目标检测发展：大道至简

Object Detection Evolution:Less is more

曾世鹏
计算机科学与技术学院
华中科技大学

Abstract

本文全面回顾并分析了目标检测流程由繁至简发展历程，着重阐述了 HOG、DPM，R-CNN、YOLO 和 DETR 等里程碑式的技术，解析了其中可变形卷积、Focal Loss 和非极大值抑制等重要模块，并深入讨论了这些方法在处理如复杂视点、光照变化、小目标检测和处理遮挡目标等复杂任务中所面临的挑战。本文旨在为目标领域领域初学者提供全局视野，以便其对目标检测基本思想和方法有初步了解。

Keywords

目标检测，深度学习，技术演进

1. 引言

目标检测是众多计算机视觉任务中的重要一环，致力于在图像或视频中识别并定位出存在的对象，并为这些对象进行分类。简而言之，是回答：” 哪里有什么对象?” 的一项技术。

在计算机视觉的方法论和技术矩阵中，目标检测的角色非常重要，既借鉴并延展了前期的技术，又铺平了道路并启发了后续的探索。早期的目标检测算法大多基于图像分类技术，后续根据特定需求做了诸多改进。此外，目标检测在推动计算机视觉的发展上起着关键作用，许多关键技术，如目标追踪和实例分割，都是在目标检测的基础上进一步研发和优化的。

在技术层面，目标检测技术主要分为两大类：早期采用基于传统机器学习的方法，如方向梯度直方图（HOG）和可变性部分模型（DPM），在当时表现出了良好的性能；另一类为基于深度学习的方法，如 R-CNN、YOLO、DETR 等。深度学习的方法目前在多数任务，尤其是大型数据集上表现更加优秀，吸引了更多的关注。

这项技术在现实生活中已被广泛应用。在自动驾驶领域，目标检测能辅助识别道路上的车辆，行人或其他障碍物，方便汽车规划路线；在医疗领域，通过医学图像分析，目标检测技术能准确识别和定位病灶，辅助医生做出精确诊断。除此之外，在安全监控，智能交通等领域，目标检测同样发挥着不可替代的作用。

然而，目前的目标检测方法仍有许多需要优化的地方，例如，对复杂场景中的小目标检测，或者目标之间有严重遮挡等，这些场景都会对目前的目标检测方法带来挑战。针对这些问题的持续改进和技术创新，是我们可以期待的未来研究方向。本文的结构如下：

- 在第二章，我们会回顾目标检测的发展历程，我们将会看到目标检测的设计思想是如何从滑动窗口分类、逐渐演变为直接回归，然后到不需要锚框的检测的。
- 在第三章，我们会提到这些检测模块算法中的一些关键模块的设计，包括可变性卷积、Focal Loss 和非极大值抑制等。
- 第四章，是对目标检测面临的困难的讨论，包括复杂视点、光照下的检测，小目标检测和处理遮挡目标等。
- 在第五章，我们会对本篇综述进行总结，并提出对未来的展望。

2. 目标检测发展历程

在本文的这一部分，我们会分阶段介绍目标检测技术的演进过程，并会特别关注其中的一些里程碑式的技术。

2.1. 传统方法

在 2014 年以前，早期的目标检测算法主要依赖于手工制作的特征，并配合使用机器学习方法。这些算法开始时主要采用 SIFT（尺度不变特征转换）这样的特征表征方式。SIFT 是一种用于检测和描述图像中局部特征的方法，通过寻找尺度空间极值点以识别关键点，并通过关键点周围像素的方向信息来生成描述符，这样可以在尺度、旋转、亮度变化等条件下保持稳定。

然而，SIFT 等基于灰度的特征描述符在某些极端光照和复杂背景下的效果不甚理想。后来，出现的 HOG 和 DPM 算法在一定程度上克服了此问题，并通过对目标不同部分检测方法的融合，彻底改变了基于机器学习的目标检测方法。

2.1.1. HOG

HOG（Histogram of Oriented Gradients, 方向梯度直方图），是由 N. Dalal 和 B. Triggs 在 2005 年的 CVPR（Computer Vision and Pattern Recognition，计算机视觉和模式识别）会议上提出的一种用于图片处理的特征描述算法 [1]。

HOG 算法的工作原理主要为提取图像的局部特征，最基本的形式是计算并统计局部区域的梯度方向直方图。具体步骤包括：归一化图像（消除光照影响）；计算图片的梯度大小和方向；将图片划分为小的连通区域，此处称之为为“单元”；对每个“单元”计算梯度方向直方图；将相邻的“单元”集成为“区块”，对区块内的梯度方向直方图进行归一化处理；然后收集所有区块的 HOG 特征，形成最终的描述子。

