCV2014 [156]。

- **Bootstrap**

Bootstrap in object detection refers to a group of training techniques in which the training starts with a small part of background samples and then iteratively add new miss-classified backgrounds during the training process. In early times object detectors, bootstrap was initially introduced with the purpose of reducing the training computations over millions of background samples [10, 29, 164]. Later it became a standard training technique in DPM and HOG detectors [12, 13] for solving the data imbalance problem.

- **Bootstrap（引导）**

目标检测中的引导是指一组训练技术，其中训练从一小部分背景样本开始，然后在训练过程中迭代地添加新的未分类背景。在早期的目标检测器中，最初引入了bootstrap，其目的是减少数百万背景样本的训练计算[10,29,164]。后来它成为 DPM 和 HOG 检测器的标准训练技术[12,13]，用于解决数据不平衡问题。

- **HNM in deep learning based detectors**

RCNN and YOLO simply balance the weights between the positive and negative windows. However, researchers later noticed that the weight-balancing cannot completely solve the imbalanced data problem [23]. To this end, after 2016, the bootstrap was re-introduced to deep learning based detectors [21, 165–168]. For example, in SSD [21] and OHEM [166], only the gradients of a very small part of samples (those with the largest loss values) will be back-propagated. In RefineDet [55], an “anchor refinement module” is designed to filter easy negatives. An alternative improvement is to design new loss functions [23, 169, 170], by reshaping the standard cross entropy loss so that it will put more focus on hard, misclassified examples [23].

- **基于深度学习的检测器中的难例挖掘**

RCNN 和 YOLO 简单地平衡正负窗口之间的权重。然而，研究人员后来注意到，权重平衡不能完全解决数据不平衡问题[23]。为此，在 2016 年之后，将引导程序重新引入基于深度学习的检测器[21,165-168]。例如，在 SSD[21]和 OHEM [166]中，只有非常小部分样本（具有最大损失值的那些）的梯度将被反向传播。在 RefineDet [55]中，“锚点细化模块”旨在过滤容易的负面。另一种改进是设计新的损失函数[23,169,170]，通过重塑标准的交叉熵损失，使其更加关注难的，错误分类的例子[23]。

3. SPEED-UP OF DETECTION

The acceleration of object detection has long been an important but challenging problem. In the past 20 years, the

3. 目标检测的提速

目标检测的提速一直是一个重要但具有挑战性的问题。在过去的 20 年中，目标检测社区已经开发出复杂的提

object detection community has developed sophisticated acceleration techniques. These techniques can be roughly divided into three levels of groups: “speed up of detection pipeline”, “speed up of detection engine”, and “speed up of numerical computation”, as shown in Fig 12.

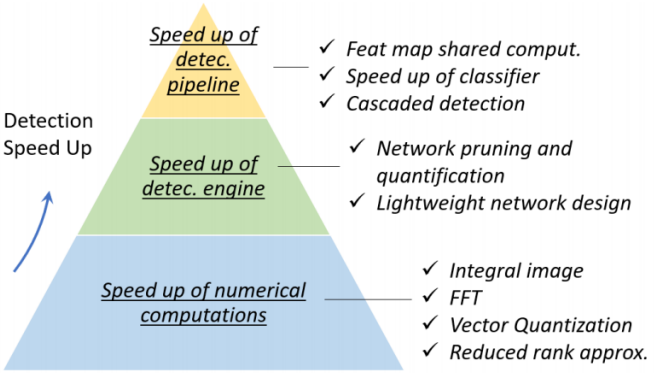


Figure 12. An overview of the speed up techniques in object detection.

速技术。这些技术可大致分为三个级别的组：“检测流程的提速”，“检测引擎的提速”和“数值计算的提速”，如图 12 所示。

图 12. 目标检测中提速技术的概述。

3.1 Feature Map Shared Computation

Among the different computational stages of an object detector, the feature extraction usually dominates the amount of computation. For a sliding window based detector, the computational redundancy starts from both positions and scales, where the former one is caused by the overlap between adjacent windows, while the later one is by the feature correlation between adjacent scales.

3.1 特征图共享计算

在目标检测器的不同计算阶段中，特征提取通常主导计算量。对于基于滑动窗口的检测器，计算冗余从位置和尺度开始，其中前者由相邻窗口之间的重叠引起，而后者是由相邻尺度之间的特征相关性引起。

3.1.1 Spatial Computational Redundancy and Speed Up

The most commonly used idea to reduce the spatial computational redundancy is feature map shared computation, i.e., to compute the feature map of the whole image only once before sliding window on it. The “image pyramid” of a traditional detector herein can be considered as a “feature pyramid”. For example, to speed up HOG pedestrian detector, researchers usually accumulate the “HOG map” of the whole input image, as shown in Fig. 13. However, the drawback of this method is also obvious, i.e., the feature map resolution (the minimum step size of the sliding window on this feature map) will be limited by the cell size. If a small object is located between two cells, it could be ignored by all detection windows. One solution to this problem is to build an integral feature pyramid, which will be introduced in Section 3.6.

3.1.1 空间计算冗余和加速

减少空间计算冗余的最常用的想法是特征图共享计算，即，在特征图上滑动窗口之前仅计算一次整个图像的特征图。这里传统检测器的“图像金字塔”可以被认为是“特征金字塔”。例如，为了加速 HOG 行人检测器，研究人员通常会累积整个输入图像的“HOG 图”，如图 13 所示。但是，这种方法的缺点也很明显，即特征图分辨率（此特征图上滑动窗口的最小步长）将受到单元格大小的限制。如果一个小对象位于两个单元之间，则所有检测窗口都可以忽略它。该问题的一个解决方案是构建一个整体特征金字塔，将在 3.6 节中介绍。

The idea of feature map shared computation has also been extensively used in convolutional based detectors.
Some related works can be traced back to the 1990s [96, 97]. Most of the CNN based detectors in recent years, e.g., SPPNet [17], Fast-RCNN [18], and Faster-RCNN [19], have applied similar ideas, which have achieved tens or even

特征映射共享计算的思想也已广泛用于基于卷积的检测器。一些相关的工作可以追溯到 20 世纪 90 年代[96,97]。近年来大多数基于 CNN 的检测器，例如 SPPNet [17]，Fast-RCNN [18]和 Faster-RCNN [19]，已经应用了类似的想法，这些想法已经实现了数十倍甚至数百倍的加速度。

hundreds of times of acceleration.

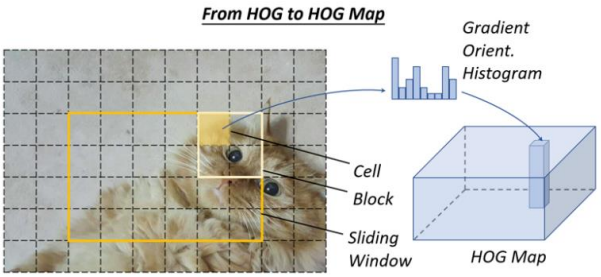


Figure 13. An illustration of how to compute the HOG map of an image.

图 13. 如何计算图像的 HOG 图的说明。

3.1.2 Scale Computational Redundancy and Speed Up

To reduce the scale computational redundancy, the most successful way is to directly scale the features rather than the images, which has been first applied in the VJ detector [10]. However, such an approach cannot be applied directly to HOG-like features because of blurring effects. For this problem, P. Dollar´ et al. discovered the strong (log-linear) correlation between the neighbor scales of the HOG and integral channel features [171] through extensive statistical analysis. This correlation can be used to accelerate the computation of a feature pyramid [172] by approximating the feature maps of adjacent scales. Besides, building "detector pyramid" is another way to avoid scale computational redundancy, i.e., to detect objects of different scales by simply sliding multiple detectors on one feature map rather than re-scaling the image or features [173].

3.1.2 尺度计算冗余和加速

为了减小尺度计算冗余，最成功的方法是直接缩放特征而不是图像，这些图像首先应用于 VJ 检测器[10]。然而，由于模糊效应，这种方法不能直接应用于类似 HOG 的特征。对于这个问题，P. Dollar´等通过广泛的统计分析，发现了 HOG 的邻域尺度与积分通道特征[171]之间的强（对数线性）相关性。该相关性可以用于通过近似相邻尺度的特征图来加速特征金字塔[172]的计算。此外，构建“检测器金字塔”是另一种避免规模计算冗余的方法，即通过简单地在一个特征图上滑动多个检测器而不是重新缩放图像或特征来检测不同尺度的物体[173]。

3.2 Speed up of Classifiers

Traditional sliding window based detectors, e.g., HOG detector and DPM, prefer using linear classifiers than nonlinear ones due to their low computational complexity. Detection with nonlinear classifiers such as kernel SVM suggests higher accuracy, but at the same time brings high computational overhead. As a standard non-parametric method, the traditional kernel method has no fixed computational complexity. When we have a very large training set, the detection speed will become extremely slow.

3.2 分类器加速

传统的基于滑动窗口的检测器，例如 HOG 检测器和 DPM，由于其低计算复杂性，更喜欢使用线性分类器而不是非线性分类器。使用诸如内核 SVM 的非线性分类器进行检测表明了更高的准确性，但同时带来了高计算开销。作为标准的非参数方法，传统的核方法没有固定的计算复杂度。当我们有一个非常大的训练集时，检测速度将变得非常慢。

In object detection, there are many ways to speed up kernelized classifiers, where the "model approximation" is most commonly used [30, 174]. Since the decision boundary of a classical kernel SVM can only be determined by a small set of its training samples (support vectors), the computational complexity at the inference stage would be proportional to the number of support vectors: $O(N_{sv})$.

Reduced Set Vectors [30] is an approximation method for kernel SVM, which aims to obtain an equivalent decision

在目标检测中，有许多方法可以加速核心化分类器，其中“模型近似”是最常用的[30,174]。由于经典核 SVM 的决策边界只能由其一小组训练样本（支持向量）确定，因此推理阶段的计算复杂度将与支持向量的数量成比例：$O(N_{sv})$。简化集合向量[30]是核 SVM 的近似方法，其目的在于根据少量合成向量获得等效判定边界。在对象检测中加速内核 SVM 的另一种方法是将其决策边界近似为分段线性形式，以便实现恒定的推理时间[174]。使用稀疏编码方法[175]也可以加速内核方法。

boundary in terms of a small number of synthetic vectors. Another way to speed up kernel SVM in object detection is to approximate its decision boundary to a piece-wise linear form so as to achieve a constant inference time [174].
The kernel method can also be accelerated with the sparse encoding methods [175].

3.3 Cascaded Detection

Cascaded detection is a commonly used technique in object detection [10, 176]. It takes a coarse to fine detection philosophy: to filter out most of the simple background windows using simple calculations, then to process those more difficult windows with complex ones. The VJ detector is a representative of cascaded detection. After that, many subsequent classical object detectors such as the HOG detector and DPM have been accelerated by using this technique [14, 38, 54, 177, 178].

In recent years, cascaded detection has also been applied to deep learning based detectors, especially for those detection tasks of "small objects in large scenes", e.g., face detection [179, 180], pedestrian detection [165, 177, 181], etc. In addition to the algorithm acceleration, cascaded detection has been applied to solve other problems, e.g., to improve the detection of hard examples [182–184], to integrate context information [143, 185], and to improve localization accuracy [104, 125].

3.4 Network Pruning and Quantification

"Network pruning" and "network quantification" are two commonly used techniques to speed up a CNN model, where the former one refers to pruning the network structure or weight to reduce its size and the latter one refers to reducing the code-length of activations or weights.

3.4.1 Network Pruning

The research of "network pruning" can be traced back to as early as the 1980s. At that time, Y. LeCun et al. proposed a method called "optimal brain damage" to compress the parameters of a multi-layer perceptron network [186]. In this method, the loss function of a network is approximated by taking the second-order derivatives so that to remove some unimportant weights. Following this idea, the network pruning methods in recent years usually take an iterative training and pruning process, i.e., to remove only a small group of unimportant weights after each stage of training, and to repeat those operations [187]. As traditional network pruning simply removes unimportant weights, which may

3.3 级联检测

级联检测是物体检测中常用的技术[10,176]。它采用粗略到精细的检测理念：使用简单的计算过滤掉大多数简单的背景窗口，然后用复杂的窗口处理那些更困难的窗口。VJ 检测器是级联检测的代表。之后，许多后续的经典物体检测器，如 HOG 检测器和 DPM，已经通过使用这种技术加速[14,18,54,177,178]。

近年来，级联检测也应用于基于深度学习的检测器，特别是对于“大场景中的小物体”的检测任务，例如人脸检测[179,180]，行人检测[165,177,181]， 除了算法加速之外，级联检测已被应用于解决其他问题，例如，改进难实例的检测[182-184]，整合上下文信息[143,185]，并提高定位精度[104,125]。

3.4 网络修剪和量化

“网络修剪”和“网络量化”是加速 CNN 模型的两种常用技术，前者指的是修剪网络结构或权重以减小其大小，后者指的是减少激活的代码长度或权重。

3.4.1 网络修剪

“网络修剪”的研究可以追溯到 20 世纪 80 年代。那时，Y.LeCun 等人提出了一种称为“最佳脑损伤”的方法来压缩多层感知器网络的参数[186]。在该方法中，通过采用二阶导数来近似网络的损失函数，以便去除一些不重要的权重。根据这一想法，近年来的网络修剪方法通常采用迭代训练和修剪过程，即在每个训练阶段后仅去除一小组不重要的权重，并重复这些操作[187]。由于传统的网络修剪只是去除了不重要的权重，这可能导致卷积滤波器中的一些稀疏连接模式，因此不能直接应用于压缩 CNN 模型。解决这个问题的一个简单方法是删除整个滤波器而不是独立的权重[188,189]。

result in some sparse connectivity patterns in a convolutional filter, it can not be directly applied to compress a CNN model. A simple solution to this problem is to remove the whole filters instead of the independent weights [188, 189].

