**Recent Advances in Object Detection**

**Contenets**

[1. Introduction 2](#_Toc26641267)

[2. Detection Components 4](#_Toc26641268)

[2.1. Detection Settings 4](#_Toc26641269)

[2.2. Detection Paradigms 4](#_Toc26641270)

[2.2.1. Two-Stage 4](#_Toc26641271)

[2.2.2. One-Stage 6](#_Toc26641272)

[2.3. Backbone Architecture 7](#_Toc26641273)

[2.4. Proposal Generation 8](#_Toc26641274)

[2.4.1. Traditional Computer Vision Methods 8](#_Toc26641275)

[2.4.2. Anchor-based Methods 8](#_Toc26641276)

[2.4.3. Keypoints-based Methods 9](#_Toc26641277)

[2.4.4. Other Methods 10](#_Toc26641278)

[2.5. Feature Representation 10](#_Toc26641279)

[2.5.1. Multi-scale Feature Learning 10](#_Toc26641280)

[2.5.2. Region Feature Encoding 11](#_Toc26641281)

[2.5.3. Contextual Reasoning 11](#_Toc26641282)

[2.5.4. Deformable Feature Learning 12](#_Toc26641283)

[3. Learning Strategy 13](#_Toc26641284)

[3.1. Training Stage 13](#_Toc26641285)

[3.1.1. Data Augmentation 13](#_Toc26641286)

[3.1.2. Imbalance Sampling 13](#_Toc26641287)

[3.1.3. Localization Refinement 13](#_Toc26641288)

[3.1.4. Cascade Learning 13](#_Toc26641289)

[3.1.5. Others 14](#_Toc26641290)

[3.2. Testing Stage 15](#_Toc26641291)

[3.2.1. Duplicate Removal 15](#_Toc26641292)

[3.2.2. Model Acceleration 15](#_Toc26641293)

[3.2.3. Others 15](#_Toc26641294)

[4. Applications and Detection Benchmarks 16](#_Toc26641295)

[4.1. Generic Detection Benchmarks 16](#_Toc26641296)

[4.2. Face Detection Benchmarks 16](#_Toc26641297)

[4.3. Pedestrian Detection Benchmarks 17](#_Toc26641298)

[4.4. Evaluation Metrics 18](#_Toc26641299)

[5. SOTA for Generic Object Detection 19](#_Toc26641300)

[6. Future Directions 21](#_Toc26641301)

**参考资料：**arXiv:1908.03673及文中提及的相关文章。

# Introduction

基础视觉任务：图像分类、目标检测、语义分割、实例分割

目标检测任务：定位、分类

传统目标检测方法的三个步骤：

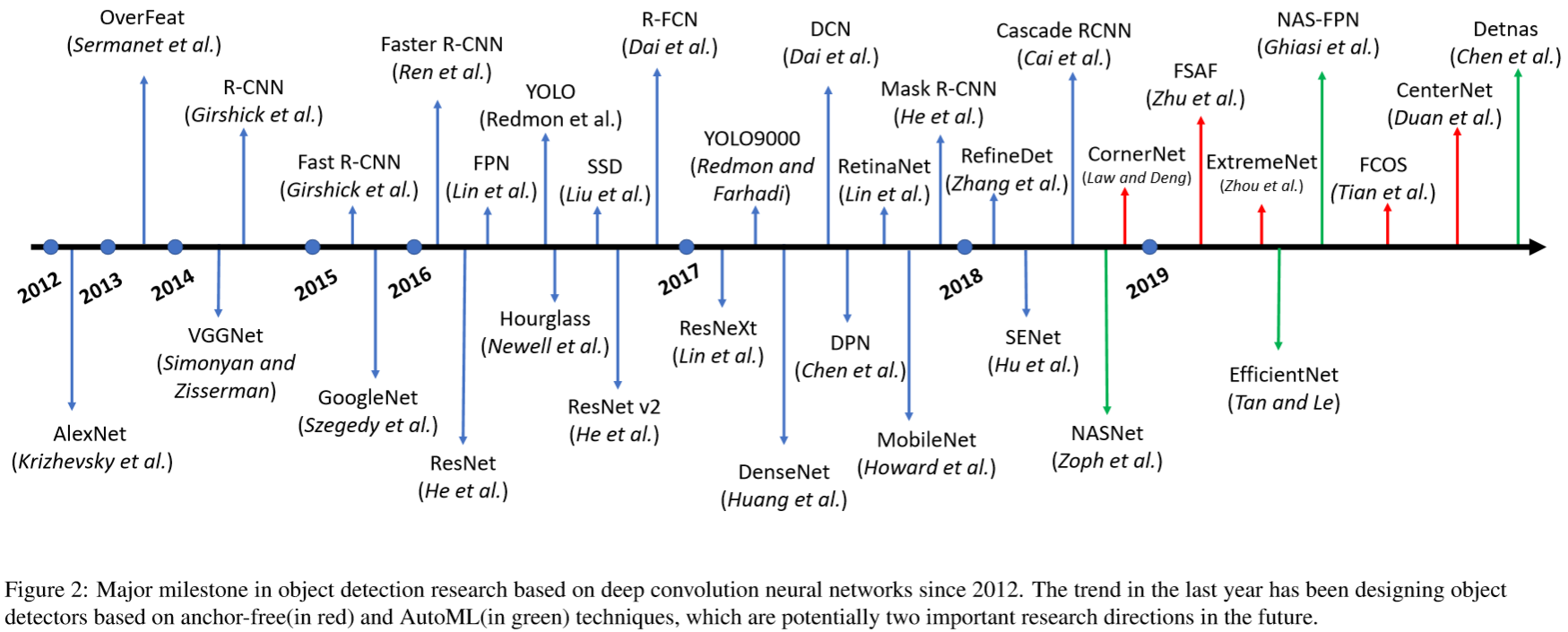
1. proposal generation：生成ROI（regions of interest）
2. feature vector extraction：提取特征
3. region classiﬁcation：分类