由于 HOG 特征能有效地抓住图像的形状特征，而且对局部的光照和阴影具有较好的适应性，使其能在物体检测（尤其是行人检测）方面取得了很好的成效。

2.1.2. DPM

DPM（Deformable Part Models，可变形部分模型）是一种用于物体检测的计算机视觉技术，由 Pedro Felzenszwalb 在 2008 年提出 [2]，后来通过 R. Girshick 等人持续进行了大量改进 [3]。

DPM 的工作原理是将线性支持向量机（SVM）与混合部件增强模型相结合，以识别不同位置、大小和形状的对象。DPM 根据对象的局部结构创建一种模型，包括一个鲁棒的低分辨率“根”滤波器以及一些高分辨率的“零件”滤波器。所创建的模型对对象的某些部分进行形状建模。这些部分的位置相对于根可能会发生形变，每种形变都有一个相应的成本。同时，在特征空间中，”根”和”零件”皆会被映射为欧几里得空间中的欧几里得距离。然后，使用带有拉普拉斯金字塔的分层模型将所有物体的形变引入到一个统一框架中。

用通俗易懂的方法理解，DPM 模型可以理解为一个有弹性的“拼图游戏”。你可以想象你要在一副含有许多物品的图像中找出特定的物体（比如一辆汽车）。在这个游戏中，你拥有一些“拼图碎片”（一组零件滤波器），这些碎片分别对应汽车的不同部位，如轮胎、车窗、车门等，此外还有一个代表整个汽车的“大图”（根滤波器）。当你开始玩这个游戏时，你会尝试用这些“拼图碎片”去匹配图片中的对应部分。由于汽车可能以不同的角度或距离出现在图片中，所以你需要把这些碎片进行一定程度地拉伸或移动以便它们能与图片中的汽车吻合。这个过程就对应 DPM 中的形状建模和形变成本计算，通过计算不同形变对应的成本，可以找到一种最能匹配图片中汽车的“拼图碎片”组合。在所有的匹配过程中，你的目标是找到一种拼图组合，使得所有的拼图部分都能以最小的成本与图片中的汽车吻合，并且这些部分合在一起能构成一个完整的汽车图像，这就是 DPM 的目标。

DPM 在众多计算机视觉任务中取得了出色的成果，特别是在 PASCAL VOC 挑战赛中一度创下最好的检测效果。然而，物体检测技术的进步以及深度学习技术的兴起，使得 DPM 的性能相较于新型技术如 R-CNN 等已有所落后。

2.2. 两阶段检测

在 2014 年以后，随着卷积神经网络开始被运用于目标检测，并取得了远超原先方法的效果，深度学习逐步成为目标检测领域的主流手段。这个时期的卷积神经网络工作主要有两个阶段，即先对候选区分类，再调整定位，可理解为先框选图片中的某一区域作为感兴趣区域，再对这个区域通过分类网络判别是否存在目标，最后再通过

某些手段调整区域以精确定位目标。

2.2.1. R-CNN

R-CNN（Regions with Convolutional Neural Networks，带卷积网络的区域），是由 Ross Girshick 等人在 2014 年提出的一种用于目标检测的计算机视觉技术 [4]。

R-CNN 的工作原理是基于图像的区域提议以及滑动窗口算法。在 R-CNN 的工作流中，首先通过非极大值抑制（Non-Maximal Suppression, NMS）和选择性搜索（Selective Search）来生成约 2000 个候选区域，这些区域是根据颜色、纹理、大小和形状兼容性在输入图片上生成的。接下来，R-CNN 将每一个提议区域调整到同一大小，并输入到预训练的卷积神经网络（如 AlexNet、VGG16 等）中进行特征提取，由此导致了大量的特征计算冗余。最后，在区域精化阶段，R-CNN 进一步使用支持向量机（SVM）对每个区域进行分类，并使用线性回归模型调整区域的大小和位置以更准确地定位目标。正因为 R-CNN 的这种工作机制，即先提出区域，然后提取特征并分类以及定位，使得其成为了两阶段目标检测的始创者。然而，R-CNN 的这种设计也让其在速度上存在局限，计算量大且冗余影响了其在大规模图像集上的运算效率。此后的研究工作，如 Fast R-CNN 和 Faster R-CNN，逐步解决了 R-CNN 的这些问题，提高了计算效率并优化了性能。