3.4.2 Network Quantification

The recent works on network quantification mainly focus on network binarization, which aims to accelerate a network by quantifying its activations or weights to binary variables (say, 0/1) so that the floating-point operation is converted to AND, OR, NOT logical operations. Network binarization can significantly speed up computations and reduce the network's storage so that it can be much easier to be deployed on mobile devices. One possible implementation of the above ideas is to approximate the convolution by binary variables with the least squares method [190]. A more accurate approximation can be obtained by using linear combinations of multiple binary convolutions [191]. In addition, some researchers have further developed GPU acceleration libraries for binarized computation, which obtained more significant acceleration results [192].

3.4.2 网络量化

最近关于网络量化的工作主要集中在网络二值化，其目的是通过将其激活或权重量化为二进制变量（例如，0/1）来加速网络，以便将浮点运算转换为 AND，OR，NOT 逻辑操作。网络二值化可以显着加快计算速度并减少网络存储，从而可以更轻松地部署在移动设备上。上述思想的一种可能的实现方式是用最小二乘法[190]用二元变量逼近卷积。通过使用多个二进制卷积的线性组合可以获得更准确的近似[191]。此外，一些研究人员进一步开发了用于二值化计算的 GPU 加速库，从而获得了更为显着的加速结果[192]。

4. RECENT ADVANCES IN OBJECT DETECTION

4. 目标检测最新进展

In this section, we will review the state of the art object detection methods in recent three years.

在本节中，我们将回顾近三年来最先进的目标检测方法。

4.1 Detection with Better Engines

In recent years, deep CNN has played a central role in many computer vision tasks. As the accuracy of a detector depends heavily on its feature extraction networks, in this paper, we refer to the backbone networks, e.g. the ResNet and VGG, as the "engine" of a detector. Fig. 17 shows the detection accuracy of three well-known detection systems: Faster RCNN [19], R-FCN [46] and SSD [21] with different choices of the engines [27].

4.1 用更好的引擎进行检测

近年来，深度 CNN 在许多计算机视觉任务中发挥了核心作用。由于检测器的准确性在很大程度上取决于其特征提取网络，因此在本文中，我们将参考骨干网络，例如 ResNet 和 VGG，作为检测器的"引擎"。图 17 显示了三种众所周知的检测系统的检测精度：Faster RCNN [19]，R-FCN [46]和 SSD [21]，它们具有不同的引擎选择[27]。

In this section, we will introduce some of the important detection engines in deep learning era. We refer readers to the following survey for more details on this topic [229].

在本节中，我们将介绍深度学习时代的一些重要检测引擎。我们向读者推荐以下调查，以获取有关该主题的更多详细信息[229]。

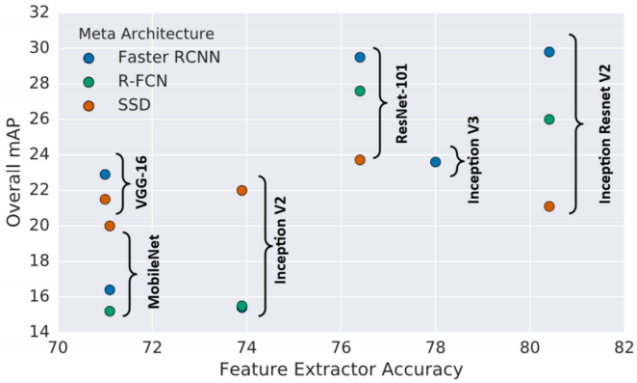


Figure 17. A comparison of detection accuracy of three detectors: Faster RCNN [19], R-FCN [46] and SSD [21] on MS-COCO dataset with different detection engines. Image from J. Huang et al. CVPR2017 [27].

图 17. 三种检测器的检测精度的比较：在具有不同检测引擎的 MS-COCO 数据集上 Faster RCNN [19]，R-FCN [46] 和 SSD [21]。来自 J. Huang 等人 CVPR2017 [27]的图片。

AlexNet: AlexNet [40], an eight-layer deep network, was the first CNN model that started the deep learning revolution in computer vision. AlexNet famously won the 2012 ImageNet LSVRC-2012 competition by a large margin [15.3% VS 26.2% (second place) error rates]. As of Feb. 2019, the Alexnet paper has been cited over 30,000 times.

AlexNet：AlexNet [40]是一个八层深度网络，是第一个启动计算机视觉深度学习革命的 CNN 模型。AlexNet 以极大的优势赢得 2012 年 ImageNet LSVRC-2012 竞赛[15.3％VS 26.2％（第二名）错误率]。截至 2019 年 2 月，Alexnet 论文已被引用超过 30,000 次。

VGG: VGG was proposed by Oxford's Visual Geometry Group (VGG) in 2014 [230]. VGG increased the model's depth to 16-19 layers and used very small (3x3) convolution filters instead of 5x5 and 7x7 those were previously used in AlexNet. VGG has achieved the state of the art performance on the ImageNet dataset of its time.

VGG：VGG 由牛津视觉几何组织（VGG）于 2014 年提出[230]。VGG 将模型的深度增加到 16-19 层，并使用非常小的（3x3）卷积滤波器，而不是之前在 AlexNet 中使用的 5x5 和 7x7。VGG 在其当时的 ImageNet 数据集上实现了最先进的性能。

GoogLeNet: GoogLeNet, a.k.a Inception [198, 231–233], is a big family of CNN models proposed by Google Inc. since 2014. GoogLeNet increased both of a CNN's width and depth (up to 22 layers). The main contribution of the Inception family is the introduction of factorizing convolution and batch normalization.

GoogLeNet：GoogLeNet，或称为 Inception [198,231-233]，是 Google Inc.自 2014 年以来提出的一大类 CNN 模型.GoogLeNet 增加了 CNN 的宽度和深度（最多 22 层）。Inception 系列的主要贡献是引入分解卷积和批量归一化。

ResNet: The Deep Residual Networks (ResNet) [234], proposed by K. He et al. in 2015, is a new type of convolutional network architecture that is substantially deeper (up to 152 layers) than those used previously. ResNet aims to ease the training of networks by reformulating its layers as learning residual functions with reference to the layer inputs. ResNet won multiple computer vision competitions in 2015, including ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

ResNet：深度残留网络（ResNet）[234]，由 K. He 等人提出。在 2015 年，是一种新型卷积网络架构，比以前使用的架构要深得多（最多 152 层）。ResNet 旨在通过参考层输入将其层重新构建为学习残差函数来简化网络训练。ResNet 在 2015 年赢得了多项计算机视觉竞赛，包括 ImageNet 检测，ImageNet 定位，COCO 检测和 COCO 分割。

DenseNet: DenseNet [235] was proposed by G. Huang and Z. Liu et al. in 2017. The success of ResNet suggested that

DenseNet：DenseNet [235]由 G. Huang 和 Z. Liu 等人提出。ResNet 的成功表明，CNN 的径直连接使我们能够

the short cut connection in CNN enables us to train deeper and more accurate models. The authors embraced this observation and introduced a densely connected block, which connects each layer to every other layer in a feedforward fashion.

训练更深入，更准确的模型。作者接受了这一发现并引入了一个密集连接的块，它以前馈方式将每一层连接到当前层。

SENet: Squeeze and Excitation Networks (SENet) was proposed by J. Hu and L. Shen et al. in 2018 [236]. Its main contribution is the integration of global pooling and shuffling to learn channel-wise importance of the feature map. SENet won the 1st place in ILSVRC 2017 classification competition.

SENet：挤压和激励网络（SENet）由 J. Hu 和 L. Shen 等人[236]在 2018 年提出。它的主要贡献是整合全局池化和改组，以学习特征通道的重要程度。SENet 在 2017 年 ILSVRC 分类比赛中获得第一名。

- **Object detectors with new engines**

In recent three years, many of the latest engines have been applied to object detection. For example, some latest object detection models such as STDN [237], DSOD [238], TinyDSOD [207], and Pelee [209] choose DenseNet [235] as their detection engine. The Mask RCNN [4], as the state of the art model for instance segmentation, applied the next generation of ResNet: ResNeXt [239] as its detection engine. Besides, to speed up detection, the depth-wise separable convolution operation, which was introduced by Xception [204], an improved version of Incepion, has also been used in detectors such as MobileNet [205] and LightHead RCNN [47].

- **采用新引擎的目标检测器**

近三年来，许多最新的引擎已应用于目标检测。例如，一些最新的目标检测模型，如 STDN[237]，DSOD[238]，TinyDSOD [207]和 Pelee [209]选择 DenseNet [235]作为它们的检测引擎。Mask RCNN [4]作为用于实例分割的最先进模型，应用下一代 ResNet：ResNeXt[239]作为其检测引擎。此外，为了加速检测，Xception [204]（Incepion 的改进版本）引入的深度可分离卷积操作也被用于诸如 MobileNet [205]和 LightHead RCNN [47]的检测器中。

4.2 Detection with Better Features

The quality of feature representations is critical for object detection. In recent years, many researchers have made efforts to further improve the quality of image features on basis of some latest engines, where the most important two groups of methods are: 1) feature fusion and 2) learning high-resolution features with large receptive fields.

4.2 用更好的特征进行检测

特征表示的质量对于目标检测是至关重要的。近年来，许多研究人员在一些最新的引擎的基础上努力进一步提高图像特征的质量，其中最重要的两组方法是：1）特征融合和 2）学习具有大感受野的高分辨率特征。

4.2.1 Why Feature Fusion is Important?

Invariance and equivariance are two important properties in image feature representations. Classification desires invariant feature representations since it aims at learning high-level semantic information. Object localization desires equivariant representations since it aims at discriminating position and scale changes. As object detection consists of two sub-tasks of object recognition and localization, it is crucial for a detector to learn both invariance and equivariance at the same time.

4.2.1 为什么特征融合如此重要？

不变性和等效性是图像特征表示中的两个重要属性。分类需要不变的特征表示，因为它旨在学习高级语义信息。目标定位需要等变的表示，因为它旨在区分位置和比例变化。由于目标检测由目标识别和定位两个子任务组成，因此检测器同时学习不变性和等效性至关重要。

Feature fusion has been widely used in object detection in the last three years. As a CNN model consists of a series of convolutional and pooling layers, features in deeper layers

特征融合在过去三年中已被广泛用于目标检测。由于 CNN 模型由一系列卷积和池化层组成，更深层中的特征将具有更强的不变性但更少的等价性。尽管这可能对类

will have stronger invariance but less equivariance. Although this could be beneficial to category recognition, it suffers from low localization accuracy in object detection. On the contrary, features in shallower layers is not conducive to learning semantics, but it helps object localization as it contains more information about edges and contours. Therefore, the integration of deep and shallow features in a CNN model helps improve both invariance and equivariance.

别识别有益，但它在目标检测中具有低定位精度。相反，较浅层中的特征不利于学习语义，但它有助于对象定位，因为它包含有关边和轮廓的更多信息。因此，CNN模型中深度和浅度特征的集成有助于改善不变性和等效性。

4.2.2 Feature Fusion in Different Ways?

There are many ways to perform feature fusion in object detection. Here we introduce some recent methods in two aspects: 1) processing flow and 2) element-wise operation.

4.2.2 特征融合的不同方式?

在目标检测中有许多方法可以执行特征融合。这里我们从两个方面介绍一些最近的方法：1）处理流程和 2）逐元素操作。

- **Processing flow**

Recent feature fusion methods in object detection can be divided into two categories: 1) bottom-up fusion, 2) topdown fusion, as shown in Fig. 18 (a)-(b). Bottom-up fusion feeds forward shallow features to deeper layers via skip connections [237, 240–242]. In comparison, top-down fusion feeds back the features of deeper layers into the shallower ones [22, 55, 243–246]. Apart from these methods, there are more complex approaches proposed recently, e.g., weaving features across different layers [247].

- **处理流程**

最近在目标检测中的特征融合方法可以分为两类：1）自下而上融合，2）自上而下融合，如图 18(a)-(b)所示。自下而上融合通过跳接将浅层特征向前推进到更深的层[237,240-242]。相比之下，自上而下的融合将较深层的特征反馈到较浅的层[22,55,243-246]。除了这些方法之外，最近提出了更复杂的方法，例如，跨不同层编织特征[247]。

As the feature maps of different layers may have different sizes both in terms of their spatial and channel dimensions, one may need to accommodate the feature maps, such as by adjusting the number of channels, up-sampling low resolution maps, or down-sampling high resolution maps to a proper size. The easiest ways to do this is to use nearestor bilinear-interpolation [22, 244]. Besides, fractional strided convolution (a.k.a. transpose convolution) [45, 248], is another recent popular way to resize the feature maps and adjust the number of channels. The advantage of using fractional strided convolution is that it can learn an appropriate way to perform up-sampling by itself [55, 212, 241– 243, 245, 246, 249].