传统方法的工作主要聚焦于设计更好的特征描述子（feature descriptors），将RoI信息映射为embedding feature。

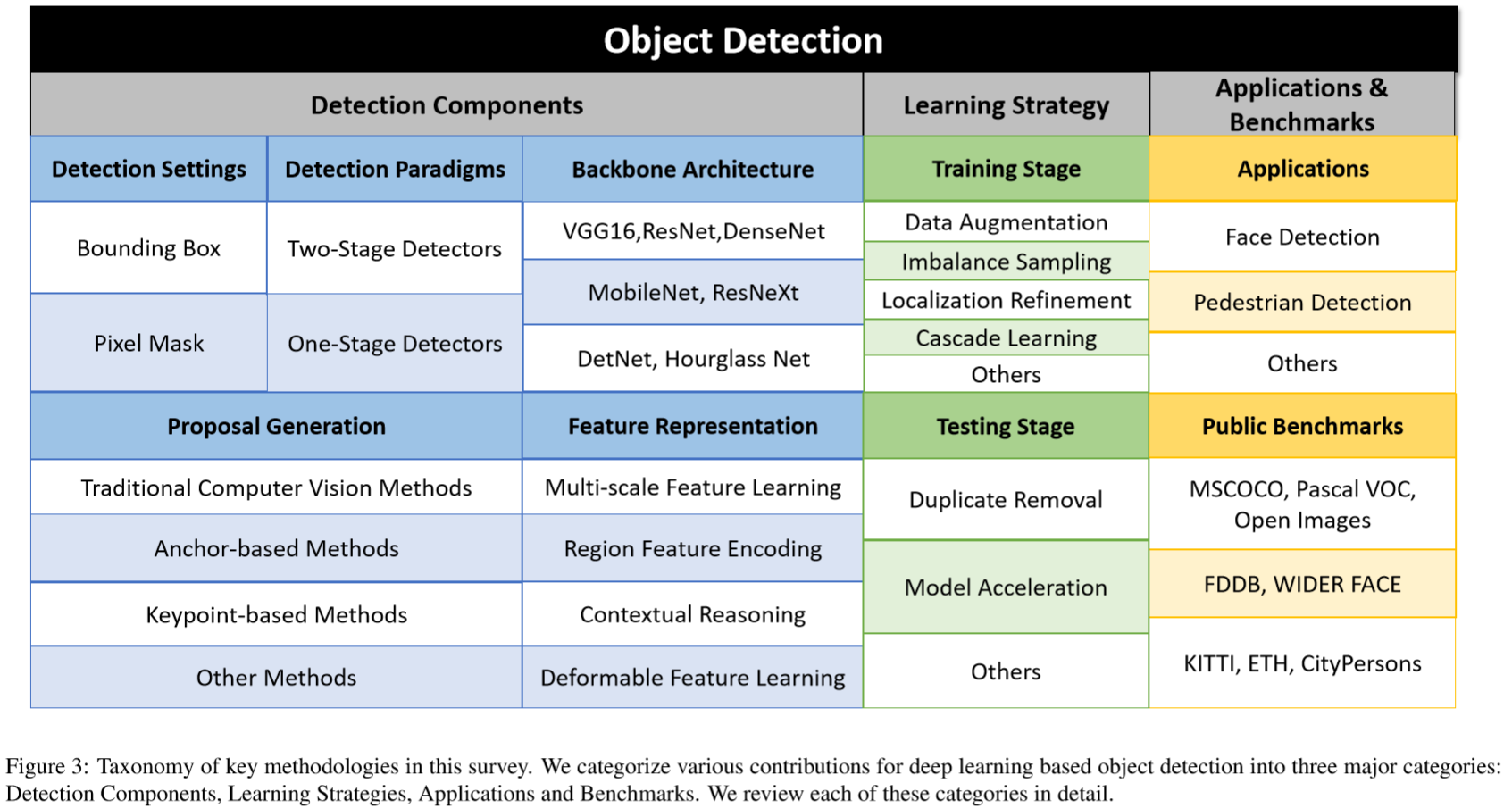
传统方法的瓶颈在于：

* 生成大量冗余的proposal，分类时存在大量FP；
* 滑动窗口基于手工、启发式设计的，效果不好；
* 特征描述子基于低级特征进行手工设计，难以捕捉高级语义特征和复杂内容；
* 检测的每个步骤是独立的，无法全局优化。