2.2.2. Fast R-CNN

Fast R-CNN 是由 Ross Girshick 在 2015 年所提出，是 R-CNN 的关键改进之一 [5]。Fast R-CNN 主要解决了 R-CNN 在执行目标检测任务时效率低下的问题，它在 R-CNN 的基础上进行升级，改变了 R-CNN 需要为每一个提议区域单独提取特征的方式。

和 R-CNN 一样，Fast R-CNN 仍然在分阶段建模过程中需要产生候选区，但 Fast R-CNN 对原图进行了一次全局的卷积运算并将结果保留成为特征图。取代原 R-CNN 对每个区域提议都进行卷积的做法，它直接在特征图上截取对应的区域，然后通过一个固定大小的空间池化层（ROI pooling layer），将任意尺寸的区域转换成相同尺寸，进一步提取特征。这样，无论区域的原始尺寸大小如何，都可以得到固定长度的特征向量。这一步的目的是为后续的分类和定位操作提供统一大小的输入，以提高处理速度。

然后，提取出的特性通过全连通网络层进行分类和回归。分类操作通过 softmax 层和交叉熵损失函数来实现，而定位操作则通过边界框回归层和平滑 L1 损失函数来实现。最后，这两项损失函数被合成到一个单一的多任务损失中，以进行联合训练。这样，既可以减小模型训练误差，又能强化模型的泛化能力。

Fast R-CNN 的提出大大提高了目标检测的速度和精度，因为整个模型只需要对原图做一次前馈操作，就能一次性获得所有区域提议的特征。这样，Fast R-CNN 在保持深度学习优秀性能的同时，显著提高了效率。但 Fast R-CNN 仍然受制于区域提议的生成，这一步依然是一个独立的、计算密集的过程。为了解决这一问题，研究者们进一步提出了 Faster R-CNN。

2.2.3. Faster R-CNN

Faster R-CNN 是由 Ross Girshick 等人在 2015 年末提出的，作为 Fast R-CNN 的进一步改进 [6]。Faster R-CNN 的主要变革在于，它提出了一种名为区域提议网络（Region Proposal Network, RPN）的方法，用于自动获得高质量的区域提议。RPN 基本上是一种全卷积网络（Fully Convolutional Network, FCN），它能扫描整个图像并输出一组矩形区域提议和概率分数，这些区域可能包含一个目标。

在 Faster R-CNN 中，同样首先通过卷积网络提取特征图，然后使用 RPN 生成区域提议。RPN 会预设多种形状和大小的锚点，然后以滑动窗口的方式在特征图上进行滑动，对每个锚点生成预测框并给予一定的分数。预测框会根据其分数进行排序，并选择分数最高的部分作为区域提议。

最后，在这些区域提议基础上，再使用 Fast R-CNN 进行目标分类和定位精修。由此，Faster R-CNN 整合了 RPN 和 Fast R-CNN 的优点，整个过程只需要进行一次卷积，大大提高了目标检测的效率。Faster R-CNN 这种将提出目标区域的任务内嵌入到网络中，使得整个目标检测过程

成为一个统一的网络，既有效的提高了运行效率，又大幅度提升了检测精度。至此，由 R-CNN 引领的计算机视觉领域的两阶段检测模式达到了一个新的高度。然而，两阶段的检测模式在处理时效与准确度之间的平衡上，还有可以改进之处。此后的一阶段检测模式如 YOLO 以及后续研究工作，试图找到一个更合理的平衡点。

2.3. 一阶段检测

一阶段检测方法在推断步骤中直接预测目标类别和位置，不再需要生成提议区域和后处理阶段，从而在一定程度上减少了计算量并加快了推理速度。最具有代表性的一阶段方法是 YOLO 和 SSD。

2.3.1. YOLO

YOLO (You Only Look Once)，由 Joseph Redmon 和 Ali Farhadi 等人在 2016 年首次提出，是早期的一阶段目标检测方法之一 [7]。YOLO 的架构非常精简，仅通过一个前馈神经网络就能预测物体的边界框和类别。在 YOLO 的框架中，输入图像被分割为 $S \times S$ 的网格，每个网格单独预测 B 个边界框及其相关置信度，并预测 C 个条件类别概率。预测结果包括每个边界框的位置、大小、类别和置信度，这种独立预测的方式可以提升预测效率，降低空间依赖，增强模型的鲁棒性。简单来说，可以理解为图像的每一个网格分别预测每个目标的部分框和类别信息，再通过 NMS 等方法处理为完整物体检测框。