由于不同层的特征图在空间和通道尺寸方面可能具有不同的尺寸，因此可能需要适应特征图，例如通过调整通道数，上采样低分辨率图或下采样高分辨率图到适当的大小。最简单的方法是使用最近或双线性插值[22,244]。此外，分数步长卷积（也称为转置卷积）[45,248]是另一种近期流行的方法，用于调整特征图的大小并调整通道数。使用分数跨步卷积的优点是它可以学习一种适当的方式来自己进行上采样[55,212,241-243,245,246,249]。

- **Element-wise operation**

From a local point of view, feature fusion can be considered as the element-wise operation between different feature maps. There are three groups of methods: 1) element-wise sum, 2) element-wise product, and 3) concatenation, as shown in Fig. 18 (c)-(e).

- 逐元素操作

从局部的角度来看，特征融合可以被认为是不同特征映射之间的逐元素操作。有三组方法：1）逐元素相加，2）逐元素相乘，3）拼接，如图 18(c)-(e)所示。

The element-wise sum is the easiest way to perform feature fusion. It has been frequently used in many recent object

逐元素相加是执行特征融合的最简单方法。它经常用于许多最近的物体检测器[22,55,241,243,246]。逐元素相乘

detectors [22, 55, 241, 243, 246]. The element-wise product [245, 249–251] is very similar to the element-wise sum, while the only difference is the use of multiplication instead of summation. An advantage of element-wise product is that it can be used to suppress or highlight the features within a certain area, which may further benefit small object detection [245, 250, 251]. Feature concatenation is another way of feature fusion [212, 237, 240, 244]. Its advantage is that it can be used to integrate context information of different regions [105, 144, 149, 161], while its disadvantage is the increase of the memory [235].

[245,249-251]与逐元素相加非常相似，而唯一的区别是使用乘法而不是求和。元素相乘的一个优点是它可以用于抑制或突出某个区域内的特征，这可以进一步有益于小物体检测[245,250,251]。特征拼接是特征融合的另一种方式[212,237,240,244]。它的优点是它可以用来整合不同区域的上下文信息[105,144,149,161]，而它的缺点是内存的增加[235]。

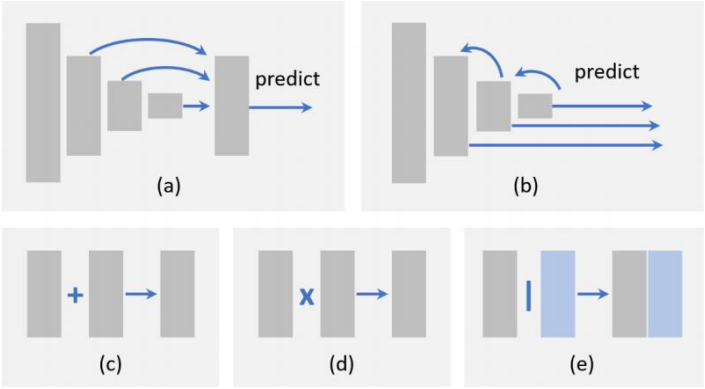


Figure 18. An illustration of different feature fusion methods: (a) bottom-up fusion, (b) top-down fusion, (c) element-wise sum, (d) element-wise product, and (e) concatenation.

图 18. 不同特征融合方法的说明：（a）自下而上融合，（b）自上而下融合，（c）逐元素相加，（d）逐元素相乘，和（e）拼接。

4.2.3 Learning High Resolution Features with Large Receptive Fields

4.2.3 采用大感受野学习高分辨率特征

The receptive field and feature resolution are two important characteristics of a CNN based detector, where the former one refers to the spatial range of input pixels that contribute to the calculation of a single pixel of the output, and the latter one corresponds to the down-sampling rate between the input and the feature map. A network with a larger receptive field is able to capture a larger scale of context information, while that with a smaller one may concentrate more on the local details.

感受野和特征分辨率是基于 CNN 的检测器的两个重要特征，其中前者指的是有助于计算输出的单个像素的输入像素的空间范围，而后者对应于输入和特征映射之间的下采样率。具有较大感受野的网络能够捕获更大规模的上下文信息，而具有较小感受野的网络可以更多地关注定位细节。

As we mentioned before, the lower the feature resolution is, the harder will be to detect small objects. The most straight forward way to increase the feature resolution is to remove pooling layer or to reduce the convolution down-sampling rate. But this will cause a new problem, the receptive field will become too small due to the decreasing of output stride. In other words, this will narrow a detector's "sight" and may result in the miss detection of some large objects.

正如我们之前提到的，特征分辨率越低，检测小对象就越困难。增加特征分辨率的最直接方法是删除池化层或降低卷积下采样率。但这会引起一个新问题，由于输出步幅的减少，感受野会变得太小。换句话说，这将缩小检测器的"视线"并可能导致一些大物体的未命中检测。

A practical method to increase both of the receptive field and feature resolution at the same time is to introduce dilated

同时增加感受野和特征分辨率的实际方法是引入扩张卷积（也被称为 atrous convolution，或 conolution with

convolution (a.k.a. atrous convolution, or convolution with holes). Dilated convolution is originally proposed in semantic segmentation tasks [252, 253]. Its main idea is to expand the convolution filter and use sparse parameters. For example, a 3x3 filter with a dilation rate of 2 will have the same receptive field as a 5x5 kernel but only have 9 parameters. Dilated convolution has now been widely used in object detection [21, 56, 254, 255], and proves to be effective for improved accuracy without any additional parameters and computational cost [56].

holes）。扩张卷积最初是在语义分割任务中提出的[252,253]。其主要思想是扩展卷积滤波器并使用稀疏参数。例如，扩张率为 2 的 3x3 滤波器将具有与 5x5 内核相同的感受野，但仅具有 9 个参数。扩张卷积现在已被广泛用于目标检测[21,56,254,255]，并证明在没有任何附加参数和计算成本的情况下有效提高准确度[56]。

4.3 Beyond Sliding Window

Although object detection has evolved from using handcrafted features to deep neural networks, the detection still follows a paradigm of "sliding window on feature maps" [137]. Recently, there are some detectors built beyond sliding windows.

4.3 超越滑动窗口

尽管目标检测已经从使用手工制作的特征演变为深度神经网络，但检测仍然遵循“特征图上的滑动窗口”的范例[137]。最近，有一些超越滑动窗户的检测器。

- **Detection as sub-region search**

Sub-region search [184, 256–258] provides a new way of performing detection. One recent method is to think of detection as a path planning process that starts from initial grids and finally converges to the desired ground truth boxes [256]. Another method is to think of detection as an iterative updating process to refine the corners of a predicted bounding box [257].

- **子区域搜索检测**

子区域搜索[184,256-258]提供了一种执行检测的新方法。最近的一种方法是将检测视为从初始网格开始并最终收敛到所需的真值框的路径规划过程[256]。另一种方法是将检测视为迭代更新过程以细化预测的边界框的角[257]。

- **Detection as key points localization**

Key points localization is an important computer vision task that has extensively broad applications, such as facial expression recognition [259], human poses identification [260], etc. As any object in an image can be uniquely determined by its upper left corner and lower right corner of the ground truth box, the detection task, therefore, can be equivalently framed as a pair-wise key points localization problem. One recent implementation of this idea is to predict a heat-map for the corners [261]. The advantage of this approach is that it can be implemented under a semantic segmentation framework, and there is no need to design multi-scale anchor boxes.

- **关键点定位检测**

关键点定位是一项重要的计算机视觉任务，具有广泛的应用，如面部表情识别[259]，人体姿势识别[260]等。因为图像中的任何对象可以由真值框的左上角和右下角唯一确定。因此，检测任务可以等效地构成一对成对的关键点定位问题。这个想法的最近实现是预测角的热图[261]。这种方法的优点是它可以在语义分割框架下实现，并且不需要设计多尺度锚框。

4.4 Improvements of Localization

To improve localization accuracy, there are two groups of methods in recent detectors: 1) bounding box refinement, and 2) designing new loss functions for accurate localization.

4.4 定位的提升

为了提高定位精度，最近的检测器中有两组方法：1）边界框细化，2）设计新的损失函数以实现精确定位。

4.4.1 Bounding Box Refinement

4.4.1 边界框细化

The most intuitive way to improve localization accuracy is bounding box refinement, which can be considered as a post-processing of the detection results. Although the bounding box regression has been integrated into most of the modern object detectors, there are still some objects with unexpected scales that cannot be well captured by any of the predefined anchors. This will inevitably lead to an inaccurate prediction of their locations. For this reason, the "iterative bounding box refinement" [262–264] has been introduced recently by iteratively feeding the detection results into a BB regressor until the prediction converges to a correct location and size. However, some researchers also claimed that this method does not guarantee the monotonicity of localization accuracy [262], in other words, the BB regression may degenerate the localization if it is applied for multiple times.

提高定位精度的最直观方法是边界框细化，可以将其视为检测结果的后处理。虽然边界框回归已经集成到大多数现代物体检测器中，但是仍然存在一些具有意外尺度的物体，这些物体不能被任何预定义的锚点很好地捕获。这将不可避免地导致对其位置的不准确预测。由于这个原因，最近通过迭代地将检测结果馈送到 BB 回归器中引入了"迭代边界框细化"[262-264]，直到预测收敛到正确的位置和大小。然而，一些研究人员还声称这种方法不能保证定位精度的单调性[262]，换句话说，如果多次应用，BB 回归可能会使定位退化。

4.4.2 Improving Loss Functions for Accurate Localization

In most modern detectors, object localization is considered as a coordinate regression problem. However, there are two drawbacks of this paradigm. First, the regression loss function does not correspond to the final evaluation of localization. For example, we can not guarantee that a lower regression error will always produce a higher IoU prediction, especially when the object has a very large aspect ratio. Second, the traditional bounding box regression method does not provide the confidence of localization. When there are multiple BB's overlapping with each other, this may lead to failure in non-maximum suppression (see more details in subsection 2.3.5).

4.4.2 改善损失函数用于精确定位

在大多数现代检测器中，目标定位被认为是坐标回归问题。但是，这种范例存在两个缺点。首先，回归损失函数与定位的最终评估不对应。例如，我们无法保证较低的回归误差始终会产生较高的 IoU 预测，尤其是当对象具有非常大的宽高比时。其次，传统的边界框回归方法不能提供定位的置信度。当多个 BB 彼此重叠时，这可能导致非最大抑制失败（请参阅第 2.3.5 小节中的更多细节）。

The above problems can be alleviated by designing new loss functions. The most intuitive design is to directly use IoU as the localization loss function [265]. Some other researchers have further proposed an IoU-guided NMS to improve localization in both training and detection stages [163]. Besides, some researchers have also tried to improve localization under a probabilistic inference framework [266]. Different from the previous methods that directly predict the box coordinates, this method predicts the probability distribution of a bounding box location.

通过设计新的损失函数可以减轻上述问题。最直观的设计是直接使用 IoU 作为本地化损失函数[265]。其他一些研究人员进一步提出了 IoU 引导的 NMS，以改善训练和检测阶段的定位[163]。此外，一些研究人员还尝试在概率推理框架下改进定位[266]。与以前直接预测框坐标的方法不同，此方法预测边界框位置的概率分布。

4.5 Learning with Segmentation

Object detection and semantic segmentation are all important tasks in computer vision. Recent researches suggest object detection can be improved by learning with semantic segmentation.

4.5 采用分割进行学习

目标检测和语义分割都是计算机视觉中的重要任务。最近的研究表明，通过语义分割学习可以改善目标检测。

4.5.1 Why Segmentation Improves Detection?

There are three reasons why the semantic segmentation improves object detection.

- **Segmentation helps category recognition**

Edges and boundaries are the basic elements that constitute human visual cognition [267, 268]. In computer vision, the difference between an object (e.g., a car, a person) and a stuff (e.g., sky, water, grass) is that the former usually has a closed and well defined boundary while the latter does not. As the feature of semantic segmentation tasks well captures the boundary of an object, segmentation may be helpful for category recognition.

- **Segmentation helps accurate localization**

The ground-truth bounding box of an object is determined by its well-defined boundary. For some objects with a special shape (e.g., imagine a cat with a very long tail), it will be difficult to predict high IoU locations. As object boundaries can be well encoded in semantic segmentation features, learning with segmentation would be helpful for accurate object localization.

- **Segmentation can be embedded as context**

Objects in daily life are surrounded by different backgrounds, such as the sky, water, grass, etc, and all these elements constitute the context of an object. Integrating the context of semantic segmentation will be helpful for object detection, say, an aircraft is more likely to appear in the sky than on the water.

4.5.2 How Segmentation Improves Detection?

There are two main approaches to improve object detection by segmentation: 1) learning with enriched features and 2) learning with multi-task loss functions.

- Learning with enriched features

The simplest way is to think of the segmentation network as a fixed feature extractor and to integrate it into a detection framework as additional features [144, 269, 270]. The advantage of this approach is that it is easy to implement, while the disadvantage is that the segmentation network may bring additional calculation.

