面向目标检测的深度学习研究进展（CVPR 2019）

摘要

分析现有的目标检测框架，分为三个部分：

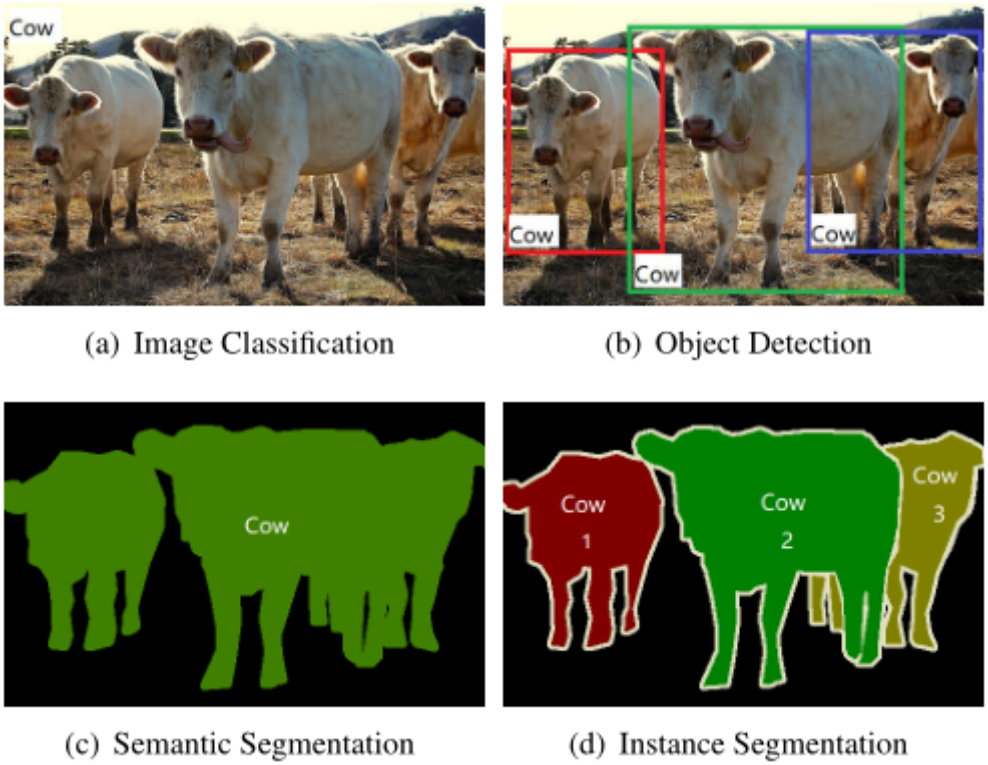
- 检测组件
- 学习策略
- 应用和基准

以及分析影响的因素，包括检测器架构、特征学习、候选生成（proposal generation）和抽样策略等。最后讨论了未来目标检测的方向。

1、介绍

计算机视觉基础问题：

- 图像分类：辨别语义类别
- 目标检测：预测类别 + 定位坐标（边界框）
- 实例分割：目标检测 + 语义分割（像素）
- 语义分割：为每个像素预测类别（不区分同类别的多个对象）



(a) Image Classification (b) Object Detection

(c) Semantic Segmentation (d) Instance Segmentation

传统目标检测流程：

- 提议生成：搜索（利用多尺度滑动窗口）图像中可能包含对象的位置，即ROI
- 特征向量提取：从滑动窗口中获得固定长度的特征向量（由低级视觉描述符编码，如SIFT、HOG和SURF等）
- 区域分类：利用SVM、级联学习和AdaBoost算法等

关注设计特征描述符，使用 Pascal VOC（用于基准对象检测的公开数据集）测试。

传统检测器的局限性：

- 提案生成阶段生成了许多多余的方案，导致误分类。此外滑动窗口是手工和启发式的，不能很好地匹配目标
- 特征描述符是基于低级视觉线索手工设计的，很难在复杂的上下文中捕捉代表性的语义信息
- 每阶段都是独立设计和优化的，不能获得全局最优解

最初深度学习的局限性：

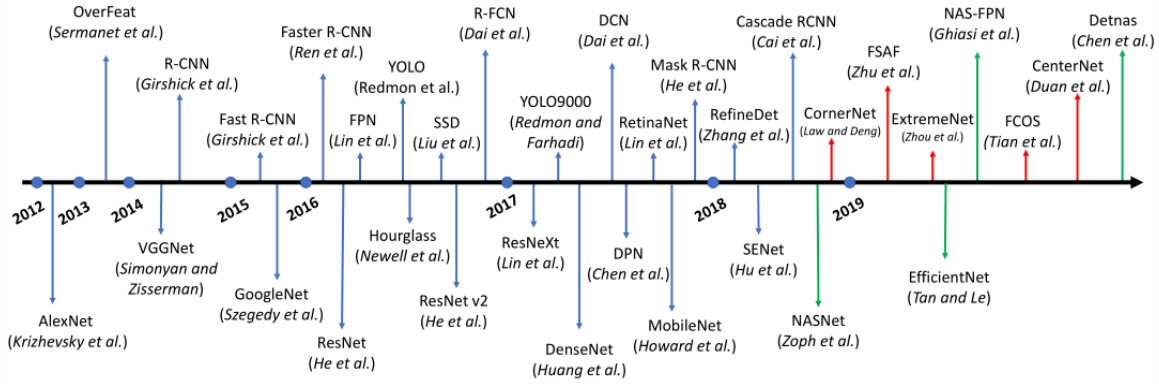
- 缺乏大量标注训练数据，导致过拟合
- 计算资源有限
- 与SVM相比，理论支持较弱

深度学习区别于传统算法：

- 从原始像素中生成高级别语义信息的分层特征
- 从训练数据中自动学习
- 复杂环境中有较好的描述表达能力

深度学习的目标检测框架主要分为两大类：

- 两阶段检测器：首先使用候选生成器生成一组候选集并提取特征，然后使用区域分类器预测该候选区域的类别。如基于区域的 CNN（R-CNN）及其变体
- 一阶段检测器：直接对特征矩阵上每个位置的对象进行分类预测，不需要级联的区域分类步骤。如 YOLO 及其变体