近年来基于深度学习的目标检测方法：

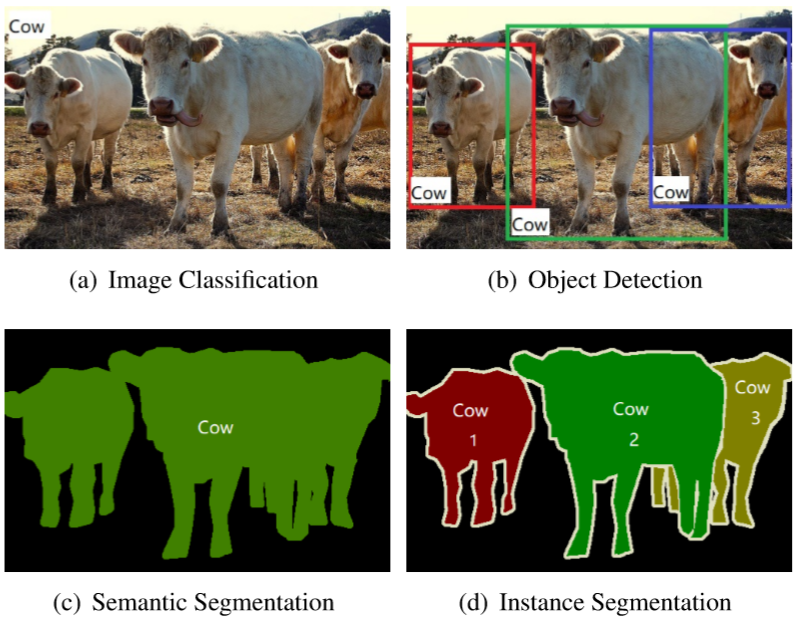


文献综述结构：



# Detection Components

## Detection Settings



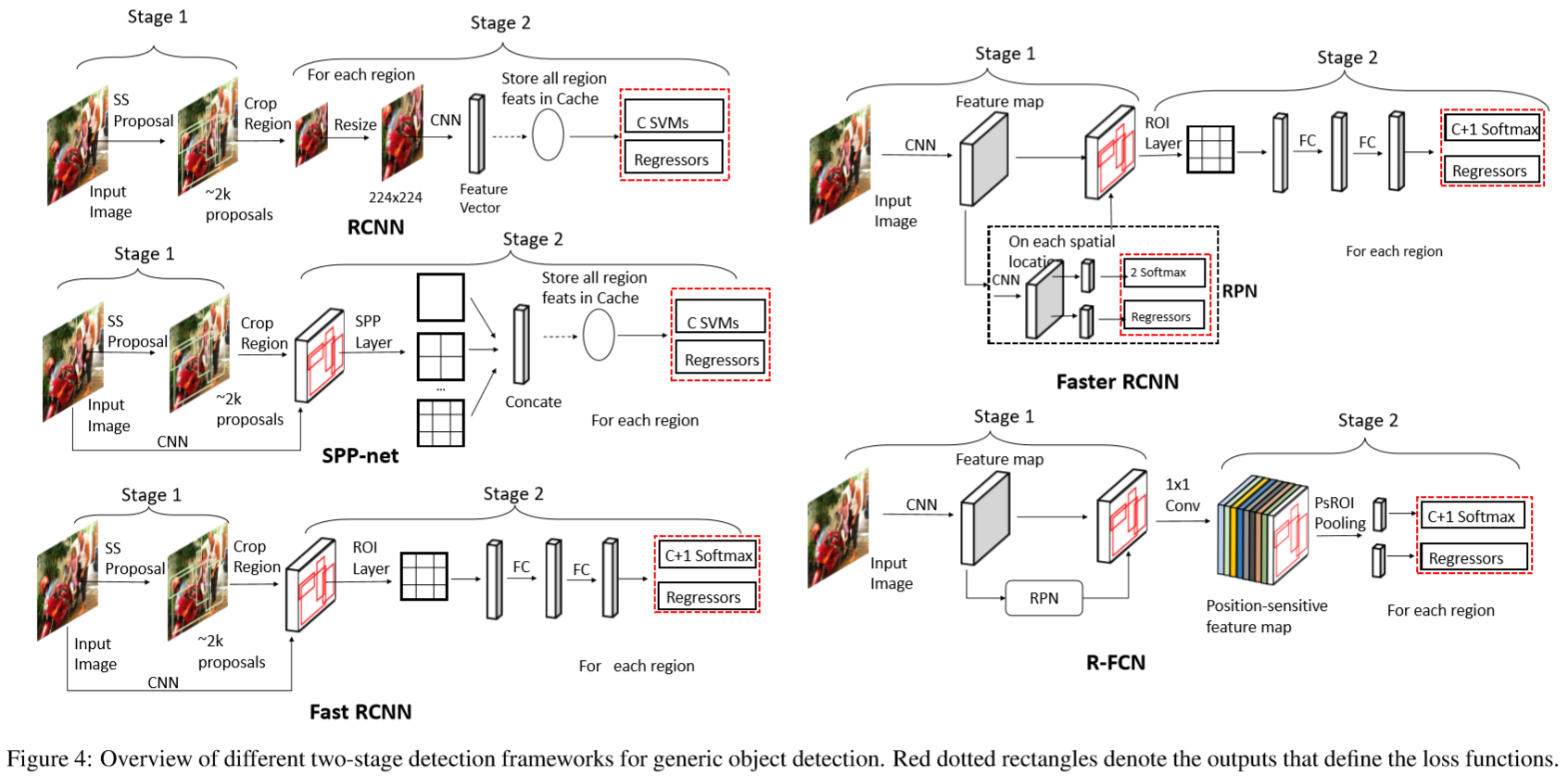
1. Bounding Box：简单目标检测，图(b)
2. Pixel Mask：实例分割，图(d)

## Detection Paradigms

### Two-Stage

思路上分两步：proposal generation；making predictions。

**方法图示：**



**方法对比：**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | 冗余 | Region proposals | Anchor | Pooling | 端到端 |
| R-CNN | 多 | Selective Search | 无 | 无 | 否 |
| SPP-Net | 少 | SPP Layer | 否 |
| Fast\_R-CNN | ROI Pooling | 是 |
| Faster\_R-CNN | RPN | 有 | 否 |
| R-FCN | 更少 | 位置敏感ROI Pooling | 否 |

（Two-Stage方法对比）

**方法核心思路：**

##### R-CNN

该系列方法的先驱。生成ROI；CNN提取特征；分类；位置回归。

##### SPP-Net

1. 将整张图像输入CNN，减少冗余计算；
2. 设计SPP Layer，处理任意大小、比例的输入；

缺点：SPP Layer不能反向传播梯度到CNN。

##### Fast\_R-CNN

1. 使用Softmax层与Regressor层，实现端到端训练；
2. ROI Pooling解决了SPP Layer不能反向传播梯度到CNN的问题。

启发：通过多目标学习同时学习目标类别与位置，相对于R-CNN或SPP-Net那样对类别与位置分别进行学习，可取得更好的mAP值；

##### Faster\_R-CNN

1. 提出Region Proposal Network (RPN)，使用深度学习的方法生成RoI；
2. 提出Anchor Box。

缺点：以端到端方式联合训练RPN和Fast\_R-CNN效果不好，作者用的训练流程较为复杂。

##### R-FCN

1. 提出Position Sensitive Score Map，Position Sensitive ROI Pooling layer (PSROI Pooling)；
2. 全卷积网络。进一步减少冗余（之前的结构每个ROI单独经过FC，造成冗余）。