- Learning with multi-task loss functions

Another way is to introduce an additional segmentation branch on top of the original detection framework and to train this model with multi-task loss functions (segmentation loss + detection loss) [4, 269]. In most cases, the

4.5.1 为什么分割能提升检测

语义分割改进对象检测有三个原因。

- **分割有助于类别识别**

边缘和边界是构成人类视觉认知的基本要素[267,268]。在计算机视觉中，物体（例如，汽车，人）和物体（例如，天空，水，草）之间的差异在于前者通常具有封闭且界限分明的边界，而后者不具有封闭且界限分明的边界。由于语义分割任务的特征很好地捕获了目标的边界，因此分割可能有助于类别识别。

- **分割有助于精确定位**

目标的真值边界框由其明确定义的边界决定。对于具有特殊形状的一些物体（例如，想象具有非常长尾巴的猫），将难以预测高 IoU 位置。由于目标边界可以在语义分割特征中很好地编码，因此使用分割进行学习将有助于准确的目标定位。

- **分割有助于整合上下文**

日常生活中的物体被不同的背景所包围，例如天空，水，草等，所有这些元素构成了物体的背景。整合语义分割的上下文将有助于物体检测，例如，飞机更可能出现在天空中而不是在水面上。

4.5.2 分割如何提升检测

通过分割来改进目标检测有两种主要方法：1）具有丰富特征的学习和 2）具有多任务损失函数的学习。

- 具有丰富特征的学习

最简单的方法是将分割网络视为固定特征提取器，并将其作为附加功能集成到检测框架中[144,269,270]。这种方法的优点是易于实现，而缺点是分割网络可能带来额外的计算。

- 具有多任务损失函数的学习

另一种方法是在原始检测框架之上引入一个额外的分割分支，并用多任务损失函数（分割损失+检测损失）训练该模型[4,269]。在大多数情况下，将在推理阶段删除分割分支。优点是检测速度不会受到影响，但缺点是训

segmentation brunch will be removed at the inference stage. The advantage is the detection speed will not be affected, but the disadvantage is that the training requires pixel-level image annotations. To this end, some researchers have followed the idea of "weakly supervised learning": instead of training based on pixel-wise annotation masks, they simply train the segmentation brunch based on the bounding-box level annotations [250, 271].

练需要像素级图像标注。为此，一些研究人员遵循"弱监督学习"的思想：他们不是基于像素方式的标注掩模进行训练，而是根据边界框级别标注训练分割分支[250,271]。

4.6 Robust Detection of Rotation and Scale Changes

4.6 旋转尺度变化的鲁棒检测

Object rotation and scale changes are important challenges in object detection. As the features learned by CNN are not invariant to rotation and large degree of scale changes, in recent years, many people have made efforts in this problem.

目标旋转和尺度变化是目标检测中的重要挑战。由于CNN学到的特征对于旋转并不是不变的，并且大规模的尺度变化，近年来，许多人已经在这个问题上做出了努力。

4.6.1 Rotation Robust Detection

4.6.1 旋转鲁棒检测

Object rotation is very common in detection tasks such as face detection, text detection, etc. The most straight forward solution to this problem is data augmentation so that an object in any orientation can be well covered by the augmented data [88]. Another solution is to train independent detectors for every orientation [272, 273]. Apart from these traditional approaches, recently, there are some new improvement methods.

目标旋转在检测任务中很常见，例如人脸检测，文本检测等。这个问题最直接的解决方案是数据增强，这样任何方向的物体都可以被增强数据很好地覆盖[88]。另一个解决方案是为每个方向训练独立的检测器[272,273]。除了这些传统方法，最近还有一些新的改进方法。

- **Rotation invariant loss functions**

- **旋转不变损失函数**

The idea of learning with rotation invariant loss function can be traced back to the 1990s [274]. Some recent works have introduced a constraint on the original detection loss function so that to make the features of rotated objects unchanged [275, 276].

学习旋转不变损失函数的想法可以追溯到20世纪90年代[274]。最近的一些作品引入了对原始检测损失函数的约束，以使旋转物体的特征不变[275,276]。

- **Rotation calibration**

- **旋转校准**

Another way of improving rotation invariant detection is to make geometric transformations of the objects candidates [277–279]. This will be especially helpful for multi-stage detectors, where the correlation at early stages will benefit the subsequent detections. The representative of this idea is Spatial Transformer Networks (STN) [278]. STN has now been used in rotated text detection [278] and rotated face detection [279].

改进旋转不变检测的另一种方法是对候选对象进行几何变换[277-279]。这对于多级检测器尤其有用，其中早期阶段的相关性将有益于后续检测。该想法的代表是空间变换器网络（STN）[278]。STN现在已用于旋转文本检测[278]和旋转面部检测[279]。

- **Rotation RoI Pooling**

- **旋转 RoI 池化**

In a two-stage detector, feature pooling aims to extract a fixed length feature representation for an object proposal with any location and size by first dividing the proposal evenly into a set of grids, and then concatenating the grid

在两阶段检测器中，特征池化旨在通过首先将候选区域均匀地划分为一组网格，然后连接网格特征来提取具有任何位置和大小的目标候选区域的固定长度特征表示。由于网格化是在笛卡尔坐标系中执行的，因此这些特征

features. As the grid meshing is performed in Cartesian coordinates, the features are not invariance to rotation transform. A recent improvement is to mesh the grids in polar coordinates so that the features could be robust to the rotation changes [272].

不是旋转变换的不变性。最近的一项改进是在极坐标中对网格进行网格划分，以使特征对旋转变化具有鲁棒性[272]。

4.6.2 Scale Robust Detection

Recent improvements have been made at both training and detection stages for scale robust detection.

4.6.2 尺度不变检测

最近在训练和检测阶段都进行了改进，以进行尺度稳健检测。

- Scale adaptive training

Most of the modern detectors re-scale the input image to a fixed size and back propagate the loss of the objects in all scales, as shown in Fig. 19 (a). However, a drawback of doing this is there will be a "scale imbalance" problem. Building an image pyramid during detection could alleviate this problem but not fundamentally [46, 234]. A recent improvement is Scale Normalization for Image Pyramids (SNIP) [280], which builds image pyramids at both of training and detection stages and only backpropagates the loss of some selected scales, as shown in Fig. 19 (b). Some researchers have further proposed a more efficient training strategy: SNIP with Efficient Resampling (SNIPER) [281], i.e. to crop and re-scale an image to a set of sub-regions so that to benefit from large batch training.

- 尺度自适应训练

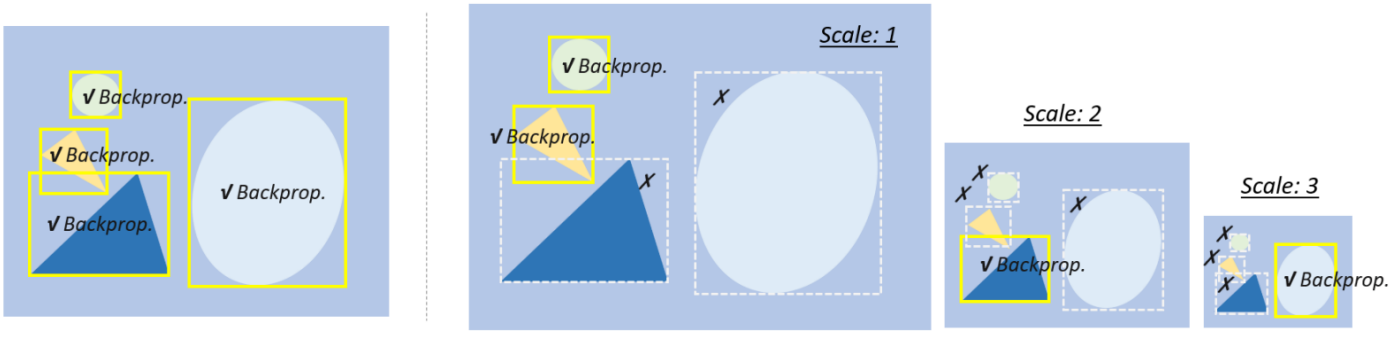
大多数现代检测器将输入图像重新缩放到固定尺寸，然后向后传播所有尺度的目标损失，如图 19(a)所示。然而，这样做的缺点是会出现"规模不平衡"问题。在检测过程中构建图像金字塔可以缓解这个问题，但不能从根本上缓解[46,234]。最近的改进是图像金字塔的尺度标准化（SNIP）[280]，它在训练和检测阶段建立图像金字塔，并且只反向传播一些选定尺度的损失，如图 19（b）所示。一些研究人员进一步提出了一种更有效的培训策略：SNIP 与高效重采样（SNIPER）[281]，即裁剪并将图像重新缩放到一组子区域，以便从大批量训练中受益。

- Scale adaptive detection

Most of the modern detectors use the fixed configurations for detecting objects of different sizes. For example, in a typical CNN based detector, we need to carefully define the size of anchors. A drawback of doing this is the configurations cannot be adaptive to unexpected scale changes. To improve the detection of small objects, some "adaptive zoom-in" techniques are proposed in some recent detectors to adaptively enlarge the small objects into the "larger ones" [184, 258]. Another recent improvement is learning to predict the scale distribution of objects in an image, and then adaptively re-scaling the image according to the distribution [282, 283].

- 尺度自适应检测

大多数现代检测器使用固定配置来检测不同尺寸的物体。例如，在典型的基于 CNN 的检测器中，我们需要仔细定义锚的大小。这样做的缺点是配置不能适应意外的比例变化。为了改善小物体的检测，在一些最近的检测器中提出了一些"自适应放大"技术，以自适应地将小物体放大为"较大物体"[184,258]。最近的另一项改进是学习预测图像中物体的比例分布，然后根据分布自适应地重新缩放图像[282,283]。



(a) Single resolution image, backprop all objects

(b) Multi-resolution images, backprop objects of selected scale

Figure 19. Different training strategies for multi-scale object detection: (a): Training on a single resolution image, back propagate objects of all scales [17–19, 21]. (b) Training on multi-resolution images (image pyramid), back propagate objects of selected scale. If an object is too large or too small, its gradient will be discarded [56, 280, 281].

图 19. 用于多尺度目标检测的不同训练策略：（a）：对单个分辨率图像进行训练，反向传播所有尺度的物体[17-19,21]。（b）对多分辨率图像（图像金字塔）进行训练，反向传播选定比例的物体。如果一个物体太大或太小，它的梯度将被丢弃[56,280,281]。

4.7 Training from Scratch

Most deep learning based detectors are first pre-trained on large scale datasets, say ImageNet, and then fine-tuned on specific detection tasks. People have always believed that pre-training helps to improve generalization ability and training speed and the question is, do we really need to pre-training a detector on ImageNet? In fact, there are some limitations when adopting the pre-trained networks in object detection. The first limitation is the divergence between ImageNet classification and object detection, including their loss functions and scale/category distributions. The second limitation is the domain mismatch. As images in ImageNet are RGB images while detection sometimes will be applied to depth image (RGB-D) or 3D medical images, the pretrained knowledge can not be well transfer to these detection tasks.

4.7 从头开始训练

大多数基于深度学习的检测器首先在大型数据集上进行预训练，比如 ImageNet，然后对特定的检测任务进行微调。人们一直认为预训练有助于提高泛化能力和训练速度，问题是，我们真的需要在 ImageNet 上预先训练检测器吗？实际上，在目标检测中采用预先训练的网络时存在一些限制。第一个限制是 ImageNet 分类和对象检测之间的差异，包括它们的损失函数和比例/类别分布。第二个限制是域不匹配。由于 ImageNet 中的图像是 RGB 图像，而检测有时会应用于深度图像（RGB-D）或 3D 医学图像，因此预训练的知识无法很好地转移到这些检测任务。

In recent years, some researchers have tried to train an object detector from scratch. To speed up training and improve stability, some researchers introduce dense connection and batch normalization to accelerate the back-propagation in shallow layers [238, 284]. The recent work by K. He et al. [285] has further questioned the paradigm of pretraining even further by exploring the opposite regime: they reported competitive results on object detection on the COCO dataset using standard models trained from random initialization, with the sole exception of increasing the number of training iterations so the randomly initialized models may converge. Training from random initialization is also surprisingly robust even using only 10% of the training data, which indicates that ImageNet pre-training may speed up convergence, but does not necessarily provide regularization or improve final detection accuracy.

近年来，一些研究人员试图从头开始训练目标检测器。为了加速训练并提高稳定性，一些研究人员引入了密集连接和批量归一化来加速浅层的反向传播[238,284]。K. He 等人最近的工作[285]通过探索相反的制度进一步质疑了预训练的范式：他们报告了使用随机初始化训练的标准模型对 COCO 数据集进行物体检测的竞争结果，唯一的例外是增加训练迭代次数，以便随机初始化的模型可能会收敛。即使仅使用 10％的训练数据，随机初始化的训练也令人惊讶地强大，这表明 ImageNet 预训练可以加速收敛，但不一定提供正则化或提高最终检测准确度。

4.8 Adversarial Training

The Generative Adversarial Networks (GAN) [286], introduced by A. Goodfellow et al. in 2014, has received great attention in recent years. A typical GAN consists of two neural networks: a generator networks and a discriminator networks, contesting with each other in a

4.8 对抗训练

生成对抗网络（GAN）[286]，由 A. Goodfellow 等人于 2014 年提出，近年来备受关注。典型的 GAN 由两个神经网络组成：发生器网络和鉴别器网络，在极小极大优化框架中相互竞争。通常，生成器学习从潜在空间映射到感兴趣的特定数据分布，而鉴别器旨在区分真实数据

minimax optimization framework. Typically, the generator learns to map from a latent space to a particular data distribution of interest, while the discriminator aims to discriminate between instances from the true data distribution and those produced by the generator. GAN has been widely used for many computer vision tasks such as image generation[286, 287], image style transfer [288], and image super-resolution [289]. In recent two years, GAN has also been applied to object detection, especially for improving the detection of small and occluded object.