本文主要探讨内容如下图所示：

<table>
<tr><th colspan="5">Object Detection</th></tr>
<tr><th colspan="3">Detection Components</th><th>Learning Strategy</th><th>Applications & Benchmarks</th></tr>
<tr><th>Detection Settings</th><th>Detection Paradigms</th><th>Backbone Architecture</th><th>Training Stage</th><th>Applications</th></tr>
<tr><td rowspan="2">Bounding Box</td><td rowspan="2">Two-Stage Detectors</td><td>VGG16,ResNet,DenseNet</td><td>Data Augmentation</td><td rowspan="2">Face Detection</td></tr>
<tr><td rowspan="2">MobileNet, ResNeXt</td><td>Imbalance Sampling</td></tr>
<tr><td rowspan="3">Pixel Mask</td><td rowspan="3">One-Stage Detectors</td><td>Localization Refinement</td><td>Pedestrian Detection</td></tr>
<tr><td>DetNet, Hourglass Net</td><td>Cascade Learning</td><td rowspan="2">Others</td></tr>
<tr><td colspan="2">Others</td></tr>
<tr><th colspan="2">Proposal Generation</th><th>Feature Representation</th><th>Testing Stage</th><th>Public Benchmarks</th></tr>
<tr><td colspan="2">Traditional Computer Vision Methods</td><td>Multi-scale Feature Learning</td><td rowspan="2">Duplicate Removal</td><td rowspan="2">MSCOCO, Pascal VOC, Open Images</td></tr>
<tr><td colspan="2">Anchor-based Methods</td><td>Region Feature Encoding</td></tr>
<tr><td colspan="2">Keypoint-based Methods</td><td>Contextual Reasoning</td><td rowspan="2">Model Acceleration</td><td rowspan="2">FDDB, WIDER FACE</td></tr>
<tr><td colspan="2">Other Methods</td><td>Deformable Feature Learning</td></tr>
<tr><td colspan="3"></td><td>Others</td><td>KITTI, ETH, CityPersons</td></tr>
</table>

本文内容划分：

- 目标检测的标准问题集 -> 第2部分
- 检测组件的细节 -> 第3部分
- 学习策略 -> 第4部分
- 对真实世界应用的检测算法和基准 -> 第5、6部分
- 通用的最新检测结果 -> 第7部分
- 总结和讨论未来方向 -> 第8部分

2、问题设置

> 目标检测涉及识别（如对象分类）和定位（位置回归）两个问题。
>
> 目标检测器需要通过精确定位和对每个目标实例的正确分类来区分图像中特定目标类别的目标和背景。预测边界框或像素遮罩来定位这些目标对象实例。

定位评估标准：

$$\mathrm{IoU}(b_{\mathrm{pred}}, b_{\mathrm{gt}}) = \frac{\mathrm{Area}(b_{\mathrm{pred}} \bigcap b_{\mathrm{gt}})}{\mathrm{Area}(b_{\mathrm{pred}} \bigcup b_{\mathrm{gt}})} \tag{4}$$

对象检测评估标准：

$$\mathrm{Prediction} = \begin{cases} \mathrm{Positive} & c_{\mathrm{pred}} = c_{\mathrm{gt}} \text{ and } \mathrm{IoU}(b_{\mathrm{pred}}, b_{\mathrm{gt}}) > \Omega \\ \mathrm{Negative} & \text{otherwise} \end{cases} \tag{5}$$

对于一般的物体检测问题，使用某类上的平均精度（mAP）来评估；对于现实世界的场景如行人检测等，使用不同的评估指标，第5节探讨。

另外，推理速度（如每秒帧数FPS等）也是目标检测算法的一个重要指标，其中达到 20 FPS 可以被认为是实时检测器。

3、检测组件

> 首先介绍两个检测概念：bbox-level 和 mask-level 算法；
>
> 接着介绍两个目标检测范例：两阶段检测器和一阶段检测器

3.1 检测设置

> 目标检测有两种设置：普通目标检测算法（bbox-level 定位）和实例分割（pixel-level 或 mask-level 定位）

普通目标检测算法是传统的检测设置，其目标是通过矩形框来定位物体。只需要 bbox 注释，使用 IoU 进行评估。

实例分割是基于传统的检测设置，使用像素级来分割对象。使用在掩码预测上的 IoU 进行评估。

3.2 检测范例

> 最新检测器分为两大类：两阶段检测器和一阶段检测器
>
> 两阶段检测器：首先生成一组稀疏的候选特征；然后由DCNN编码预测。特点是准确率高，推理时间长
>
> 单阶段检测器：通常将图像上所有位置视为潜在对象，并尝试将每个ROI分类为背景或目标对象。特点是推理时间短，准确率低

3.2.1 两阶段检测器

> 两阶段检测器使用步骤：
>
> 1、生成候选。检测器尝试识别图像中可能是对象的区域，其思想是提出具有高召回率的区域，使得图像中所有对象属于这些区域中的至少一个；
>
> 2、为候选生成预测。为每个区域预测为背景或对象，并细化原始候选定位。

1、R-CNN（Region-based CNN） 在2014年由 Girshick 等人提出，可分为三个部分：候选生成、特征提取和区域分类。

原理：对于每张图像，R-CNN生成一组候选（约2000个），生成候选的方式是选择性搜索，即拒绝容易被识别为背景的区域；然后每个候选被裁减为固定大小的区域并被送入DCNN中编码为特征向量（约4096维），之后是1对多的SVM分类器；最后使用提取的特征作为输入来训练边界框回归器。

对比：相对于传统提取特征，DCNN可以通过不同层来捕获不同尺度的信息，从而产生鲁棒的、有辨识度的特征

缺点：

- 每个候选的特征由DCNN分别提取（即计算不共享），导致大量重复计算
- 3个步骤是独立的，整个检测框架无法进行端到端的优化，难以获得全局最优解
- 选择性搜索依赖底层视觉信息，难以在复杂环境中生成高质量候选
- 无法使用GPU加速

2、SPP-net（Spatial Pyramid Pooling） 是由 He 等人受空间金字塔匹配（SPM）启发提出的加速 R-CNN 并学习更多区分特征的网络。