启发：全连接层对回归空间信息很重要，简单地去掉FC会严重影响性能。

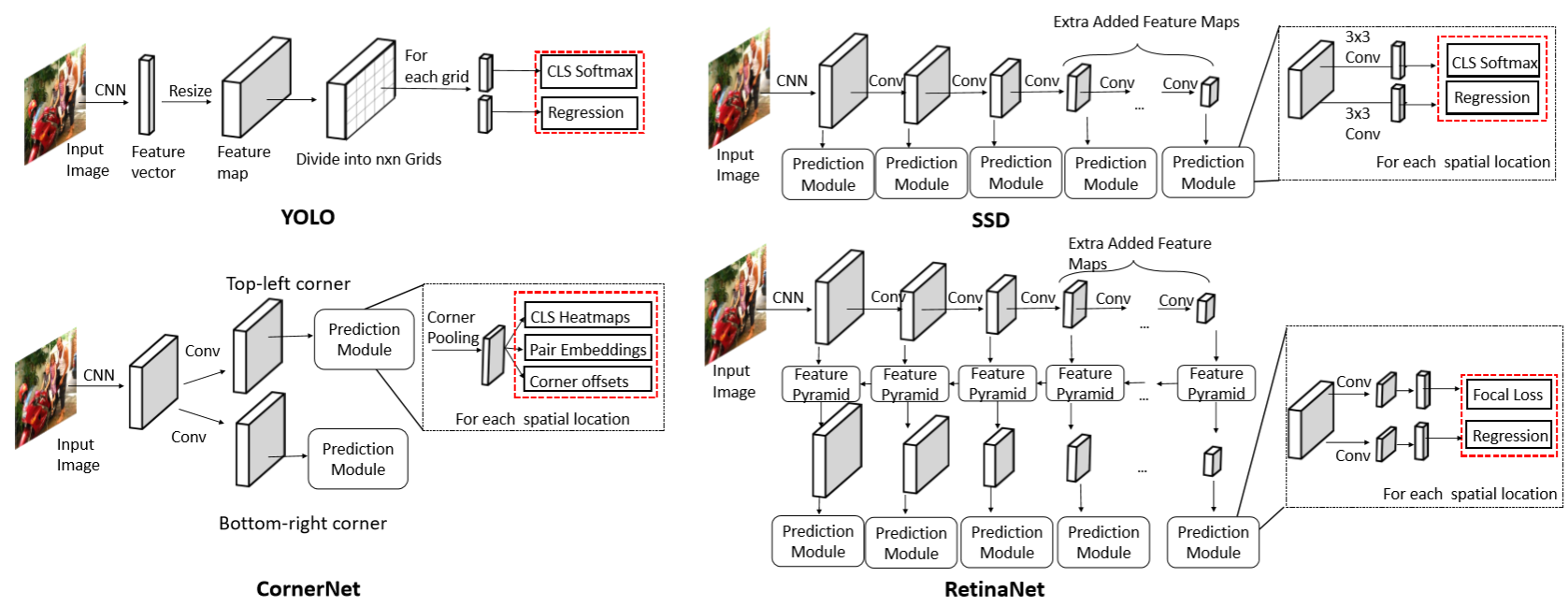
##### 其他

1. FPN的作者在论文中提出了一个two-stage网络
2. Mask\_R-CNN等语义分割算法也属于two-stage方法

### One-Stage

将所有点看作潜在的RoI。与two-stage相比精度低，速度高。目标检测任务中，20FPS一般被认为达到实时

部分方法图示：



方法核心思路：

OverFeat

早期基于CNN的one-stage方法。关于检测框架的描述不多。

YOLOv1

将图像划分为网格，预测每个网格里是否有目标、目标位置、目标类别。

缺点：一个网格只能预测一个物体，难以检测小物体和密集物体。

**YOLOv2：**借鉴SSD，使用Anchor。通过对训练集聚类生成合适的Anchor。

**YOLOv3：**一些针对v2的改进。

特点：相比于two-stage方法追求精度，YOLO系列追求速度。

SSD

1. 借鉴YOLO，将图像划分为网格。在每个网格中设置Anchor；
2. 使用多个特征图联合预测。

缺点：小目标的检测效果一般。

特点：SSD系列致力于精度和速度的折中。

DSSD

1. 基于SSD，使用residual模块设计新的预测模块；
2. 使用反卷积做进一步的特征融合。

RSSD

基于SSD。用了新的特征融合方式，使各个特征图具有相同的通道数。实现各个classifier network间参数共享，降低了训练难度。

CornerNet

anchor-free，预测对角点坐标。

CenterNet

anchor-free。预测每个像素点是目标中心的概率，并回归其宽高。相较于CornerNet，中心点的语义信息更明显。

LFFD

1. 针对人脸检测，利用人脸比例接近1:1这一特点，设计类似anchor-free的检测框架；
2. 结合上下文信息以检测小目标。

RON & RefineDet

属于one-stage和two-stage结合的方法。two-stage是用一个比较复杂的方法生成ROI，RON & RefineDet是用一个非常简单的方法生成ROI，再用一个比较简单的方法过滤这些ROI

其他

RetinaNet也属于one-stage方法，本质上是Resnet + FPN + 两个FCN子网络

## Backbone Architecture

主流分类网络

* VGG16：更深的网络，更好的表达能力；
* ResNet：降低了深层神经网络的训练难度；
* ResNetv2：BN能够提升模型效果；
* DenseNet：ResNet的越层链接中，浅层的原始特征在像素级的操作上会丢失信息，不能很好利用所有的特征；DenseNet采用密集连接以及concat方式充分利用之前各层的特征信息。
* DPN：DenseNet在信息组合过程中产生大量冗余，DPN对ResNet+DenseNet进行concat，既能组合新的特征，也能降低冗余；
* ResNeXt：设计group convolution layers，减少计算量的同时保持模型准确度；
* MobileNet：精度换时间；
* GoogleNet：更宽的网络，融合多尺度特征；
* DLA：更优的连接方式，减少计算量的同时提升性能。

分类与检测的潜在冲突

1. 分类需要较大的RF，并希望保持空间不变性。因此，应用了多个下采样操作（例如池化层）以降低特征图的分辨率。生成的特征图是低分辨率且空间不变的，并且具有较大的接收范围。然而，在检测中，需要高分辨率的空间信息以正确地定位对象
2. 分类在单个特征图上做出预测，而检测则需要具有多种表示形式的特征图以多尺度检测物体

其他网络结构

* DetNet：检测特化网络。保持高分辨率；使用扩展卷积扩大RF；使用多尺度特征
* Hourglass：HPR网络。下采样+上采样+skip connect

## Proposal Generation

### Traditional Computer Vision Methods

传统方法基于低级特征如方向、边缘、色彩等。

分为三种方式：

1. Objectness Score based：给每个候选框打分。
2. Superpixels Merging：如Selective Search。不断合并小区域。
3. Seed Segmentation：不断分割大区域。

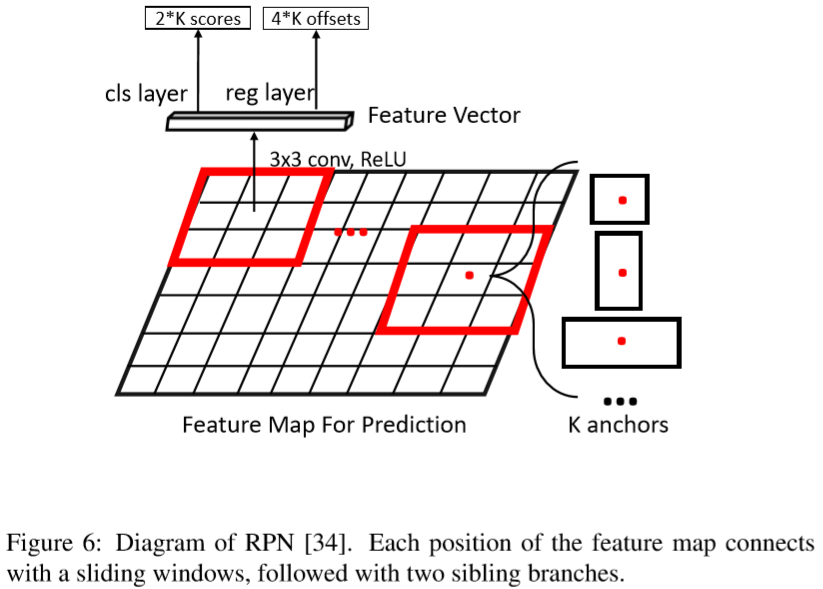
缺点

1. 无法联合优化；
2. 虽然这些方法很简单，但是耗时未必就短；
3. 难以应用到复杂的大型数据集中。

### Anchor-based Methods

基于预定义的anchor先验，使用卷积网络以监督的方式生成ROI。anchor先验的设置是手动启发式设计的。该系列方法的核心在于设计好的anchor。

该系列先驱方法RPN的示意图如下：



主要方法

* Faster\_R-CNN：RPN。
* SSD：采用了和RPN类似的设计。不同的是，SSD给每个anchor分配了类别概率。
* S3FD：设计anchor时考虑了不同特征图上的RF
* Seeing small faces from robust anchors perspective：放大输入图像，减小anchor步长，以识别小物体
* DeRPN：将anchor回归的四维信息分解成两个线段信息，易于匹配尺度范围大的物体，减少搜索空间
* DeepProposal（ICCV2015）从低分辨率深层特征图生成anchor，用高分辨率浅层特征图微调；
* YOLOv2：对训练数据进行k-means聚类
* RefineDet（CVPR2018）：对手工设计的anchor进行逐级微调
* Cascade\_RCNN（CVPR2018）：RefineDet类似，与通过Iou阈值的match观察，通过iou阈值的筛选逐渐提高高质量proposal的比例；
* MetaAnchor（NIPS2018）：使用神经网络学习如何生成anchor，但是其自定义的基础anchor组仍是手工设计的