分布的实例和生成器产生的实例。GAN 已广泛用于许多计算机视觉任务，如图像生成[286,287]，图像样式转移[288]和图像超分辨率[289]。近两年来，GAN 也被应用于目标检测，特别是用于改善小物体和被遮挡物体的检测。

GAN has been used to enhance the detection on small objects by narrowing the representations between small and large ones [290, 291]. To improve the detection of occluded objects, one recent idea is to generate occlusion masks by using adversarial training [292]. Instead of generating examples in pixel space, the adversarial network directly modifies the features to mimic occlusion.

GAN 已被用于通过缩小小型和大型物体之间的表示来增强对小物体的检测[290,291]。为了改进对被遮挡物体的检测，最近的一个想法是通过使用对抗训练来生成遮挡掩模[292]。对抗网络不是在像素空间中生成示例，而是直接修改特征以模仿遮挡。

In addition to these works, "adversarial attack" [293], which aims to study how to attack a detector with adversarial examples, has drawn increasing attention recently. The research on this topic is especially important for autonomous driving, as it cannot be fully trusted before guaranteeing the robustness to adversarial attacks.

除了这些作品之外，“对抗性攻击”[293]旨在研究如何用对抗性例子攻击检测器，最近引起了越来越多的关注。关于这一主题的研究对于自动驾驶尤为重要，因为在保证对抗性攻击的鲁棒性之前，它无法完全受到信任。

4.9 Weakly Supervised Object Detection

The training of a modern object detector usually requires a large amount of manually labeled data, while the labeling process is time-consuming, expensive, and inefficient. Weakly Supervised Object Detection (WSOD) aims to solve this problem by training a detector with only image level annotations instead of bounding boxes.

4.9 弱监督的目标检测

现代目标检测器的训练通常需要大量手动标注的数据，而标注过程耗时，昂贵且低效。弱监督目标检测（WSOD）旨在通过训练仅具有图像级标注而不是边界框的检测器来解决该问题。

Recently, multi-instance learning has been used for WSOD [294, 295]. Multi-instance learning is a group of supervised learning method [39, 296]. Instead of learning with a set of instances which are individually labeled, a multi-instance learning model receives a set of labeled bags, each containing many instances. If we consider object candidates in one image as a bag, and image-level annotation as the label, then the WSOD can be formulated as a multi-instance learning process.

最近，多实例学习已被用于 WSOD [294,295]。多实例学习是一组有监督的学习方法[39,296]。多实例学习模型不是使用一组单独标记的实例进行学习，而是接收一组标注的包，每个包含许多实例。如果我们将一个图像中的候选对象视为包，并将图像级标注视为标签，则可以将 WSOD 表示为多实例学习过程。

Class activation mapping is another recently group of methods for WSOD [297, 298]. The research on CNN visualization has shown that the convolution layer of a CNN behaves as object detectors despite there is no supervision on the location of the object. Class activation mapping shed

类别激活映射是另一组最近的 WSOD 方法[297,298]。对 CNN 可视化的研究表明，尽管没有对对象位置的监督，但 CNN 的卷积层表现为目标检测器。尽管在图像级标签上进行了训练，但类别激活映射揭示了如何使 CNN 具有定位能力[299]。

light on how to enable a CNN to have localization ability despite being trained on image level labels [299].

In addition to the above approaches, some other researchers considered the WSOD as a proposal ranking process by selecting the most informative regions and then training these regions with image-level annotation [300]. Another simple method for WSOD is to mask out different parts of the image. If the detection score drops sharply, then an object would be covered with high probability [301]. Besides, interactive annotation [295] takes human feedback into consideration during training so that to im prove WSOD. More recently, generative adversarial training has been used for WSOD [302].

除了上述方法之外，其他一些研究人员通过选择信息量最大的区域，然后用图像级标注训练这些区域，将 WSOD 视为候选区域排名过程[300]。WSOD 的另一个简单方法是屏蔽图像的不同部分。如果检测得分急剧下降，那么物体将被高概率地覆盖[301]。此外，交互式标注[295]在训练期间考虑人工反馈，以便证明 WSOD。最近，生成对抗训练已被用于 WSOD [302]。

5. CONCLUSION AND FUTURE DIRECTIONS

5. 结论和展望

Remarkable achievements have been made in object detection over the past 20 years. This paper not only extensively reviews some milestone detectors (e.g. VJ detector, HOG detector, DPM, Faster-RCNN, YOLO, SSD, etc), key technologies, speed up methods, detection applications, datasets, and metrics in its 20 years of history, but also discusses the challenges currently met by the community, and how these detectors can be further extended and improved.

在过去的 20 年中，在目标检测方面取得了显着的成就。本文不仅对其 20 年历史中的一些里程碑检测器（如 VJ 检测器，HOG 检测器，DPM，Faster-RCNN，YOLO，SSD 等），关键技术，加速方法，检测应用，数据集和指标进行了广泛的评论。还讨论了社区目前遇到的挑战，以及如何进一步扩展和改进这些检测器。

The future research of object detection may focus but is not limited to the following aspects:

目标检测的未来研究可能集中但不限于以下几个方面：

Lightweight object detection: To speed up the detection algorithm so that it can run smoothly on mobile devices. Some important applications include mobile augmented reality, smart cameras, face verification, etc. Although a great effort has been made in recent years, the speed gap between a machine and human eyes still remains large, especially for detecting some small objects.

轻量级目标检测：加速检测算法，使其可以在移动设备上平稳运行。一些重要的应用包括移动增强现实，智能相机，面部验证等。虽然近年来已经做出了巨大努力，但机器和人眼之间的速度差距仍然很大，特别是对于检测一些小物体。

Detection meets AutoML: Recent deep learning based detectors are becoming more and more sophisticated and heavily relies on experiences. A future direction is to reduce human intervention when designing the detection model (e.g., how to design the engine and how to set anchor boxes) by using neural architecture search. AutoML could be the future of object detection.

检测符合 AutoML：最近基于深度学习的检测器变得越来越复杂，并且在很大程度上依赖于经验。未来的方向是在设计检测模型（例如，如何设计引擎以及如何设置锚箱）时通过使用神经架构搜索来减少人为干预。AutoML 可能是对象检测的未来。

Detection meets domain adaptation: The training process

检测满足域自适应：在独立且相同分布（i.i.d.）数据的

of any target detector can be essentially considered as a likelihood estimation process under the assumption of independent and identically distributed (i.i.d.) data. Object detection with non-i.i.d. data, especially for some real-world applications, still remains a challenge. GAN has shown promising results in domain adaptation and may be of great help to object detection in the future.

假设下，任何目标检测器的训练过程基本上可以被视为似然估计过程。使用非 i.i.d 进行对象检测。数据，特别是对于某些实际应用，仍然是一个挑战。GAN 已经在域适应中显示出有希望的结果，并且可能对将来的对象检测有很大帮助。

Weakly supervised detection: The training of a deep learning based detector usually relies on a large amount of well-annotated images. The annotation process is timeconsuming, expensive, and inefficient. Developing weakly supervised detection techniques where the detectors are only trained with image-level annotations, or partially with bounding box annotations is of great importance for reducing labor costs and improving detection flexibility.

弱监督检测：基于深度学习的检测器的训练通常依赖于大量标注良好的图像。标注过程耗时，昂贵且效率低下。开发弱监督检测技术，其中检测器仅使用图像级注释进行训练，或部分使用边界框检测

Small object detection: Detecting small objects in large scenes has long been a challenge. Some potential application of this research direction includes counting the population of wild animals with remote sensing images and detecting the state of some important military targets. Some further directions may include the integration of the visual attention mechanisms and the design of high resolution lightweight networks.

小物体检测：在大型场景中检测小物体一直是一个挑战。该研究方向的一些潜在应用包括用遥感图像计算野生动物种群并检测一些重要军事目标的状态。一些进一步的方向可能包括视觉注意机制的集成和高分辨率轻量级网络的设计。

Detection in videos: Real-time object detection/tracking in HD videos is of great importance for video surveillance and autonomous driving. Traditional object detectors are usually designed under for image-wise detection, while simply ignores the correlations between videos frames. Improving detection by exploring the spatial and temporal correlation is an important research direction.

视频中的检测：高清视频中的实时物体检测/跟踪对于视频监控和自动驾驶非常重要。传统的物体检测器通常设计用于图像方式检测，而忽略视频帧之间的相关性。通过探索空间和时间相关性来改进检测是一个重要的研究方向。

Detection with information fusion: Object detection with multiple sources/modalities of data, e.g., RGB-D image, 3d point cloud, LIDAR, etc, is of great importance for autonomous driving and drone applications. Some open questions include: how to immigrate well-trained detectors to different modalities of data, how to make information fusion to improve detection, etc.

利用信息融合进行检测：具有多个数据源/数据模式的目标检测，例如 RGB-D 图像，3d 点云，激光雷达等，对于自动驾驶和无人机应用非常重要。一些开放性问题包括：如何将训练有素的检测器移植到不同的数据模式，如何进行信息融合以改善检测等。

Standing on the highway of technical evolutions, we believe this paper will help readers to build a big picture of object detection and to find future directions of this fastmoving research field.

站在技术演变的高速公路上，我们相信本文将帮助读者建立一个物体检测的大图，并找到这个快速研究领域的未来方向。

6. References

[1] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in European Conference on Computer Vision. Springer, 2014, pp. 297–312.
[2] ——, "Hypercolumns for object segmentation and finegrained localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 447–456.
[3] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3150–3158.
[4] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask rcnn," in Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 2980–2988.
[5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.
[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in International conference on machine learning, 2015, pp. 2048–2057.
[7] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 6, pp. 1367–1381, 2018.
[8] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang et al., "T-cnn: Tubelets with convolutional neural networks for object detection from videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896–2907, 2018.
[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.
[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1. IEEE, 2001, pp. I–I.
[11] P. Viola and M. J. Jones, "Robust real-time face detection," International journal of computer vision, vol. 57, no. 2, pp. 137–154, 2004.
[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 886–893.
[13] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.
[14] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in Computer vision and pattern recognition (CVPR), 2010 IEEE conference on. IEEE, 2010, pp. 2241–2248.
[15] 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.
[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in European conference on computer vision. Springer, 2014, pp. 346–361.
[18] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.
[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016,pp. 21–37.
[22] T.-Y. Lin, P. Dollar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in CVPR, vol. 1, no. 2, 2017, p. 4.
[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," IEEE transactions

on pattern analysis and machine intelligence, 2018.
[24] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikainen, "Deep learning for generic object de- ¨ tection: A survey," arXiv preprint arXiv:1809.02165, 2018.
[25] S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," arXiv preprint arXiv:1809.03193, 2018.
[26] A. Andreopoulos and J. K. Tsotsos, "50 years of object recognition: Directions forward," Computer vision and image understanding, vol. 117, no. 8, pp. 827–891, 2013.
[27] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in IEEE CVPR, vol. 4, 2017.
[28] K. Grauman and B. Leibe, "Visual object recognition (synthesis lectures on artificial intelligence and machine learning)," Morgan & Claypool, 2011.
[29] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in Computer vision, 1998. sixth international conference on. IEEE, 1998, pp. 555–562.
[30] C. Papageorgiou and T. Poggio, "A trainable system for object detection," International journal of computer vision, vol. 38, no. 1, pp. 15–33, 2000.
[31] A. Mohan, C. Papageorgiou, and T. Poggio, "Examplebased object detection in images by components," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 4, pp. 349–361, 2001.
[32] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," Journal-Japanese Society For Artificial Intelligence, vol. 14, no. 771-780, p. 1612, 1999.
[33] D. G. Lowe, "Object recognition from local scale-invariant features," in Computer vision, 1999. The proceedings of the seventh IEEE international conference on, vol. 2. Ieee, 1999, pp. 1150–1157.
[34] ——, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.
[35] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," CALIFORNIA UNIV SAN DIEGO LA JOLLA DEPT OF COMPUTER SCIENCE AND ENGINEERING, Tech. Rep., 2002.
[36] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 89–96.
[37] R. B. Girshick, P. F. Felzenszwalb, and D. A. Mcallester, "Object detection with grammar models," in Advances in Neural Information Processing Systems, 2011, pp. 442–450.
[38] R. B. Girshick, From rigid templates to grammars: Object detection with structured models. Citeseer, 2012.
[39] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in Advances in neural information processing systems, 2003, pp.577–584.
[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012,pp. 1097–1105.
[41] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Regionbased convolutional networks for accurate object detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, pp. 142–158, 2016.
[42] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W.Smeulders, "Segmentation as selective search for object recognition," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 1879–1886.
[43] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," http://people.cs.uchicago.edu/ rbg/latentrelease5/.
[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 6, pp. 1137–1149, 2017.
[45] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision. Springer, 2014, pp. 818–833.
[46] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in Advances in neural information processing systems, 2016, pp. 379–387.
[47] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head r-cnn: In defense of two-stage object detector," arXiv preprint arXiv:1711.07264, 2017.
[48] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," arXiv preprint, 2017.