原理：无需裁减区域和将特征送入CNN，而是直接使用DCNN计算整张图像的特征图，并通过空间金字塔池化（SPP）在特征图中提取固定长度的特征向量。其中SPP将图中图划分为多个N组成的 N x N 网络，并对每个N中每个单元进行池化、连接来获取不同比例区域信息的特征向量。最后将特征送入SVM进行分类和矩阵框回归。

对比：相对于R-CNN，该网络也能够提取不同比例的图像特征，但是由于没有裁减而避免了信息丢失和不必要的几何失真；有更快的推理速度

缺点：

- 多阶段，因此不能被端到端地优化
- 需要额外的高速缓存来存储提取的特征
- SSP层没有将梯度反向传播到卷积核，因此SSP层之前的所有参数都被冻结

3、Fast R-CNN 是由Girshick等人提出的多任务快速学习检测器，克服了SPP-net的两个局限性。

原理：同样从整张图像计算特征图并提取固定长度的区域特征。不同的是Fast R-CNN使用ROI池化层提取区域特征，即使用单个尺度对提取固定数量的特征并反向传播到卷积核上。之后依次经过同级的分类层和回归层被传入一系列的全连接层。其中分类层用 softmax 生成 C+1 类（含背景类），而回归层编码4个实数参数来细化边界框。

对比：相对于SPP-net，特征提取、区域分类和边界框回归等都可以进行端到端优化，且无需额外的缓存空间来存储特征。

缺点：

- 生成候选步骤仍然依赖于传统方法（如选择性搜索或边界框等），这些方法基于低级视觉线索，不能以数据驱动的方式学习

4、Faster R-CNN 依赖于新的候选生成器，即区域候选网络（RPN），可以通过监督学习方法来学习。

原理：RPN是一个全卷积网络，可以使用任意大小的图像在特征矩阵的每个位置生成一组对象候选，即使用 N x N 滑动窗口在特征矩阵上滑动，并为每个位置生成特征向量。之后特征向量被送入对象分类层和边界框回归层，最后被馈送到最终层进行对象分类和边界框定位。

对比：相对于Fast R-CNN，RPN能够以数据驱动的方式生成候选。在特征矩阵上提取区域特征时不同区域之间共享特征提取计算。

缺点：

- 区域分类步骤中没有共享计算，每个特征向量仍然分别需要通过一系列FC层，这种额外的计算可能非常大，而简单移除了FC层会导致检测性能急剧下降，因为深层网络会减少候选的空间信息
- 使用单个深层特征矩阵做最后的预测，对于不同尺度的对象检测变得困难，很难发现小的对象。

5、R-FCN（Region-based Fully Convolutional Networks） 由Dai等人提出的基于区域的全卷积网络，该网络在区域分类步骤中共享计算成本。

原理： FCN生成一个位置敏感分数矩阵，它对不同类别的相对位置信息进行编码，并使用位置敏感ROI池化层（PSROI Pooling）通过对目标区域的每个相对位置进行编码来提取空间感知区域特征。

对比：特征向量保持了空间信息，因此比没有基于区域的全连接层操作的 Faster R-CNN 效果更好。

6、FPN（Feature Pyramid Networks） 由Lin等人利用“深层特征语义强空间弱，而浅层特征语义弱空间强”这一特性提出，该网络将深层特征与浅层特征相结合，从而能够在不同比例的特征矩阵中进行目标检测。适用于检测多尺度对象，可用于其他领域，如视频检测和人体姿态识别等。

原理：主要思想是利用来自更深层的丰富语义信息来增强空间强的浅层特征。

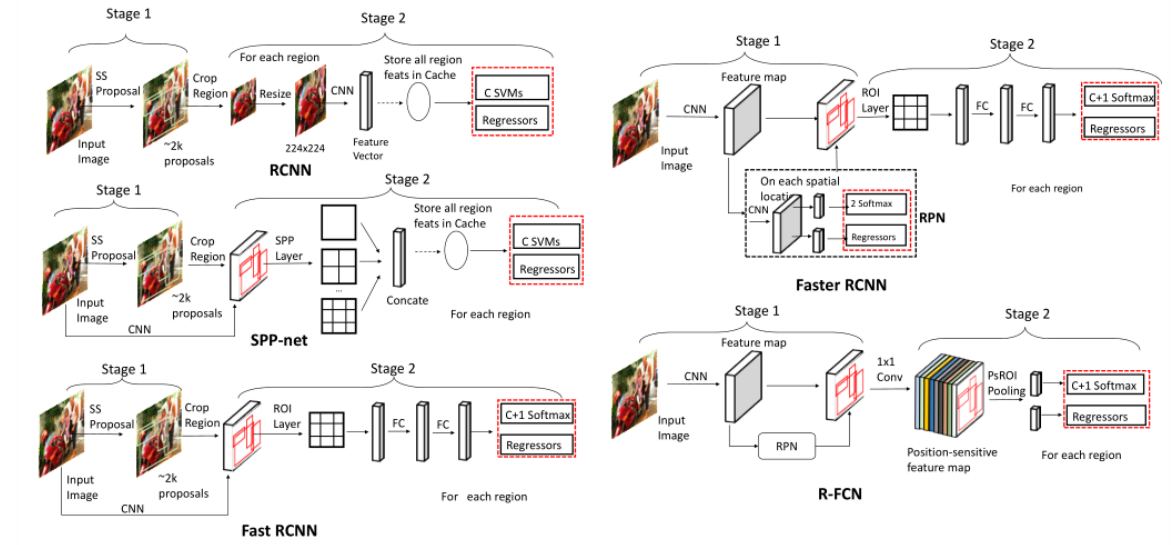


Figure 4: Overview of different two-stage detection frameworks for generic object detection. Red dotted rectangles denote the outputs that define the loss functions.