### Keypoints-based Methods

一般又称anchor-free方法，分为两类

corner-based：Denet预测四个角；CornerNet预测对角。需要对角点进行分组。

center-based：预测每个位置是目标中心点的概率并在此基础上预测宽高。如FSAF，CenterNet

两种思路结合：arXiv:1904.08189先用对角预测，再用中心预测筛选。

### Other Methods

AZNet（CVPR2016），通过将整张图片进行不断划分子区域来生成proposal，适用于检测稀疏的小物体，对于一般检测不是特别好用。

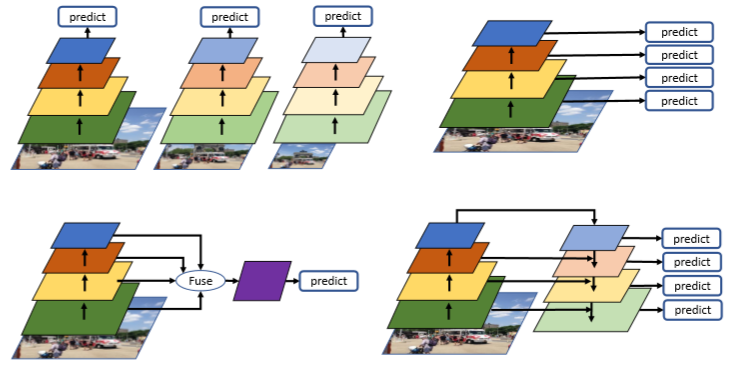
## Feature Representation

### Multi-scale Feature Learning

在单个特征图上检测尺度和比例范围较大的对象非常困难。浅层特征具有丰富空间信息，更高的分辨率和较小的RF，更适合于检测小物体。深层特征语义丰富鲁棒性较强，具有较大的RF，适合检测大物体。

检测小物体需要高分辨率的特征，这在深层特征中困难无法获取。对此dilated/atrous convolutions被提出用来避免下采样，可以在深层中保留高分辨率信息。另外，没有足够大的RF，在浅层检测大型物体也比较困难。

目前有四种主流方法来获取多尺度特征：



**图像金字塔**

左上图所示。训练多个检测器，每个检测器用于处理不同大小的图像，检测时将原图像resize成多个大小。计算消耗巨大，一般不用。类似的方法有SNIP，训练多个检测器，每个检测器用于检测不同大小的目标。

**特征整合**

左下图所示。合并不同层的特征。如ION将不同层的特征通过RoI pooling进行裁剪融合；另外还有HyperNet，MLKP等。现在用的不多。

**预测金字塔**

右上所示。使用多层特征一起预测。最早在SSD中使用。Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classiﬁers使用特征图生成特定尺度的目标提案，这些提案被送入多个scale-dependent的分类器做联合预测；MSCNN使用反卷积改善了特征图的分辨率。这种多尺度预测现在用的很多。

**特征金字塔**

可以理解成特征整合+预测金字塔。生成一组尺度不变的特征图，送入多个scale-dependent的分类器做联合预测。经典方法是FPN；针对FPN的变体很多，思路大同小异。

### Region Feature Encoding

* Resize+CNN：早期使用，已被淘汰；
* RoI Pooling：Fast\_R-CNN中使用。使用下采样的特征图，难以处理小物体；
* RoI Warping：使用双线插值。使用下采样的特征图，pooling和warping会产生定位偏差；
* RoI Align Pooling：在每个网格内的部分采样位置处的双线性插值解决了量化问题；
* PrRoI pooling：避免了量化问题，在bbox坐标上有连续的梯度；
* PSRoI Pooling：R-FCN中使用。在下采样的特征图中保留了相对空间信息；
* Feature selective networks for object detection：利用子区域和纵横比之间的差异来学习鲁棒的区域特征；
* Couplenet：整合ROI Pooling的全局信息和PSROI Pooling的局部信息；
* Deformable RoI Pooling：学习每个网格的offset并将其添加到网格中心来泛化Align Pooling。

### Contextual Reasoning

有效地使用上下文信息可以帮助提高检测性能，尤其是在检测线索不足（小对象，遮挡等）的对象时。基于深度学习的对象检测尚未广泛地探索上下文的使用，这是因为卷积网络已经隐式地从分层特征表示中捕获了上下文信息。但是其实现实世界背景的关联性没有那么强，比如人可以出现在很多很复杂的环境中，这和人脸检测的强结构关联不同，所以也有文献指出盲目加上下文会使效果变差。

Global context reasoning

学习整张图片的上下文信息，并通过这些上下文信息进行目标的推断。代表工作有：

* ION：使用RNN从四个方向对整个图像的上下文信息进行编码
* DeepId：学习每个图像的分类分数，将其作为上下文特征与检测结果关联
* 改进版的Faster R-CNN：提取了整个图像的特征，将其与区域特征关联以改善检测结果
* 一些其他方法使用语义分割来获取全局信息，如Mask\_R-CNN

Region Context Reasoning

区域上下文推理只编码候选区域附近的上下文信息，即检测目标与其周围环境的关系。直接建模物体与其周围环境的关系是很难的，因为目标的周围环境具有很大的变化性。相关工作有：

* SMN（ICCV2017）：将对象实例重新组合为伪“图像”来获取实例级上下文，然后将其用于对象关系推理；
* SIN（CVPR2018）：将物体作为节点，关系作为边，把对象检测转换为图推理问题；
* Relation Network（CVPR2018）：建模不同对象之间的外观和图像位置之间的相互作用；
* Learning region features for object detection（ECCV2018）替换了ROI Polling中的启发式选择，自动选择了最重要的部分，包括ROI之外的上下文。

有一类方法是添加ROI周围区域的图像特征来隐式编码上下文信息：

* Object detection via a multi-region and semantic segmentation-aware cnn model（ICCV2015）：从ROI的多个不同子区域（边界，中央区域，上下文）中提取特征，并将这些特征与原ROI特征连接起来；
* A uniﬁed multi-scale deep convolutional neural network for fast object detection（ECCV2016）扩大ROI窗口的大小并将这些特征与原ROI连接起来
* GBDNet：从多尺度的子区域提取特征，学习门函数控控制不同区域信息的传输，因为并非所有上下文信息都有助于检测