[49] ——, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
[50] M. Everingham, L. Van Gool, C. K. 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.
[51] M. Everingham, S. A. Eslami, L. Van Gool, C. K. 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.
[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.
[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: ´ Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.
[54] M. A. Sadeghi and D. Forsyth, "30hz object detection with dpm v5," in European Conference on Computer Vision. Springer, 2014, pp. 65–79.
[55] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Singleshot refinement neural network for object detection," in IEEE CVPR, 2018.
[56] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," arXiv preprint arXiv:1901.01892, 2019.
[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee, 2009, pp. 248–255.
[58] I. Krasin and T. e. a. Duerig, "Openimages: A public dataset for large-scale multi-label and multiclass image classification." Dataset available from https://storage.googleapis.com/openimages/web/index.html,2017.
[59] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 304–311.
[60] 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.
[61] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 3354–3361.
[62] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, no. 2, 2017, p. 3.
[63] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.
[64] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "The eurocity persons dataset: A novel benchmark for object detection," arXiv preprint arXiv:1805.07193, 2018.
[65] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, Tech. Rep., 2010.
[66] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. IEEE, 2011, pp. 2144–2151.
[67] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015,pp. 1931–1939.
[68] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5525–5533.
[69] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel, "Pushing the limits of unconstrained face detection: a challenge dataset and baseline results," arXiv preprint arXiv:1804.10275, 2018.
[70] M. K. Yucel, Y. C. Bilge, O. Oguz, N. Ikizler-Cinbis, P. Duygulu, and R. G. Cinbis, "Wildest faces: Face detection and recognition in violent settings," arXiv preprint arXiv:1805.07566, 2018.
[71] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "Icdar 2003 robust reading competitions," in null. IEEE, 2003, p. 682.

[72] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu et al., "Icdar 2015 competition on robust reading," in Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. IEEE, 2015, pp. 1156–1160.
[73] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "Icdar2017 competition on reading chinese text in the wild (rctw-17)," in Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on, vol. 1. IEEE, 2017, pp. 1429–1434.
[74] K. Wang and S. Belongie, "Word spotting in the wild," in European Conference on Computer Vision. Springer, 2010, pp. 591–604.
[75] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 1083–1090.
[76] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in BMVC British Machine Vision Conference. BMVA, 2012.
[77] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," arXiv preprint arXiv:1406.2227, 2014.
[78] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," arXiv preprint arXiv:1601.07140, 2016.
[79] R. De Charette and F. Nashashibi, "Real time visual traffic lights recognition based on spot light detection and adaptive traffic lights templates," in Intelligent Vehicles Symposium, 2009 IEEE. IEEE, 2009, pp. 358–363.
[80] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey." IEEE Trans. Intelligent Transportation Systems, vol. 13, no. 4, pp. 1484–1497, 2012.
[81] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The german traffic sign detection benchmark," in Neural Networks (IJCNN), The 2013 International Joint Conference on. IEEE, 2013, pp. 1–8.
[82] R. Timofte, K. Zimmermann, and L. Van Gool, "Multiview traffic sign detection, recognition, and 3d localisation," Machine vision and applications, vol. 25, no. 3, pp. 633–647, 2014.
[83] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2110–2118.
[84] K. Behrendt, L. Novak, and R. Botros, "A deep learning approach to traffic lights: Detection, tracking, and classification," in Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017, pp. 1370–1377.
[85] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in European conference on computer vision. Springer, 2008, pp. 30–43.
[86] F. Tanner, B. Colder, C. Pullen, D. Heagy, M. Eppolito, V. Carlan, C. Oertel, and P. Sallee, "Overhead imagery research data setan annotated data library & tools to aid in the development of computer vision algorithms," in 2009 IEEE Applied Imagery Pattern Recognition Workshop (AIPR 2009). IEEE, 2009, pp. 1–8.
[87] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images." IEEE Geosci. Remote Sensing Lett., vol. 12, no. 9, pp. 1938–1942, 2015.
[88] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015, pp. 3735–3739.
[89] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," Journal of Visual Communication and Image Representation, vol. 34, pp. 187–203, 2016.
[90] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 117, pp. 11–28, 2016.
[91] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1100–1111, 2018.
[92] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in Proc. CVPR, 2018.
[93] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xview: Objects in

context in overhead imagery," arXiv preprint arXiv:1802.07856, 2018.
[94] K. Oksuz, B. C. Cam, E. Akbas, and S. Kalkan, "Localization recall precision (lrp): A new performance metric for object detection," in European Conference on Computer Vision (ECCV), vol. 6, 2018.
[95] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of cognitive neuroscience, vol. 3, no. 1, pp. 71–86, 1991.
[96] R. Vaillant, C. Monrocq, and Y. Le Cun, "Original approach for the localisation of objects in images," IEE Proceedings-Vision, Image and Signal Processing, vol. 141, no. 4, pp. 245–250, 1994.
[97] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradientbased learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
[98] I. Biederman, "Recognition-by-components: a theory of human image understanding." Psychological review, vol. 94, no. 2, p. 115, 1987.
[99] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," IEEE Transactions on computers, vol. 100, no. 1, pp. 67–92, 1973.
[100] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," International journal of computer vision, vol. 77, no.1-3, pp. 259–289, 2008.
[101] D. M. Gavrila and V. Philomin, "Real-time object detection for" smart" vehicles," in Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 1. IEEE, 1999, pp. 87–93.
[102] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in null. IEEE, 2005, pp. 90–97.
[103] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229, 2013.
[104] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in Advances in neural information processing systems, 2013, pp. 2553–2561.
[105] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in European Conference on Computer Vision. Springer, 2016, pp. 354–370.
[106] A. Pentland, B. Moghaddam, T. Starner et al., "Viewbased and modular eigenspaces for face recognition," 1994.
[107] G. Yang and T. S. Huang, "Human face detection in a complex background," Pattern recognition, vol. 27, no. 1, pp. 53–63, 1994.
[108] I. Craw, D. Tock, and A. Bennett, "Finding face features," in European Conference on Computer Vision. Springer, 1992, pp. 92–96.
[109] R. Xiao, L. Zhu, and H.-J. Zhang, "Boosting chain learning for object detection," in Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003, pp. 709–715.
[110] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
[111] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," arXiv preprint arXiv:1412.7062, 2014.
[112] C. Garcia and M. Delakis, "A neural architecture for fast and robust face detection," in Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 2. IEEE, 2002, pp. 44–47.
[113] M. Osadchy, M. L. Miller, and Y. L. Cun, "Synergistic face detection and pose estimation with energy-based models," in Advances in Neural Information Processing Systems, 2005, pp. 1017–1024.
[114] S. J. Nowlan and J. C. Platt, "A convolutional neural network hand tracker," Advances in neural information processing systems, pp. 901–908, 1995.
[115] T. Malisiewicz, Exemplar-based representations for object detection, association and beyond. Carnegie Mellon University, 2011.
[116] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 73–80.
[117] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," International journal of computer vision, vol. 104, no. 2, pp. 154–171, 2013.
[118] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in Computer

Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3241–3248.
[119] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 328–335.
[120] 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.
[121] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 3286–3293.
[122] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in European conference on computer vision. Springer, 2014, pp. 391–405.
[123] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe, "Scalable, high-quality object detection," arXiv preprint arXiv:1412.1441, 2014.
[124] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2147–2154.
[125] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, "Deepproposal: Hunting objects by cascading deep convolutional layers," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2578–2586.
[126] W. Kuo, B. Hariharan, and J. Malik, "Deepbox: Learning objectness with convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2479–2487.
[127] S. Gidaris and N. Komodakis, "Attend refine repeat: Active box proposal generation via in-out localization," arXiv preprint arXiv:1606.04446, 2016.
[128] H. Li, Y. Liu, W. Ouyang, and X. Wang, "Zoom out-andin network with recursive training for object proposal," arXiv preprint arXiv:1702.05711, 2017.
[129] J. Hosang, R. Benenson, P. Dollar, and B. Schiele, "What makes for effective detection proposals?" IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 4, pp. 814–830, 2016.
[130] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" arXiv preprint arXiv:1406.6962, 2014.
[131] J. Carreira and C. Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric min-cuts," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 7, pp. 1312–1328, 2011.
[132] N. Chavali, H. Agrawal, A. Mahendru, and D. Batra, "Object-proposal evaluation protocol is' gameable'," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 835–844.
[133] K. Lenc and A. Vedaldi, "R-cnn minus r," arXiv preprint arXiv:1506.06981, 2015.
[134] P.-A. Savalle, S. Tsogkas, G. Papandreou, and I. Kokkinos, "Deformable part models with cnn features," in European Conference on Computer Vision, Parts and Attributes Workshop, 2014.
[135] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Partbased r-cnns for fine-grained category detection," in European conference on computer vision. Springer, 2014, pp. 834–849.
[136] L. Wan, D. Eigen, and R. Fergus, "End-to-end integration of a convolution network, deformable parts model and non-maximum suppression," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 851–859.
[137] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2015, pp. 437–446.
[138] B. Li, T. Wu, S. Shao, L. Zhang, and R. Chu, "Object detection via end-to-end integration of aspect ratio and context aware part-based models and fully convolutional networks," arXiv preprint arXiv:1612.00534, 2016.
[139] A. Torralba and P. Sinha, "Detecting faces in impoverished images," MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, Tech. Rep., 2001.
[140] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollar, "A multipath network for object detection," arXiv preprint arXiv:1604.02135, 2016.
[141] X. Zeng, W. Ouyang, B. Yang, J. Yan, and X. Wang, "Gated bi-directional cnn for object detection," in European Conference on Computer Vision. Springer, 2016, pp. 354–369.
[142] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang et al., "Crafting gbd-net for

object detection," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 9, pp. 2109–2123, 2018.
[143] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Learning chained deep features and classifiers for cascade in object detection," arXiv preprint arXiv:1702.07054, 2017.
[144] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1134–1142.
[145] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, H. Lu et al., "Couplenet: Coupling global structure with local parts for object detection," in Proc. of Intl Conf. on Computer Vision (ICCV), vol. 2, 2017.
[146] 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.
[147] Z. Li, Y. Chen, G. Yu, and Y. Deng, "R-fcn++: Towards accurate region-based fully convolutional networks for object detection." in AAAI, 2018.
[148] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2874–2883.
[149] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," IEEE Transactions on Multimedia, vol. 19, no. 5, pp. 944–954, 2017.
[150] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 1, pp. 13–27, 2015.
[151] S. Gupta, B. Hariharan, and J. Malik, "Exploring person context and local scene context for object detection," arXiv preprint arXiv:1511.08177, 2015.
[152] X. Chen and A. Gupta, "Spatial memory for context reasoning in object detection," arXiv preprint arXiv:1704.04224, 2017.
[153] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6985–6994.
[154] J. H. Hosang, R. Benenson, and B. Schiele, "Learning nonmaximum suppression." in CVPR, 2017, pp. 6469–6477.
[155] P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," in Asian Conference on Computer Vision. Springer, 2016, pp. 198–213.
[156] R. Rothe, M. Guillaumin, and L. Van Gool, "Nonmaximum suppression for object detection by passing messages between windows," in Asian Conference on Computer Vision. Springer, 2014, pp. 290–306.
[157] D. Mrowca, M. Rohrbach, J. Hoffman, R. Hu, K. Saenko, and T. Darrell, "Spatial semantic regularisation for large scale object detection," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2003–2011.
[158] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Softnmsimproving object detection with one line of code," in Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 5562–5570.
[159] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness nms and bounded iou loss," arXiv preprint arXiv:1711.00164, 2017.
[160] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 1271–1278.
[161] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-cnn for small object detection," in Asian conference on computer vision. Springer, 2016, pp. 214–230.
[162] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in Computer Vision and Pattern Recognition (CVPR), vol. 2, no. 3, 2018.
[163] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018, pp. 8–14.
[164] H. A. Rowley, S. Baluja, and T. Kanade, "Human face detection in visual scenes," in Advances in Neural Information Processing Systems, 1996, pp. 875–881.
[165] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster rcnn doing well for pedestrian detection?" in European Conference on Computer Vision. Springer, 2016, pp. 443–457.
[166] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining,"