3.2.2 单阶段检测器

> 通常将图像上的所有位置视为潜在对象，并试图将每个ROI分类为背景或目标对象

1、OverFeat 由Sermanet等人提出， 通过将casting神经网络分类器转换为全卷积目标检测器来检测目标。

原理：目标检测可以看作是一个“多区域分类”问题，因此该算法将最后的FC层看做 1x1 的卷积层以允许任意的输入，从而将原始分类器扩展为检测器。分类网络在输入的每个区域上输出预测网络，以指示对象的存在，在识别出目标后，基于分类器相同的DCNN学习边界框回归器来细化预测区域。为了检测多尺度图像，输入图像被调整为多个尺度输入到网络中，最后跨越所有尺度的预测被合并在一起。

对比：相对于R-CNN，该模型通过使用卷积层来共享重叠区域的计算，并且只需要通过网络的一次前向传播，提高了推理速度

缺点：分类器和回归器的训练是分开的，没有共同优化

2、YOLO（You Only Look Once） 由Redmon等人提出的实时检测器。

原理：YOLO认为目标检测是一个回归问题，在空间上将整个图像分成固定数量的网格（如 7x7 网络），每个网格都是一个候选，用来检测一个或多个对象。最初实现中每个网格被认为包含两个对象的中心，对每个网格检测是否有对象、边界框坐标和大小以及对象的类别。

优点：整个框架是单一的网络，省略了可以端到端方式优化的候选生成步骤。

缺点：

- 在给定位置只能探测到两个物体，很难检测到小物体和拥挤的物体
- 仅适用最后的特征矩阵进行预测，不适用于预测多个比例和纵横比的物体

3、SSD（Single-Shot Mulibox Detector） 由Liu等人提出的单次多盒检测器。

原理： SSD将图像划分为多个网格，对于每个网格使用一组具有多种比例和纵横比的锚点，来离散化边界框的输出空间。其中每个锚由回归器学习的 4-value 偏移来细化，并由分类器分配 C+1 个类别。此外，SSD在多个特征矩阵上预测对象，并且每个特征矩阵负责根据其感受野来检测特定比例的对象。通过端到端训练，利用所有的预测矩阵的定位损失和分类损失的加权和来优化整个网络，最后合并来自不同特征矩阵的所有检测结果做最终预测。

不足：

- 为了避免大量负样本在训练梯度占比较大，使用硬否定挖掘来训练检测器
- 使用密集数据增强来提高检测精度

4、RetinaNet 由Lin等人提出的一种以更灵活的方式解决了类不平衡问题的检测器。

原理： RetinaNet 使用聚焦损失来抑制建档负样本梯度，而不是简单丢弃。使用特征金字塔网络在不同级别的特征矩阵上检测多尺度对象

5、YOLO v2 由Redmon提出的YOLO改进版检测器，保持推理速度的同时提高了检测性能。

原理： 采用了从 ImageNet 图像预训练的模型，学习的权重对捕获细粒度信息更敏感。受SSD启发使用通过训练数据的 k-均值聚类来设定锚优先级。最后结合批标准化层（BN）和多尺度训练技术获得了较好的检测结果。

6、CornerNet 由Law和Deng提出的一种无锚框架，将物体检测为一对角。

原理： 在特征矩阵的每个位置，预测了类热图、对嵌入和角偏移。其中类热图计算拐角的概率，角偏移用于回归拐角的位置，而对嵌入用于对属于同一对象的一对角进行分组。

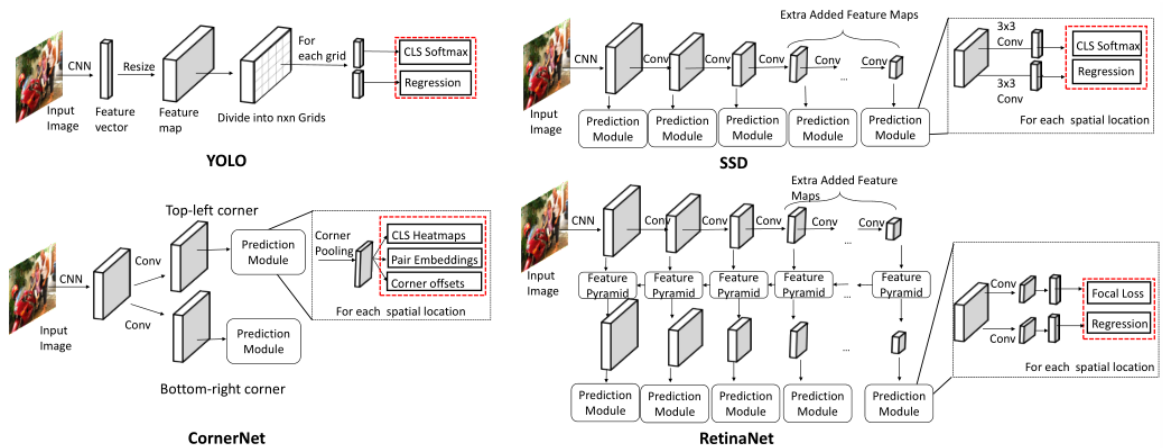


Figure 5: Overview of different one-stage detection frameworks for generic object detection. Red rectangles denotes the outputs that define the objective functions.

3.3 主干架构

> 采用大规模图像分类问题的预训练模型的卷积权重可以为训练检测器提供更丰富的语义信息，提高检测性能。

3.3.1 一个CNN的基本架构

> DCNN通常由一系列卷积层、池化层、非线性激活层和全连接层组成

卷积层： 通过卷积层生成的特征矩阵可以被视为多通道图像，每个通道表示关于图像的不同信息。特征矩阵中的每个像素（即神经元）都与前一个图中相邻神经元的一部分相连，这部分就是感受野。（后可使用非线性激活层）

池化层： 用于汇总感受野内的信息，以扩大感受野并降低计算成本

3.3.2 用于目标检测的CNN主干网

> 概述用于目标检测的CNN主干网，如 VGG16、ResNet、ResNeXt 和 Hourglass 等
>
> **注：**
>
> 1、分类需要大的感受野并要保持空间一致性，因此许多下采样操作（如池化层）被应用于降低特征图分辨率。而在检测中，需要高分辨率的空间信息来正确定位物体。
>
> 2、分类在单个特征图上进行预测，而检测需要具有多个表示的特征图来检测多个比例的对象