### Deformable Feature Learning

一个好的检测器应该对物体的非刚性变形具有鲁棒性。

* DPM：非深度学习方法。使用可变形编码方法由多个组成部分表示对象；
* DeepIDNet：设计deformable-aware pooling以编码不同对象类别之间的变形信息；
* DCN：替换普通卷积模块，自动学习辅助位置偏移量，以增强在特征图的常规采样位置的采样信息。

# Learning Strategy

## Training Stage

### Data Augmentation

水平翻转；旋转；随机裁剪；缩放；色彩抖动；

### Imbalance Sampling

proposal的样本不平衡可分为两类：类别不平衡、难度不平衡。

类别不平衡

* Fast\_RCNN对正负样本1:3采样，缺点是不能重复利用负样本信息；
* SSD中提出困难负样本采样策略，固定了正负样本比例，选择难度大（即分类损失高）的负样本训练模型。

难度不平衡

* focal loss：设置衰减参数，抑制了简单样本和负样本的信号；
* GHM：抑制简单样本的同时抑制异常值的影响；
* OHE：类似于SSD，不同之处在于仅考虑难度而不考虑类别，即正负样本比例不固定。

### Localization Refinement

精确的定位具有挑战性，因为检测通常集中在对象具有区分度的区域，而不一定是完整包含对象的区域。目前对定位的优化方法有：

* 生成更高质量的proposal
* R-CNN的L2回归和Fast R-CNN的smooth-L1回归
* Object detection via a multi-region and semantic segmentation-aware cnn model：迭代多次边界框回归
* LocNet：对每个边界框的分布进行了建模以优化预测结果
* Multipath Network：针对不同阈值IoU训练多个分类器，联合预测
* FitnessNMS：基于一组IoU上限导出一种新颖的边界框回归损失，该损失更好地匹配IoU最大化的目标。旨在寻找最优目标而不是合适的目标。
* Grid RCNN：用定位角点代替了bbox回归

### Cascade Learning

由粗到精的学习过程

* （CVPR2001）在人脸检测中提出级联学习的思想。在早期过滤大部分简单负样本，把困难样本送到下一个stage
* CRAFT：Fast\_R-CNN+联级学习
* layer-wise cascade classifier（CVPR2016）：将多个分类器放在不同层的特征图上，逐步过滤样本
* ReﬁneDet & Cascade RCNN：使用联级方法优化bbox位置
* Revisiting rcnn: On awakening the classiﬁcation power of faster rcnn：Faster RCNN和RCNN相互补充的级联检测系统

### Others

Adversarial Learning

典型的对抗学习就是GAN。在目标检测领域的应用有：

* Perceptual GAN：学习生成小物体的高分辨率特征表示；
* A-Fast-R-CNN：用对抗学习生成具有遮挡和变形的特征来训练分类器，作用类似于数据增量。

Training from Scratch

最早提出Training from Scratch的是DSOD，论文中指出：

微调的优点：

* 有很多好的模型是开源的，可以直接将他们用于目标检测；
* 微调可以快速得到最终的模型，需要的训练数据也相对较少。

微调的缺点：

* Limited structure design space，基于ImageNet 分类数据库的预训练模型大，参数太多。直接将它们用于目标检测，导致模型结构灵活性差，难以改变网络结构，计算量也大，限制其应用场合；
* Learning bias，因为分类和检测任务的损失函数和类别分布是不一样的，作者认为两者的搜索/优化空间是不一样的，所以在分类上预训练可能会导致在检测上的得到局部最优，而不是全局最优解；
* Domain mismatch，尽管微调可以减少不同目标类别分布的差异性，但是对于一些目标域为深度图、医学图像和ImageNet的差异性太大，微调效果不理想。

Training from Scratch方法：

* DSOD：紧密连接的网络结构可以极大地减少优化难度；
* gated recurrent feature pyramid：定义了一种递归特征金字塔结构，将空间和语义信息压缩到单个预测层中，进一步减少了参数数量，从而加快了收敛速度。

Kaiming He的Rethinking imagenet pre-training指出只要数据够大（10k），常规结构Training from Scratch不逊色于fine-tune，无关具体结构。这间接否定了一些这方面的工作。

Knowledge Distillation

通过teacher-student training scheme，用简单的模型学习复杂的模型。可以在几乎不损失精度的情况下提升速度。两篇论文：

* Revisiting rcnn: On awakening the classification power of faster rcnn
* Mimicking very efficient network for object detection

## Testing Stage

### Duplicate Removal

* NMS：非极大值抑制；
* soft-NMS：NMS在检测密集对象时容易发生误删除。不同于NMS直接删除非极大值预测，soft-NMS对相应框的置信度做连续衰减；
* Learning NMS：将NMS作为网络的优化目标，鼓励检测器对每个对象做一个精确检测。

### Model Acceleration

* one-stage速度上优于two-stage
* R-FCN：position sensitive ROI Pooling，共享计算；
* Light Head R-CNN：减少最终特征图的通道数；
* 速度更快的backbone：MobileNet采用深度可分离卷积结构；PVANet采用CReLU激活函数减少非线性计算；
* 离线加速：模型压缩量化、剪枝、TensorRT等。

### Others

测试时的图像变换，如图像金字塔、数据增强。缺点是时间耗费大。

# Applications and Detection Benchmarks

目标检测技术的应用除了人脸检测和行人检测，还有商标检测，视频检测，车辆检测，交通标志检测，骨架检测等

## Generic Detection Benchmarks

数据集：

* Pascal VOC2007：包含20种类别的用于目标检测的中型数据集。分别使用2501、2510和5011图像进行训练，验证和测试；
* Pascal VOC2012：用于目标检测的中型数据集，包含与Pascal VOC2007相同的20个类别。分别使用5717、5823和10991图像进行训练，验证和测试。VOC2012测试集的注释信息不公开；
* MSCOCO：包含80个类别的大规模数据集。分别使用118287、5000和40670图像进行训练，验证和测试。MSCOCO测试集的标注信息不公开；
* Open Images：包含1.9M张图像和600个类别的15M个对象。500个最常见的类别用于评估检测基准；
* LVIS：新数据集，包含164K张图像和1K多个类别；
* ImageNet：包含200个类别，规模巨大。对象尺度范围类似于VOC。一般用于分类而非检测。

## Face Detection Benchmarks

人脸检测和一般性检测的区别在于：

* 人脸检测具有更大的尺度变化性、遮挡和模糊的情况；
* 具有很强的结构联系信息。

代表工作：

* 传统方法：滑动窗口+手工特征；

深度学习方法的关注点在于多尺度鲁棒性和上下文信息。

* S3FD（ICCV2017）：侧重尺度不变性的单阶段检测器；
* SSH（ICCV2017）：尺度无关性的快速单阶段人脸检测；
* Finding tiny faces（CVPR2017）:级联RPN的变尺度小目标人脸检测器，速度慢；
* Face MegNet（WACV2018）：使用上下文，小目标转置卷积等方法得到大特征图，实现更好的定位；
* ……原文列举的方法多而杂，不一一列出。