in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 761–769.
[167] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," Sensors, vol. 17, no. 2, p. 336, 2017.
[168] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster rcnn approach," Neurocomputing, vol. 299, pp. 42–50, 2018.
[169] J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 5, pp. 1991–2000, 2014.
[170] M. Zhou, M. Jing, D. Liu, Z. Xia, Z. Zou, and Z. Shi, "Multi-resolution networks for ship detection in infrared remote sensing images," Infrared Physics & Technology, 2018.
[171] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009.
[172] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 8, pp. 1532–1545, 2014.
[173] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2903–2910.
[174] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.
[175] A. Vedaldi and A. Zisserman, "Sparse kernel approximations for efficient classification and detection," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2320–2327.
[176] F. Fleuret and D. Geman, "Coarse-to-fine face detection," International Journal of computer vision, vol. 41, no. 1-2, pp.85–107, 2001.
[177] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2. IEEE, 2006, pp. 1491–1498.
[178] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 606–613.
[179] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325–5334.
[180] 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.
[181] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3361–3369.
[182] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Craft objects from images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 6043–6051.
[183] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2129–2137.
[184] M. Gao, R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Dynamic zoom-in network for fast object detection in large images," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[185] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Chained cascade network for object detection," in Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 1956–1964.
[186] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in Advances in neural information processing systems, 1990, pp. 598–605.
[187] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," arXiv preprint arXiv:1510.00149, 2015.
[188] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," arXiv preprint arXiv:1608.08710, 2016.
[189] G. Huang, S. Liu, L. van der Maaten, and K. Q. Weinberger, "Condensenet: An efficient densenet using learned group convolutions," group, vol. 3, no. 12, p. 11, 2017.
[190] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional

neural networks," in European Conference on Computer Vision. Springer, 2016, pp. 525–542.
[191] X. Lin, C. Zhao, and W. Pan, "Towards accurate binary convolutional neural network," in Advances in Neural Information Processing Systems, 2017, pp. 345–353.
[192] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in Advances in neural information processing systems, 2016, pp. 4107–4115.
[193] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[194] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," arXiv preprint arXiv:1412.6550, 2014.
[195] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in Advances in Neural Information Processing Systems, 2017, pp. 742–751.
[196] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 7341–7349.
[197] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5353–5360.
[198] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
[199] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel mattersimprove semantic segmentation by global convolutional network," in Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017, pp. 1743–1751.
[200] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, "Pvanet: deep but lightweight neural networks for realtime object detection," arXiv preprint arXiv:1608.08021, 2016.
[201] X. Zhang, J. Zou, X. Ming, K. He, and J. Sun, "Efficient and accurate approximations of nonlinear convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1984–1992.
[202] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 1943–1955, 2016.
[203] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," 2017.
[204] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," arXiv preprint, pp. 1610–02 357, 2017.
[205] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861,2017.
[206] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018, pp. 4510–4520.
[207] Y. Li, J. Li, W. Lin, and J. Li, "Tiny-dsod: Lightweight object detection for resource-restricted usages," arXiv preprint arXiv:1807.11013, 2018.
[208] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," science, vol. 313, no. 5786, pp. 504–507, 2006.
[209] R. J. Wang, X. Li, S. Ao, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," arXiv preprint arXiv:1804.06882, 2018.
[210] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.
[211] B. Wu, F. N. Iandola, P. H. Jin, and K. Keutzer, "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving." in CVPR Workshops, 2017, pp. 446–454.
[212] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 845–853.
[213] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8697–8710.
[214] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," arXiv preprint arXiv:1611.01578, 2016.

[215] Y. Chen, T. Yang, X. Zhang, G. Meng, C. Pan, and J. Sun, "Detnas: Neural architecture search on object detection," arXiv preprint arXiv:1903.10979, 2019.
[216] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. Yuille, and L. Fei-Fei, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," arXiv preprint arXiv:1901.02985, 2019.
[217] X. Chu, B. Zhang, R. Xu, and H. Ma, "Multi-objective reinforced evolution in mobile neural architecture search," arXiv preprint arXiv:1901.01074, 2019.
[218] C.-H. Hsu, S.-H. Chang, D.-C. Juan, J.-Y. Pan, Y.-T. Chen, W. Wei, and S.-C. Chang, "Monas: Multi-objective neural architecture search using reinforcement learning," arXiv preprint arXiv:1806.10332, 2018.
[219] P. Simard, L. Bottou, P. Haffner, and Y. LeCun, "Boxlets: a fast convolution algorithm for signal processing and neural networks," in Advances in Neural Information Processing Systems, 1999, pp. 571–577.
[220] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 32–39.
[221] F. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 829–836.
[222] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through ffts," arXiv preprint arXiv:1312.5851, 2013.
[223] H. Pratt, B. Williams, F. Coenen, and Y. Zheng, "Fcnn: Fourier convolutional neural networks," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2017, pp. 786–798.
[224] N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, and Y. LeCun, "Fast convolutional nets with fbfft: A gpu performance evaluation," arXiv preprint arXiv:1412.7580, 2014.
[225] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," in Advances in neural information processing systems, 2015, pp. 2449–2457.
[226] C. Dubout and F. Fleuret, "Exact acceleration of linear object detectors," in European Conference on Computer Vision. Springer, 2012, pp. 301–311.
[227] M. A. Sadeghi and D. Forsyth, "Fast template evaluation with vector quantization," in Advances in neural information processing systems, 2013, pp. 2949–2957.
[228] I. Kokkinos, "Bounding part scores for rapid detection with deformable part models," in European Conference on Computer Vision. Springer, 2012, pp. 41–50.
[229] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang et al., "Recent advances in convolutional neural networks," arXiv preprint arXiv:1512.07108, 2015.
[230] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[231] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[232] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.
[233] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in AAAI, vol. 4, 2017, p. 12.
[234] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[235] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." In CVPR, vol. 1, no. 2, 2017, p. 3.
[236] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," arXiv preprint arXiv:1709.01507, vol. 7, 2017.
[237] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scaletransferrable object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 528–537.
[238] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "Dsod: Learning deeply supervised object detectors from scratch," in The IEEE International Conference on Computer Vision (ICCV), vol. 3, no. 6, 2017, p. 7.
[239] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in

Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017, pp. 5987–5995.
[240] J. Jeong, H. Park, and N. Kwak, "Enhancement of ssd by concatenating feature maps for object detection," arXiv preprint arXiv:1705.09587, 2017.
[241] K. Lee, J. Choi, J. Jeong, and N. Kwak, "Residual features and unified prediction network for single stage detection," arXiv preprint arXiv:1707.05031, 2017.
[242] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, "Feature-fused ssd: fast detection for small objects," in Ninth International Conference on Graphic and Image Processing (ICGIP 2017), vol. 10615. International Society for Optics and Photonics, 2018, p. 106151E.
[243] L. Zheng, C. Fu, and Y. Zhao, "Extend the shallow part of single shot multibox detector via convolutional neural network," arXiv preprint arXiv:1801.05918, 2018.
[244] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta, "Beyond skip connections: Top-down modulation for object detection," arXiv preprint arXiv:1612.06851, 2016.
[245] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "Ron: Reverse connection with objectness prior networks for object detection," in IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2017, p. 2.
[246] S. Woo, S. Hwang, and I. S. Kweon, "Stairnet: Top-down semantic aggregation for accurate one shot detection," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 1093–1102.
[247] Y. Chen, J. Li, B. Zhou, J. Feng, and S. Yan, "Weaving multi-scale context for single shot detector," arXiv preprint arXiv:1712.03149, 2017.
[248] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 2018–2025.
[249] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," arXiv preprint arXiv:1701.06659, 2017.
[250] J. Wang, Y. Yuan, and G. Yu, "Face attention network: An effective face detector for the occluded faces," arXiv preprint arXiv:1711.07246, 2017.
[251] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in The IEEE International Conference on Computer Vision (ICCV), vol. 6, no. 7, 2017.
[252] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.
[253] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated residual networks." in CVPR, vol. 2, 2017, p. 3.
[254] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Detnet: A backbone network for object detection," arXiv preprint arXiv:1804.06215, 2018.
[255] S. Liu, D. Huang, and Y. Wang, "Receptive field blocknet for accurate and fast object detection," arXiv preprint arXiv:1711.07767, 2017.
[256] M. Najibi, M. Rastegari, and L. S. Davis, "G-cnn: an iterative grid based object detector," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2369–2377.
[257] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. So Kweon, "Attentionnet: Aggregating weak directions for accurate object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2659–2667.
[258] Y. Lu, T. Javidi, and S. Lazebnik, "Adaptive object detection using adjacency and zoom prediction," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2351–2359.
[259] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 1, pp. 121–135, 2019.
[260] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," arXiv preprint arXiv:1611.08050, 2016.
[261] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in Proceedings of the European Conference on Computer Vision (ECCV), vol. 6, 2018.
[262] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, no. 2, 2018, p. 10.
[263] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "Refinenet: Iterative refinement for accurate object localization," in

Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. IEEE, 2016, pp. 1528–1533.
[264] M.-C. Roh and J.-y. Lee, "Refining faster-rcnn for accurate object detection," in Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on. IEEE, 2017, pp. 514–517.
[265] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016, pp. 516–520.
[266] S. Gidaris and N. Komodakis, "Locnet: Improving localization accuracy for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 789–798.
[267] B. A. Olshausen and D. J. Field, "Emergence of simplecell receptive field properties by learning a sparse code for natural images," Nature, vol. 381, no. 6583, p. 607, 1996.
[268] A. J. Bell and T. J. Sejnowski, "The independent components of natural scenes are edge filters," Vision research, vol. 37, no. 23, pp. 3327–3338, 1997.
[269] S. Brahmbhatt, H. I. Christensen, and J. Hays, "Stuffnet: Using stuffto improve object detection," in Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. IEEE, 2017, pp. 934–943.
[270] A. Shrivastava and A. Gupta, "Contextual priming and feedback for faster r-cnn," in European Conference on Computer Vision. Springer, 2016, pp. 330–348.
[271] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, "Single-shot object detection with enriched semantics," Center for Brains, Minds and Machines (CBMM), Tech. Rep., 2018.
[272] B. Cai, Z. Jiang, H. Zhang, Y. Yao, and S. Nie, "Online exemplar-based fully convolutional network for aircraft detection in remote sensing images," IEEE Geoscience and Remote Sensing Letters, no. 99, pp. 1–5, 2018.
[273] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 98, pp. 119–132, 2014.
[274] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognitiontangent distance and tangent propagation," in Neural networks: tricks of the trade. Springer, 1998, pp. 239–274.
[275] G. Cheng, P. Zhou, and J. Han, "Rifd-cnn: Rotationinvariant and fisher discriminative convolutional neural networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2884–2893.
[276] ——, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 12, pp. 7405–7415, 2016.
[277] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, "Real-time rotation-invariant face detection with progressive calibration networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2295–2303.
[278] M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in Advances in neural information processing systems, 2015, pp. 2017–2025.
[279] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in European Conference on Computer Vision. Springer, 2016, pp. 122–138.
[280] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection–snip," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3578–3587.
[281] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," arXiv preprint arXiv:1805.09300, 2018.
[282] S. Qiao, W. Shen, W. Qiu, C. Liu, and A. L. Yuille "Scalenet: Guiding object proposal generation in supermarkets and beyond." in ICCV, 2017, pp. 1809–1818.
[283] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, "Scaleaware face detection," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 3, 2017.
[284] R. Zhu, S. Zhang, X. Wang, L. Wen, H. Shi, L. Bo, and T. Mei, "Scratchdet: Exploring to train single-shot object detectors from scratch," arXiv preprint arXiv:1810.08425, 2018.
[285] K. He, R. Girshick, and P. Dollar, "Rethinking imagenet ´ pre-training," arXiv preprint arXiv:1811.08883, 2018.
[286] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,

D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,
"Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.
[287] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434,
2015.
[288] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired
image-to-image translation using cycle-consistent adversarial networks," arXiv preprint, 2017.
[289] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunning- ´
ham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang
et al., "Photo-realistic single image super-resolution using
a generative adversarial network." in CVPR, vol. 2, no. 3,
2017, p. 4.
[290] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in IEEE CVPR, 2017.
[291] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan:
Small object detection via multi-task generative adversarial network," Computer Vision-ECCV, pp. 8–14, 2018.
[292] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn:
Hard positive generation via adversary for object detection," in IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[293] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau,
"Robust physical adversarial attack on faster r-cnn object
detector," arXiv preprint arXiv:1804.05810, 2018.
[294] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," IEEE transactions on pattern analysis and machine
intelligence, vol. 39, no. 1, pp. 189–203, 2017.
[295] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari,
"We don't need no bounding-boxes: Training object class
detectors using only human verification," in Proceedings
of the IEEE Conference on Computer Vision and Pattern
Recognition, 2016, pp. 854–863.
[296] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, ´
"Solving the multiple instance problem with axis-parallel
rectangles," Artificial intelligence, vol. 89, no. 1-2, pp. 31–
71, 1997.
[297] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Soft proposal
networks for weakly supervised object localization," in
Proc. IEEE Int. Conf. Comput. Vis.(ICCV), 2017, pp. 1841–
1850.
[298] A. Diba, V. Sharma, A. M. Pazandeh, H. Pirsiavash, and
L. Van Gool, "Weakly supervised cascaded convolutional
networks." in CVPR, vol. 1, no. 2, 2017, p. 8.
[299] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[300] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2846–2854.
[301] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani, "Self-taught object localization with deep networks," in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016, pp. 1–9.
[302] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang, "Generative adversarial learning towards fast weakly supervised detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5764–5773.