1、VGG16 基于AlexNet，由5组卷积层和3个全连接层组成，其中前2组有2个卷积层，后3组有3个卷积层，且每组之间用最大池化层来降低空间维度。

2、ResNet 引入快捷连接降低了优化难度，使得网络深度可以进一步增加。

3、DenseNet 认为相加不能保留浅层信息，通过将输入与剩余输出连接起来，而不是逐元素相加，保留了分层特征。 *（下图中的∘表示连接）*

$$x_{l+1} = x_l \circ f_{l+1}(x_l, \theta) \qquad (7)$$

4、DPN（Dual Path Network） 综合上述两个网络，将通道 x 分为两部分，一部分用于密集连接计算，一部分用于元素求和，最后是两个分支的串联输出。

$$x_{l+1} = (x_l^r + f_{l+1}^r(x_l^r, \theta^r)) \circ (x_l^d \circ f_{l+1}^d(x_l^d, \theta^d)) \qquad (8)$$

5、ResNeXt 基于ResNet，采用稀疏连接特征矩阵通道的组卷积层来降低计算成本，通过增加组的数量来保持计算成本与原ResNet一致，ResNeXt 从训练数据中捕获更丰富的语义特征表示，从而提高主干网的准确性。

6、MobileNet 将坐标设置为每个特征矩阵的通道数，显著降低了计算成本和参数数量，而不会降低分类精度。该模型是专为在移动平台而设计的。

7、GoogleNet 使用 inception 模型（在宽度上扩展），该模块在给定层的相同特征矩阵上应用不同比例的卷积核。通过这种方式，它捕获了多尺度特征，并汇总为一个输出特征矩阵。

8、DetNet 专为检测而设计（解决上述 [注] 中说明的问题），保留了用于预测的高分辨率特征图，具有扩大的卷积以增加感受野。在多尺度特征图上检测对象，提供了更丰富的信息。

9、Hourglass Network 不是专门为图像分类设计的， 最早用于人体姿态识别，是一个全卷积结构网络。该网络首先通过一系列卷积核池化对输入图像进行下采样，然后通过反卷积运算对特征图进行上采样，为了避免下采样阶段信息丢失，在上下采样之间使用了跳跃连接。该网络可以捕捉局部和全局信息，因此适合物体检测且广泛应用于检测框架。

3.4 候选生成

> 候选生成器生成一组矩阵边界框（可能包含对象），然后被用于分类和定位细化。
>
> 两阶段检测器和单阶段检测器都生成候选，**主要区别**是两阶段检测器生成的候选只包含前景或背景的稀疏候选集，而单阶段检测器将图像的每个区域都视为潜在候选，并相应的评估每个位置处的潜在对象的类别和边界框坐标，

候选生成方法分为四类：

- 传统计算机视觉方法
- 基于锚点的监督学习方法
- 基于关键点的方法
- 其他方法

3.4.1 传统计算机视觉方法

> 使用传统的基于低级线索的计算机视觉方法，如边缘、角和颜色等，在图像中生成候选
>
> **方法**：
>
> - 计算候选框的目标分数
> - 合并来自原始图像的超像素
> - 生成多个前景和背景片段

目标分数：基于方法预测每个候选框包含目标的概率分数，如视觉线索、级联学习方法等

超像素合并：基于合并从分割结果中生成超像素。如选择性搜索等

种子分割：从多个种子区域开始，为每个种子生成前景和背景分割，如CPMC等

特点：

- 简单，且可以生成具有高召回率的候选
- 基于颜色或边缘等低级视觉线索，无法与整体共同联合优化，因此无法利用大规模数据集的力量来改进学习

3.4.2 基于锚方法

> 基于预定义的锚点生成候选

区域候选网络（RPN）：基于深度卷积特征图的监督方式生成候选。使用 3x3 卷积滤波器在特征图上滑动，对于每个位置，考虑不同大小和纵横比的 k 个锚（或边界框的初始估计值），这些尺寸和比例允许在整个图像中以不同的比例匹配对象，找到最匹配的锚框。每个锚匹配256维特征向量，并将其输入并列的分类层和回归层。其中分类分支负责建模对象分数，而回归分支编码4个实值，以根据原始锚点估计来细化边界框的位置。

SSD： 利用多尺度锚来匹配对象，与 RPN 的主要区别在于，SSD 为每个锚点候选分配分类概率，而 RPN 首先评估锚点候选类型（前景或背景），在下一阶段执行类别分类

特定：手工设计，可能不是最佳的，不同的数据集需要不同的锚设计策略

3.4.3 基于关键点方法

> 基于关键点的方法分为两类：
>
> - 基于角点的方法：通过合并从特征图中学习的角点对来预测边界框
> - 基于中心的方法：在特征图的每个位置上预测作为对象中心的概率，并且高度和宽度被直接回归而没有任何锚定先验

3.4.4 其他方法

AZnet：采用一种自适应的搜索策略，将计算资源定向到可能包含对象的子区域，对于每个区域，该算法预测了两个值：缩放指示器和邻接分数。其中缩放指示器确定是否进一步划分该区域（起点是整个图像），该区域可能包含较小的对象；而邻接分数表示其对象性。

3.5 特征表示学习

> 目标对象位于复杂的环境中，在比例和纵横比上有很大的差异，需要训练对象的鲁棒和有区别的特征嵌入来获得良好的检测性能。
>
> **特征学习策略：**
>
> - 多尺度特征学习
> - 上下文推理
> - 可变形特征学习

3.5.1 多尺度特征学习

> **深度卷积网络学习不同层中的分层信息，这些分层特征捕获不同的尺度信息：**
>
> - 具有空间丰富信息的浅层特征具有更高的分辨率和更小的感受野，因此更适合检测小物体；
> - 而深层的语义丰富特征对光照、平移和具有更大的感受野更鲁棒，更适合检测大物体
>
> **解决多尺度特征学习问题的主要范式有四种：**
>
> - 图像金字塔
> - 预测金字塔
> - 集成特征
> - 特征金字塔