数据集：

* WIDER FACE：共32203张图像，包含约40K张面部，覆盖各种尺度。包含三个子集：40％用于训练，10％用于验证，50％用于测试。测试集标注不公开。根据检测难度分为三个部分：简单，中等和困难；
* FDDB：2845张图像，5171张人脸。通常用于测试；
* PASCAL FACE：851张图像，1335张人脸。仅用作测试集。

## Pedestrian Detection Benchmarks

行人检测和一般性检测的区别在于：

* 行人检测具有近乎固定的宽高比（2:3），更大的尺度变化性；
* 行人检测更容易面临密集人群、遮挡、模糊等问题；
* 由于行人所处的复杂环境，容易产生更多的困难负样本（如交通灯、邮筒等）。

代表工作：

* 传统方法：滑动窗口+手工特征。早期工作集中在设计具有鲁棒性的特征描述子上，其中HOG取得了很好的效果；

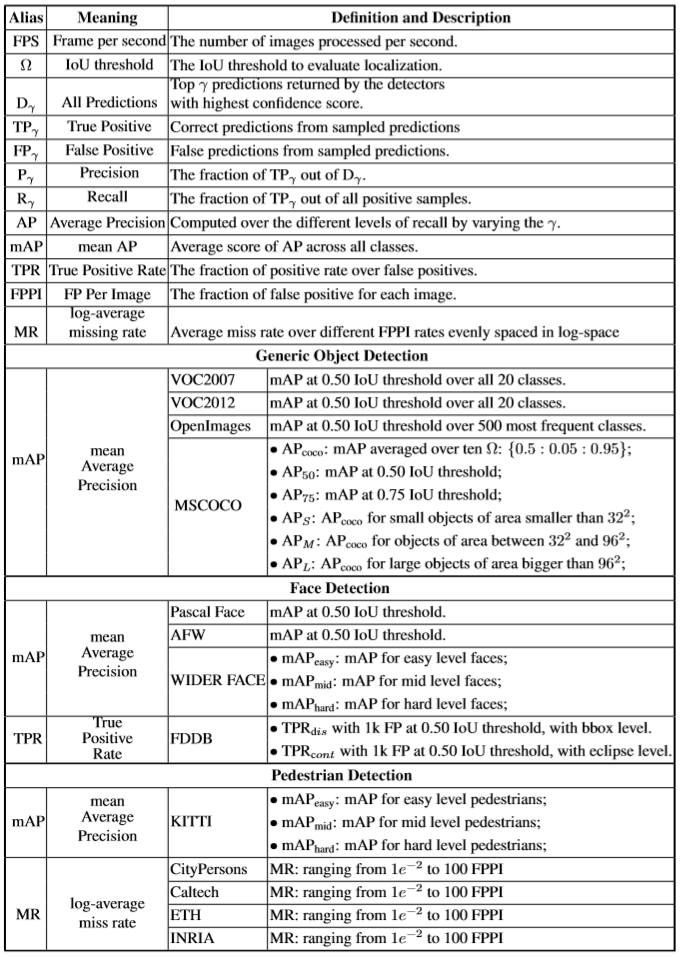
深度学习方法的关注点在于多尺度鲁棒性、样本不均衡问题，和遮挡问题。

* Repulsion Loss（CVPR2018）；
* OR-CNN（ECCV2018）；
* ……原文列举的方法多而杂，不一一列出。

数据集：

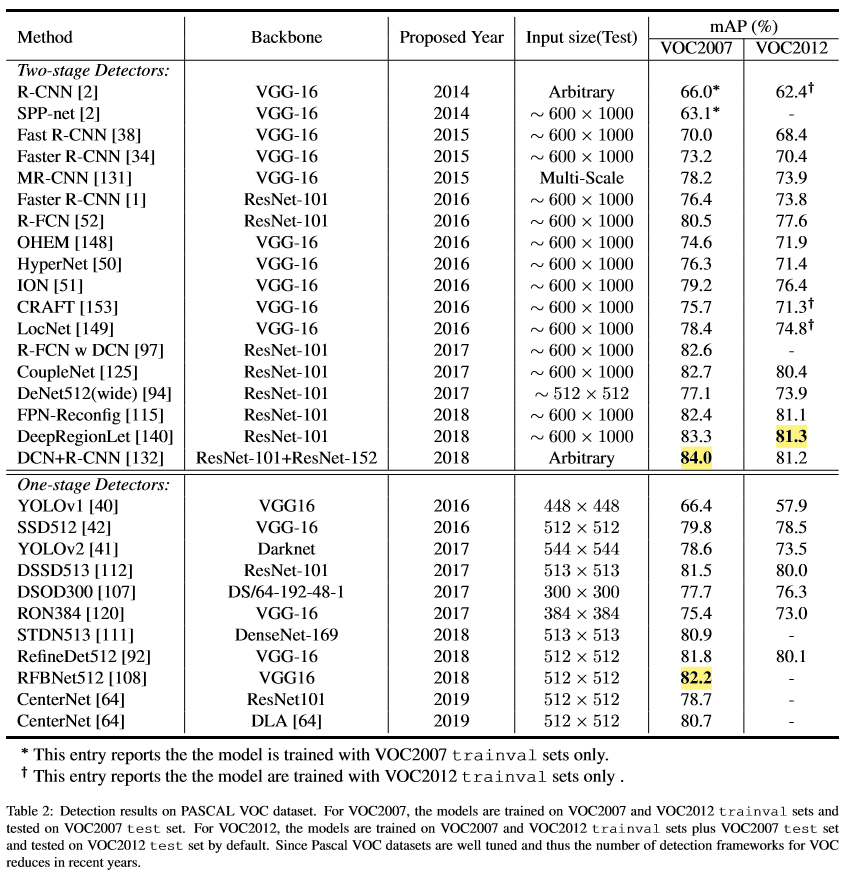
* CityPersons：在语义分割数据集CityScapes [243]之上的新的行人检测数据集。在德国的几个城市中捕获了5000张图像，共35000人和13000个被忽略的区域。同时提供完整bbox和可见部分bbox；
* Caltech：来自约10小时30Hz VGA视频。训练集42782帧，测试集4024帧；
* ETH：来自三个视频的1804帧，通常用作测试集；
* INRIA：包含高分辨率的行人图像，主要是从假日照片中收集的。包括2120张图像，训练集1832张，测试集288张。训练集中有614张正样本和1218张负样本；
* KITTI：包含7481张分辨率为1250\*375的标注图像和7518张测试图像。KITTI中的行人分为两类：行人和骑自行车者。KITTI包含三个评估指标：简单，中度和困难，bbox的最小值和最大遮程度不同。