[303] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 12, pp. 2179–2195, 2008.
[304] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 7, pp. 1239–1258, 2010.
[305] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in European Conference on Computer Vision. Springer, 2014, pp. 613–627.
[306] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1259–1267.
[307] ——, "Towards reaching human performance in pedestrian detection," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 973–986, 2018.
[308] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," International Journal of Computer Vision, vol. 63, no. 2, pp. 153–161, 2005.
[309] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007, pp. 1–8.
[310] J. Cao, Y. Pang, and X. Li, "Pedestrian detection inspired by appearance constancy and shape symmetry," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1316–1324.
[311] R. Benenson, R. Timofte, and L. Van Gool, "Stixels estimation without depth map computation," in Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. IEEE, 2011, pp. 2010–2017.
[312] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4073–4082.
[313] 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.
[314] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 6034–6043.
[315] Q. Hu, P. Wang, C. Shen, A. van den 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.
[316] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5079–5087.
[317] 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 (CVPR), 2017.
[318] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," arXiv preprint arXiv:1711.07752, 2017.
[319] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1904–1912.
[320] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "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.
[321] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6995–7003.
[322] P. Hu and D. Ramanan, "Finding tiny faces," in Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017, pp. 1522–1530.
[323] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 1, pp. 34–58, 2002.
[324] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: past, present and future," Computer

Vision and Image Understanding, vol. 138, pp. 1–24, 2015.
[325] H. A. Rowley, S. Baluja, and T. Kanade, "Neural networkbased face detection," IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 1, pp. 23–38, 1998.
[326] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in Computer vision and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on. IEEE, 1997, pp. 130–136.
[327] M. Osadchy, Y. L. Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," Journal of Machine Learning Research, vol. 8, no. May, pp. 1197–1215, 2007.
[328] 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.
[329] S. Yang, Y. Xiong, C. C. Loy, and X. Tang, "Face detection through scale-friendly deep convolutional networks," arXiv preprint arXiv:1706.02863, 2017.
[330] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "Ssh: Single stage headless face detector." in ICCV, 2017, pp. 4885–4894.
[331] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Sˆ 3fd: Single shot scale-invariant face detector," in Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 192–201.
[332] X. Liu, "A camera phone based currency reader for the visually impaired," in Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility. ACM, 2008, pp. 305–306.
[333] 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 Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on. IEEE, 2005, pp. 257–261.
[334] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," in Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012, pp. 3288–3291.
[335] Z. Wojna, A. Gorban, D.-S. Lee, K. Murphy, Q. Yu, Y. Li, and J. Ibarz, "Attention-based extraction of structured information from street view imagery," arXiv preprint arXiv:1704.03549, 2017.
[336] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in Proc. CVPR, 2017, pp. 3454–3461.
[337] Y. Wu and P. Natarajan, "Self-organized text detection with minimal post-processing via border learning," in Proc. ICCV, 2017.
[338] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," Frontiers of Computer Science, vol. 10, no. 1, pp. 19–36, 2016.
[339] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 7, pp. 1480–1500, 2015.
[340] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 97–104.
[341] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 5, pp.970–983, 2014.
[342] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 1457–1464.
[343] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012, pp. 3304–3308.
[344] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. Lim Tan, "Text flow: A unified text detection system in natural scene images," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4651–4659.
[345] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in European conference on computer vision. Springer, 2014, pp. 512–528.
[346] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multiorientation scene text detection with adaptive clustering," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 9, pp. 1930–1937, 2015.
[347] Z. Zhang, W. Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2558–2567.

[348] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," International Journal of Computer Vision, vol. 116, no. 1, pp. 1–20, 2016.
[349] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in European Conference on Computer Vision. Springer, 2014, pp. 497–511.
[350] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-attentional convolutional neural network for scene text detection," IEEE transactions on image processing, vol. 25, no. 6, pp. 2529–2541, 2016.
[351] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," IEEE Transactions on Multimedia, 2018.
[352] ——, "Arbitrary-oriented scene text detection via rotation proposals," IEEE Transactions on Multimedia, 2018.
[353] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2cnn: rotational region cnn for orientation robust scene text detection," arXiv preprint arXiv:1706.09579, 2017.
[354] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network." in AAAI, 2017, pp. 4161–4167.
[355] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," arXiv preprint arXiv:1703.08289, 2017.
[356] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in Proc. CVPR, 2017, pp. 3454–3461.
[357] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in Proc. CVPR, 2017, pp. 2642–2651.
[358] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," arXiv preprint arXiv:1606.09002, 2016.
[359] C. Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in European Conference on Computer Vision. Springer, 2018,pp. 370–387.
[360] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7553–7563.
[361] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in European conference on computer vision. Springer, 2016, pp. 56–72.
[362] A. d. l. Escalera, L. Moreno, M. A. Salichs, and J. M. Armingol, "Road traffic sign detection and classification," 1997.
[363] D. M. Gavrila, U. Franke, C. Wohler, and S. Gorzig, "Real time vision for intelligent vehicles," IEEE Instrumentation & Measurement Magazine, vol. 4, no. 2, pp. 22–27, 2001.
[364] C. F. Paulo and P. L. Correia, "Automatic detection and classification of traffic signs," in Image Analysis for Multimedia Interactive Services, 2007. WIAMIS'07. Eighth International Workshop on. IEEE, 2007, pp. 11–11.
[365] A. De la Escalera, J. M. Armingol, and M. Mata, "Traffic sign recognition and analysis for intelligent vehicles," Image and vision computing, vol. 21, no. 3, pp. 247–258, 2003.
[366] W. Shadeed, D. I. Abu-Al-Nadi, and M. J. Mismar, "Road traffic sign detection in color images," in Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003 10th IEEE International Conference on, vol. 2. IEEE, 2003, pp. 890–893.
[367] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gomez-Moreno, and F. L ´ opez-Ferreras, "Road-sign detection and recognition based on support vector machines," IEEE transactions on intelligent transportation systems, vol. 8, no. 2, pp. 264–278, 2007.
[368] M. Omachi and S. Omachi, "Traffic light detection with color and edge information," 2009.
[369] Y. Xie, L.-f. Liu, C.-h. Li, and Y.-y. Qu, "Unifying visual saliency with hog feature learning for traffic sign detection," in Intelligent Vehicles Symposium, 2009 IEEE. IEEE, 2009, pp. 24–29.
[370] S. Houben, "A single target voting scheme for traffic sign detection," in Intelligent Vehicles Symposium (IV), 2011 EEE. IEEE, 2011, pp. 124–129.
[371] A. Soetedjo and K. Yamada, "Fast and robust traffic sign etection," in Systems, Man and Cybernetics, 2005 IEEE nternational Conference on, vol. 2. IEEE, 2005, pp. 1341–1346.
[372] N. Fairfield and C. Urmson, "Traffic light mapping and detection," in Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011, pp. 5421–5426.

[373] J. Levinson, J. Askeland, J. Dolson, and S. Thrun, "Traffic light mapping, localization, and state detection for autonomous vehicles," in Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011, pp.5784–5791.
[374] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler, "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information," in Intelligent Vehicles Symposium, 2005. Proceedings. IEEE. IEEE, 2005, pp. 255–260.
[375] I. M. Creusen, R. G. Wijnhoven, E. Herbschleb, and P. de With, "Color exploitation in hog-based traffic sign detection," in 2010 IEEE International Conference on Image Processing. IEEE, 2010, pp. 2669–2672.
[376] G. Wang, G. Ren, Z. Wu, Y. Zhao, and L. Jiang, "A robust, coarse-to-fine traffic sign detection method," in Neural Networks (IJCNN), The 2013 International Joint Conference on. IEEE, 2013, pp. 1–5.
[377] Z. Shi, Z. Zou, and C. Zhang, "Real-time traffic light detection with adaptive background suppression filter," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 3, pp. 690–700, 2016.
[378] Y. Lu, J. Lu, S. Zhang, and P. Hall, "Traffic signal detection and classification in street views using an attention model," Computational Visual Media, vol. 4, no. 3, pp. 253–266, 2018.
[379] M. Bach, D. Stumper, and K. Dietmayer, "Deep convolutional traffic light recognition for automated driving," in 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 851–858.
[380] S. Qiu, G. Wen, and Y. Fan, "Occluded object detection in high-resolution remote sensing images using partial configuration object model," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 5, pp. 1909–1925, 2017.
[381] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with svd networks," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 10, pp. 5832–5845,2016.
[382] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," IEEE Geoscience and Remote Sensing Magazine, vol. 4, no. 2, pp. 22–40, 2016.
[383] N. Proia and V. Page, "Characterization of a bayesian ship detection method in optical satellite images," IEEE Geoscience and Remote Sensing Letters, vol. 7, no. 2, pp. 226–230, 2010.
[384] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," IEEE Transactions on geoscience and remote sensing, vol. 48, no. 9, pp. 3446–3456, 2010.
[385] S. Qi, J. Ma, J. Lin, Y. Li, and J. Tian, "Unsupervised ship detection based on saliency and s-hog descriptor from optical satellite images," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 7, pp. 1451–1455, 2015.
[386] F. Bi, B. Zhu, L. Gao, and M. Bian, "A visual search inspired computational model for ship detection in optical satellite images," IEEE Geoscience and Remote Sensing Letters, vol. 9, no. 4, pp. 749–753, 2012.
[387] J. Han, P. Zhou, D. Zhang, G. Cheng, L. Guo, Z. Liu, S. Bu, and J. Wu, "Efficient, simultaneous detection of multiclass geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 89, pp.37–48, 2014.
[388] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 6, pp. 3325–3337, 2015.
[389] J. Tang, C. Deng, G.-B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 3, pp. 1174–1185, 2015.
[390] Z. Shi, X. Yu, Z. Jiang, and B. Li, "Ship detection in highresolution optical imagery based on anomaly detector and local shape feature," IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 8, pp. 4511–4523, 2014.
[391] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 6, pp. 1250–1265, 2011.
[392] L. Wan, L. Zheng, H. Huo, and T. Fang, "Affine invariant description and large-margin dimensionality reduction for target detection in optical remote sensing images," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 7, pp. 1116–1120, 2017.
[393] H. Zhou, L. Wei, C. P. Lim, D. Creighton, and S. Nahavandi, "Robust vehicle detection in aerial images using bag-of-

words and orientation aware scanning," IEEE Transactions on Geoscience and Remote Sensing, no. 99, pp.1–12, 2018.
[394] M. ElMikaty and T. Stathaki, "Detection of cars in high-resolution aerial images of complex urban environments," IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 10, pp. 5913–5924, 2017.
[395] L. Zhang, Z. Shi, and J. Wu, "A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 10, pp. 4895–4909, 2015.
[396] C. Zhu, B. Liu, Y. Zhou, Q. Yu, X. Liu, and W. Yu, "Framework design and implementation for oil tank detection in optical satellite imagery," in Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International. IEEE, 2012, pp. 6016–6019.
[397] G. Liu, Y. Zhang, X. Zheng, X. Sun, K. Fu, and H. Wang, "A new method on inshore ship detection in highresolution satellite images using shape and context information," IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 3, pp. 617–621, 2014.
[398] J. Xu, X. Sun, D. Zhang, and K. Fu, "Automatic detection of inshore ships in high-resolution remote sensing images using robust invariant generalized hough transform," IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 12, pp. 2070–2074, 2014.
[399] J. Zhang, C. Tao, and Z. Zou, "An on-road vehicle detection method for high-resolution aerial images based on local and global structure learning," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 8, pp. 1198–1202, 2017.
[400] W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang, and K. Fu, "Efficient saliency-based object detection in remote sensing images using deep belief networks," IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 2, pp. 137–141, 2016.
[401] P. Zhang, X. Niu, Y. Dou, and F. Xia, "Airport detection on optical satellite images using deep convolutional neural networks," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 8, pp. 1183–1187, 2017.
[402] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 6, pp. 3623–3634, 2017.
[403] X. Han, Y. Zhong, and L. Zhang, "An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery," Remote Sensing, vol. 9, no. 7, p. 666, 2017.
[404] Z. Xu, X. Xu, L. Wang, R. Yang, and F. Pu, "Deformable convnet with aspect ratio constrained nms for object detection in remote sensing imagery," Remote Sensing, vol. 9, no. 12, p. 1312, 2017.
[405] W. Li, K. Fu, H. Sun, X. Sun, Z. Guo, M. Yan, and X. Zheng, "Integrated localization and recognition for inshore ships in large scene remote sensing images," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 6, pp.936–940, 2017.
[406] O. A. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 44–51.
[407] L. W. Sommer, T. Schuchert, and J. Beyerer, "Fast deep vehicle detection in aerial images," in Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. IEEE, 2017, pp. 311–319.
[408] L. Sommer, T. Schuchert, and J. Beyerer, "Comprehensive analysis of deep learning based vehicle detection in aerial images," IEEE Transactions on Circuits and Systems for Video Technology, 2018.
[409] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based cnn for ship detection," in Image Processing (ICIP), 2017 IEEE International Conference on. IEEE, 2017, pp. 900–904.
[410] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 10, pp. 1665–1669, 2017.
[411] ——, "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network," Remote Sensing, vol. 9, no. 5, p. 480, 2017.