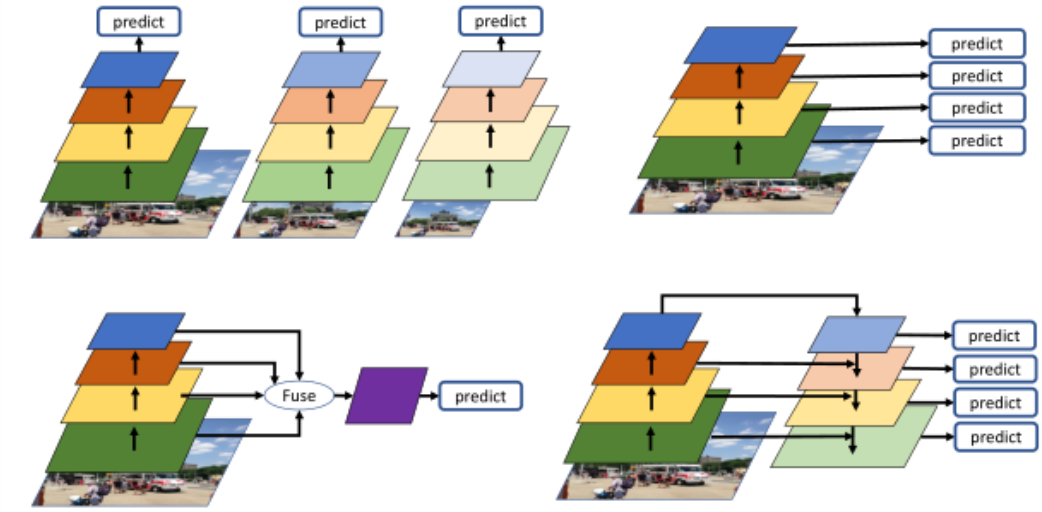


Figure 7: Four paradigms for multi-scale feature learning. Top Left: **Image Pyramid**, which learns multiple detectors from different scale images; Top Right: **Prediction Pyramid**, which predicts on multiple feature maps; Bottom Left: **Integrated Features**, which predicts on single feature map generated from multiple features; Bottom Right: **Feature Pyramid** which combines the structure of **Prediction Pyramid** and **Integrated Features**.

1、图像金字塔（IP）： 将输入图像调整到不同的比例并训练多个检测器，每个检测器负责一定范围的比例，最后检测结果被合并。

2、集成特征（IF）： 组合多个图层中的特征来构建单个特征图，并根据新构建的特征图进行最终预测。其中需要跳跃连接和特征归一化。

3、预测金字塔（PP）： 将多层的粗、细特征结合在一起。

4、特征金字塔（FP）： 为了结合集成特征和预测金字塔的优点，该网络以自上而下的方式集成具有横向连接的不同尺度特征以构建一组尺度不变的特征图，并且在这些特征金字塔上学习多个尺度相关的分类器。这些自上而下的横向特征通过元素求和或连接结合在一起，通过小的卷积来减少维度。

3.5.2 区域特征编码

> 对于两阶段检测器，区域特征编码是将候选中的特征提取为固定长度特征向量的关键步骤

3.5.3 上下文推理

> 目标往往出现在特定的环境中，有时也与其他对象共存。有效使用上下文信息有助于提高检测性能，尤其是检测线索不足的对象（小对象或遮挡等）。
>
> 存在两方面：
>
> - 全局上下文推理：指整个图像中的上下文学习
> - 区域上下文推理：对周围区域的上下文信息进行编码，并学习对象与其周围区域之间的交互

3.5.4 可变形特征学习

> 一个好的检测器应该对物体的非刚性变形具有鲁棒性

4、学习策略

4.1 训练阶段

> 讨论数据增强、不平衡采样、级联学习、定位细化等学习策略

4.1.1 数据增强

- 水平翻转（Faster R-CNN）
- 旋转、随机裁减、扩展和颜色抖动（单阶段检测器）

4.1.2 不平衡采样

> 大多数被估计为候选的感兴趣区域实际上只是背景图像，由此引发两个问题：
>
> - 类不平衡：大部分候选属于背景引起。
> - 难度不平衡：由于类不平衡，大多数背景更容易分类，而对象更难分类

类不平衡： 拒绝大部分阴性样本，保留2000个候选（两阶段检测器如 R-CNN 和 Fast R-CNN）；随机抽样可以解决类不平衡问题，但不能充分利用来自负候选的信息；一些负候选可能包含关于图像的丰富上下文信息，而一些硬候选可以帮助提高检测精度

难度不平衡： 基于精心设计的损失函数，如抑制简单样本的梯度信号，在训练过程中更多地集中在硬候选上。

4.1.3 定位细化

> 目标检测器必须为每个目标提供一个严格的定位预测（bbox 或 mask）

- 学习 L-2 辅助边界框回归器以细化定位（R-CNN）
- 通过端到端训练学习 L1 回归器（Fast R-CNN）
- 优化损失函数

4.1.4 级联学习

> 级联学习是一种由粗到精的学习策略，它从给定分类器的输出中收集信息，以级联方式构建更强的分类器。

4.1.5 其他

> 有四类：
>
> - 对抗性学习
> - 从零开始的训练
> - 知识蒸馏

1、对抗性学习： 如GAN，其中一个生成器和一个鉴别器进行竞争。

2、从零开始学习： 不使用预训练模型来减少训练偏差，可通过紧密连接的网络结构进行深度监控可以显著降低优化难度。

3、知识蒸馏： 是一种训练策略，通过师生训练方案将模型集合中的知识提炼为单个模型。

4.2 测试阶段

> 目标检测算法产生一组密集的预测，因此这些预测由于大量重复而不能直接用于评估。此外，还需要一些其他的学习策略来进一步提高检测精度。
>
> 重点包括重复删除、模型加速等。

4.2.1 重复删除

非极大值抑制（NMS）是目标检测的一个组成部分，用于消除重复的假阳性预测。候选框设定为M，然后计算与其他 M 框的 IoU 值，如果值大于预定义的阈值，这些框将被移除（置信度设置为0）。

$$\text{Score}_B = \begin{cases} \text{Score}_B & \text{IoU}(B, M) < \Omega_{\text{test}} \\ 0 & \text{IoU}(B, M) \geq \Omega_{\text{test}} \end{cases} \tag{13}$$