## Evaluation Metrics

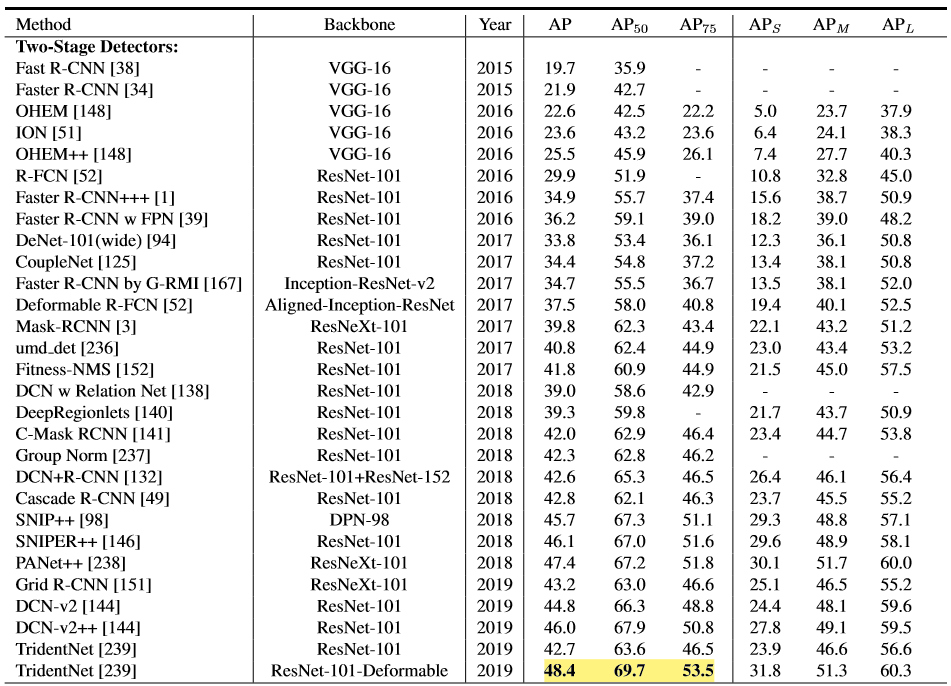


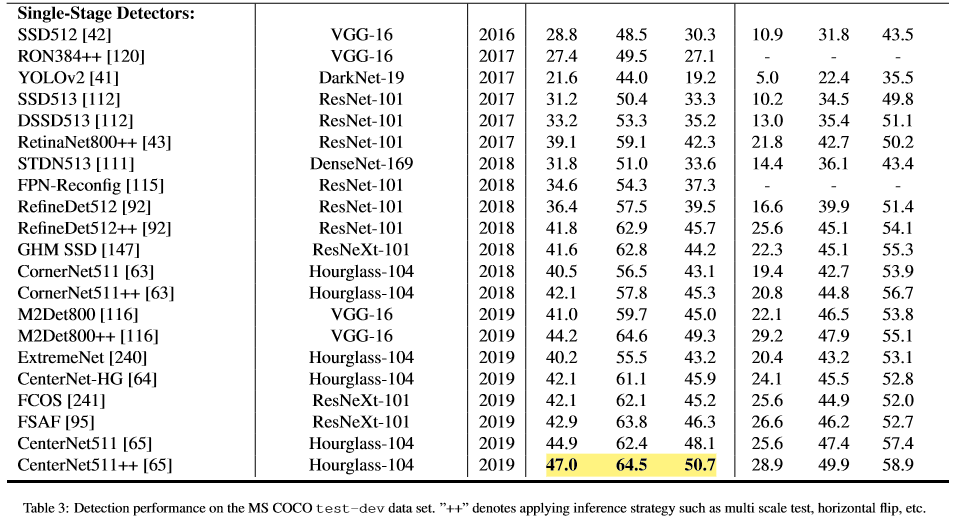
# SOTA for Generic Object Detection

PASCAL VOC：



MSCOCO：





# Future Directions

可拓展的提案生成策略

目前大多数检测器是anchor-base。anchor是手动设计的，很难匹配多尺度对象，并且基于IoU的匹配策略也是启发式的。最近anchor-free方法成为目标检测中非常热门的话题（如基于关键点的方法），但是仍然存在一些局限性（如高计算成本），还有很大的改进空间。设计有效的提案生成策略在未来仍然是非常重要的研究方向。

高效的上下文信息编码

上下文可能帮助或阻碍目标检测效果，如何正确、高效使用上下文信息值得研究。

基于自动机器学习的检测

为特定任务设计最佳的骨干架构可以显着改善结果，但也需要大量的工程工作。因此，直接在数据集上学习骨干架构是一个非常有趣且重要的研究方向。受AutoML在图像分类方面的开创性工作的启发，已经有相关工作通过AutoML解决检测问题，例如学习FPN结构和学习数据增强策略。这些工作成果相比baseline有显着改善。AutoML的缺点是需要庞大的计算资源，如何设计一个低计算量的框架也值得关注。

新的目标检测基准

当前MSCOCO是最常用的检测基准测试平台。但MSCOCO只有80个类别，相比于现实世界中更复杂的场景仍然太少。最新的LVIS包含164K幅图像，1K多个类别，2.2M个高质量的实例分割蒙板。此外，LVIS模拟了现实世界的Low-shot场景（存在大量类别，但每个类别的数据不多）。

Low-shot检测

使用有限标注数据训练的检测器称为Low-shot检测。对于分类任务，Low-shot学习已得到积极研究，但只有少数研究集中在检测任务上。MSPLD用半监督学习解决了Low-shot检测问题；RepMet采用深度度量学习（DML），同时学习了特征空间和训练集类别的数据分布，但RepMet仅在具有相似概念（动物）的数据集上进行了测试；LSTD采用迁移学习，使用知识正则化从大型外部标注数据集学习获取知识，但存在过拟合问题。Low-shot检测仍有很大改善空间。

用于检测的骨干架构

将在大规模数据集上预训练的分类模型的权重引入检测问题已成为一种范式。但是分类任务和检测任务之间仍然存在冲突，因此这并不是一个最佳的解决方案。目前大多数检测算法都是基于分类骨干架构的，只有少数尝试不同的选择（如Hourglass）。因此，开发针对检测任务的骨干架构也是未来的重要研究方向。

其他

如large batch learning和incremental learning。batch大小是DCNN训练中的关键因素，但在检测任务中尚未进行充分研究。对于增量学习，在没有初始训练数据的情况下适应新任务时，存在灾难性遗忘问题。这些开放和基础性的问题也值关注。另外，这篇文章提出了其他一些值得关注的方向（<https://www.zhihu.com/question/280703314>）。