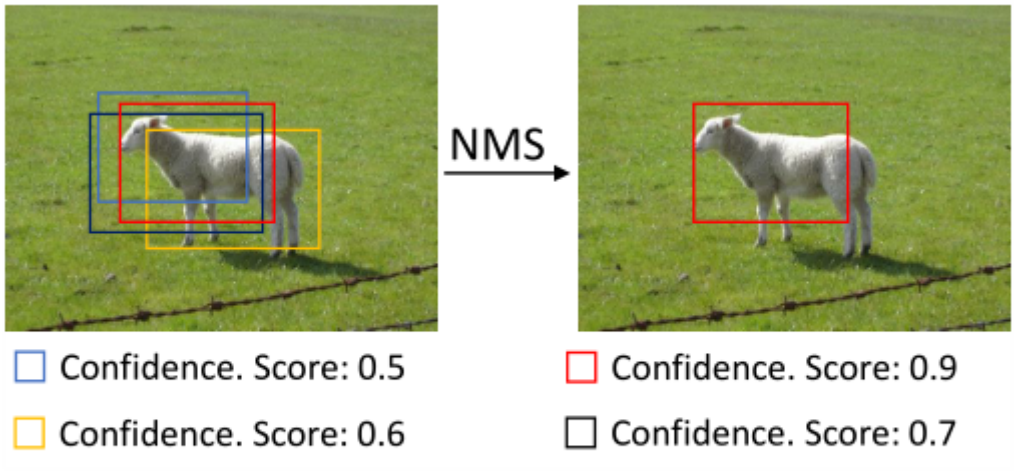


Figure 10: Duplicate predictions are eliminated by NMS operation. The most-confident box is kept, and all other boxes surrounding it will be removed.

改进的NMS为了不误删候选框（避免了消除聚集对象的预测），则不直接将其置信度设置为0，采用如下公式进行处理：

$$\text{Score}_B = \begin{cases} \text{Score}_B & \text{IoU}(B, M) < \Omega_{\text{test}} \\ F(\text{IoU}(B, M)) & \text{IoU}(B, M) \geq \Omega_{\text{test}} \end{cases} \tag{14}$$

4.2.2 模型加速

> 加快检测速度的一个简单方法是用更高效的主干替换检测主干，如 MobileNet 是一个高效的CNN模型，具有深度方向的卷积层。
>
> 另一个方法是离线优化模型，如对学习的模型进行模型压缩和量化

4.2.3 其他

> 检测阶段的其他学习策略主要包括对输入图像的变换，以提高检测精度

5、应用

5.1 人脸检测

> 与一般检测之间的区别：
>
> - 人脸检测中对象的尺度范围比一般的检测中的对象大得多，且遮挡和模糊的情况比较常见
> - 人脸对象包含较强的结构信息，只有一个目标类，可以利用先验改进人脸检测

5.2 行人检测

> 特点：
>
> - 行人对象是结构良好的对象，具有几乎固定的纵横比（约1:5），但它们也位于大范围的比例上
> - 拥挤、遮挡和模糊是常见的
> - 硬阴性样本较多，如红绿灯、邮箱等

5.3 其他

> 如 Logo 检测、视频对象检测、车辆检测、交通标志检测和骨架检测等

6、 检测基准

> 展示一些通用对象检测、人脸检测和行人检测的常见基准，首先为每个任务展示一些广泛使用的数据集，然后介绍评估指标

6.1 通用检测基准

1、Pascal VOC 2007： 是一个中等规模的对象检测数据集，有20个类别。有三个图像拆分，分别用2501、2510和5011进行训练、验证和测试。

2、Pascal VOC 2012： 是一个中等规模的目标检测数据集，与上一个数据集共享20个类别。有三个图像拆分，分别用5717、5823和10991进行训练、验证和测试，其中测试集的标注信息不可用。

3、MSCOCO： 是一个包含80个类别的大规模数据集，分别用118287、5000和40670张图像划分训练、验证和测试集，其中测试集的注释信息不可用。

4、Open Images： 包含190万张图像，包含600个类别的15M对象。其中500个最常见的类别用于评估检测基准，70%以上的类别有1000多个训练样本。

5、LVIS： 是一个新收集的基准，有164000张图像和1000多类别，总共有220万个高质量的实例分割掩模。

6、ImageNet： 有200个类别，规模巨大，对象类似于VOC数据集，因此不是常用的检测算法基准。

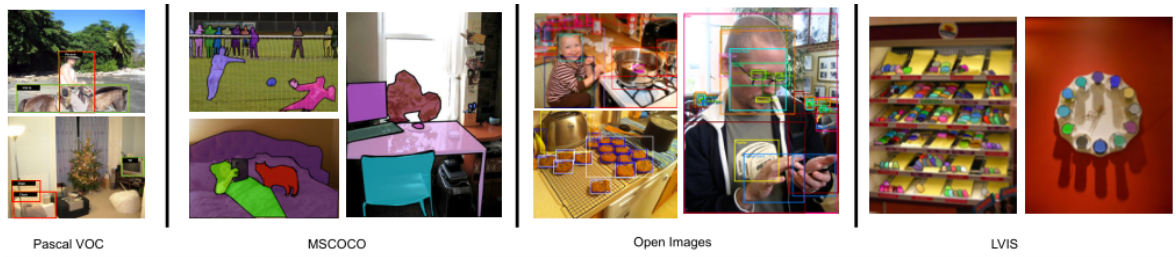


Figure 11: Some examples of Pascal VOC, MSCOCO, Open Images and LVIS.

评估指标： 检测精度和推理速度都用于评估检测算法

Alias	Meaning	Definition and Description	
FPS	Frame per second	The number of images processed per second.	
Ω	IoU threshold	The IoU threshold to evaluate localization.	
D_γ	All Predictions	Top γ predictions returned by the detectors with highest confidence score.	
TP_γ	True Positive	Correct predictions from sampled predictions	
FP_γ	False Positive	False predictions from sampled predictions.	
P_γ	Precision	The fraction of TP_γ out of D_γ.	
R_γ	Recall	The fraction of TP_γ out of all positive samples.	
AP	Average Precision	Computed over the different levels of recall by varying the γ.	
mAP	mean AP	Average score of AP across all classes.	
TPR	True Positive Rate	The fraction of positive rate over false positives.	
FPPI	FP Per Image	The fraction of false positive for each image.	
MR	log-average missing rate	Average miss rate over different FPPI rates evenly spaced in log-space	
Generic Object Detection			
mAP	mean Average Precision	VOC2007	mAP at 0.50 IoU threshold over all 20 classes.
		VOC2012	mAP at 0.50 IoU threshold over all 20 classes.
		OpenImages	mAP at 0.50 IoU threshold over 500 most frequent classes.
		MSCOCO	• AP_{coco}: mAP averaged over ten Ω: $\{0.5 : 0.05 : 0.95\}$; • AP_{50}: mAP at 0.50 IoU threshold; • AP_{75}: mAP at 0.75 IoU threshold; • AP_S: AP_{coco} for small objects of area smaller than 32^2; • AP_M: AP_{coco} for objects of area between 32^2 and 96^2; • AP_L: AP_{coco} for large objects of area bigger than 96^2;
Face Detection			
mAP	mean Average Precision	Pascal Face	mAP at 0.50 IoU threshold.
		AFW	mAP at 0.50 IoU threshold.
		WIDER FACE	• mAP_{easy}: mAP for easy level faces; • mAP_{mid}: mAP for mid level faces; • mAP_{hard}: mAP for hard level faces;
TPR	True Positive Rate	FDDB	• TPR_{dis} with 1k FP at 0.50 IoU threshold, with bbox level. • TPR_{cont} with 1k FP at 0.50 IoU threshold, with eclipse level.
Pedestrian Detection			
mAP	mean Average Precision	KITTI	• mAP_{easy}: mAP for easy level pedestrians; • mAP_{mid}: mAP for mid level pedestrians; • mAP_{hard}: mAP for hard level pedestrians;
MR	log-average miss rate	CityPersons	MR: ranging from $1e^{-2}$ to 100 FPPI
		Caltech	MR: ranging from $1e^{-2}$ to 100 FPPI
		ETH	MR: ranging from $1e^{-2}$ to 100 FPPI
		INRIA	MR: ranging from $1e^{-2}$ to 100 FPPI

Table 1: Summary of common evaluation metrics for various detection tasks including generic object detection, face detection and pedestrian detection.

6.2 人脸检测基准

1、WIDER FACE： 共 32203 幅图像，约 400k 张人脸，适用于大范围的缩放。其中40%用于训练，10%用于验证，50%用于测试。训练和验证集的注释可在线获得。根据检测任务的难度，有容易、中等和难三种难度。

2、FDDB： 是一个众所周知的基准，在2845张图像中有5171张人脸。通常人脸检测器首先在大规模数据集上进行训练，然后在这个数据集上进行测试。

3、PASCAL FACE： 是从PASCAL个人布局测试集中收集的，在851张图像中有1335个标记的人脸，与FDDB类似，它通常仅用作测试集。

评估指标： 如上图所示。

6.3 行人检测基准

1、CityPersons： 是在语义分割数据集CityScapes之上的一个新的行人检测数据集，其中5000张图像是在德国的城市中捕获的，共35000人，外加13000个忽略区域，提供所有人的边界框注释和可见部分的注释。

2、Caltech： 是最受欢迎和最具挑战性的行人检测数据集之一，它来自洛杉矶都市一辆汽车在街道上行驶大约10小时30Hz VGA视频记录。训练集和测试集分别包含42782和4024帧。

3、ETH： 包含三个视频片段中的1804帧，通常用作测试集来评估在大规模数据集（如CityPersons）上训练的模型性能。

4、INRIA： 包含从假日图像中收集的高分辨率行人图像，由2120张图像组成，包括1832张用于训练的图像和288张图像。具体来说，训练集有614个正图像和1218个负图像。

5、KITTI： 包含7481张分辨率为1250x375的标签图像和用于测试的7518张图像。其中人分为行人和骑自行车的人，都是用mAP方法评估的。含有三个评估指标：简单、中等和困难，还有最小差异、边界框最大高度和遮挡水平等。

评估标准：对于前三个数据集，使用从1e-2到100 FPPI（每幅图像的假阳性）范围内超过9个点的对数平均失败率来评估检测器性能，越低越好。对于KITTI，标准平均精度用作评估指标，阈值为0.5 IoU。

7、通用对象检测的最新技术

> 见论文表3（32页）

8、结束语和未来方向

> 讨论开放的挑战和未来的方向

1、可扩展的候选生成策略： 大多数检测器都是基于锚的方法，并且存在一些限制检测精度的关键缺点。目前的锚先验主要是人工设计的，难以匹配多尺寸对象，基于 IoU 的匹配策略也是启发式的。虽然已提出一些方法将基于锚的方法换成无锚的方法（如基于关键点的方法），但是仍然存在一些限制（如高计算成本等），有较大的提升空间。

2、语境信息的有效编码： 背景可能有助于或阻碍对象检测结果，因为视觉世界中的对象具有很强的关系，并且背景对于更好的理解视觉世界至关重要。然后很少有人关注如何正确使用上下文信息，如何有效的结合上下文进行目标检测可能是一个有前途的方向。

3、基于自动机器学习的检测： 为某项任务设计最佳主干架构可以显著提高结果，但是也需要巨大的努力。直接在数据集上学习主干架构是一个重要的研究方向，然而对于大多数研究人员来说自动机器学习所需的计算资源是负担不起的，因此开发一个低计算量的框架将对目标检测产生巨大的影响。此外，检测任务的新的结构策略（如候选生成和区域编码）可以在将来被探索。

4、对象检测的新兴基准： 目前MSCOCO是最常用的检测基准测试平台，然而其只有80类，无法理解现实世界中更复杂的场景，而最新的LVIS可能为将来更具挑战性的检测、分割和低射学习任务打开一个新的基准。

5、低射目标检测： 用有限的标记数据训练检测器被称为低射检测器。低射学习已经被积极地研究用于分类任务，但是只有少数研究集中于检测任务。

6、检测任务的主干架构： 将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经成为一种范式。然后分类和检测任务之间仍然存在冲突，因此这样做不是最佳解决方案。大多数最先进的检测算法都是基于分类主干的，只有少数尝试了不同选择。因此，如何开发一个检测感知的主干架构也是未来的一个重要研究方向。

7、其他研究问题： 如大批量学习和增量学习。批量大小是DCNN训练的一个关键因素，但在检测任务中还没有得到很好的研究。对于增量学习，如果在没有初始训练数据的情况下适应新任务，检测算法仍然会无法使用。