与两阶段模型相比，YOLO 更注重全局信息，在保持较高精度的同时，实现了实时的目标检测。然而，YOLO 在检测一些密集或小规模的目标时效果不够理想，其定位精度也略低于两阶段的检测方法。在 YOLO 的后续版本，如 YOLOv3[8] 和 YOLOv4[9]，都对这一问题进行了改进，并保持了较快的检测速度。

值得一提的是，尽管 Joseph Redmon 在 2020 年退出了 YOLO 项目，但 Ultralytics 公司在他们的基础上发布了 YOLOv5。这是一个在 YOLO 原始框架上做出众多优化和改进的版本，尽管并非官方版本，但在目标检测的精度、速度和效率方面都做出了显著的提升，因此仍被广泛应用。

得益于 Ultralytics 不断的努力和改进，YOLO 仍在不断优化更新，目前已推出的 YOLOv8 版本在目标检测领域实现了相当的实时性和准确性，其代码封装和可读性也值得肯定，对初学者来说，它是一个极好的学习资源。

2.3.2. SSD

SSD (Single Shot MultiBox Detector) 是由 Wei Liu 等人在 2016 年提出的一种一阶段目标检测方法 [10]。不同于 YOLO，SSD 在不同尺度的特征图上都进行预测，这使得 SSD 可以检测到不同大小的目标。同时，每一层的预测都使用一系列固定大小的先验框作基础，通过回归微调边界框的形状以获得更准确的结果。并且，SSD 也在各级先验框的置信度上采用了 softmax 分布，这使得它在目标置信度预测上有更自然的表现。

SSD 的提出改变了常规的思路，多尺度特征图和先验框的引入，对 SSD 在保持快速准确的检测性能上做出了巨大贡献。但它的一些缺点，如对于高精度的定位效果一般，并且在处理小目标上仍有所欠缺。但无论如何，SSD 仍然是一阶段检测方法中极为重要的一个里程碑。

总结起来，两阶段模型在检测准确性上通常表现优秀，但计算效率是其主要难题；而一阶段模型亦有其优点和短板：它们能以更高的速度进行推断，一些方法如 SSD 也能达到很高的准确率，但其特点是速度和精度通常很难同时做到最优。新的改进和模型仍在不断的挑战这个问题，以期找到更好的平衡点。

2.4. “无阶段”检测

当我们谈论目标检测的历史发展时，我们可能会怀疑是否存在一种超越“一阶段”和“两阶段”之间的方法，既可以获取高质量的准确性，又能保证有效的处理速度。不出所料，正如 CNN 让目标检测进入深度学习时代，Transformer 架构的出现让目标检测逐步踏入“无阶段”时代 [11]。

若将 R-CNN 那样显式地生成候选区域，再进行分类和框回归看作“两阶段”。在这个意义上，YOLO 可以认为是一个“一阶段”的模型。因为它虽然剪短了这个过程，直接在卷积特征图上预测框和类别，但对每个图像位置都进行预测并

产生多个纵向重叠的候选框，需要使用非极大值抑制来筛选并且去除重叠的预测框，仍需要一个后处理阶段。而无阶段模型, 例如 DETR 和 RT-DETR，被设计成既不预设生成阶段，也不需要后处理阶段，全程以一个单一、全局优化的网络进行处理。尽管它需要更多的参数来捕捉依赖性，但捋直了整个流程后，让模型可以更方便地进行端到端的优化。

2.4.1. DETR

DETR (Detection Transformer) 是由 Facebook AI 的研究团队（包括 Nicolas Carion 等人）在 2020 年提出的端到端的目标检测模型 [12]。DETR 将注意力模型 Transformer 从自然语言处理领域引入到计算机视觉任务中，以全新的方式解决目标检测问题。

在 DETR 的设计中，输入图像经过 CNN 编码后，被送入 Transformer 网络进行全局特征整合。然后，DETR 直接在全局上进行推断，通过预先固定数量的目标查询，生成像素级的预测与感知边界之间的二分匹配问题，然后利用匈牙利算法（Hungarian algorithm，一种寻找二分图中最大匹配的经典算法）进行优化求解。最后，得到了物体的边界框和类别。

值得注意的是，DETR 摒弃了许多传统目标检测方法中常用的技术，如样本选择，滑动窗口，锚框设计及非极大值抑制等。通过这种方式，DETR 具备了真正的端到端及无锚框（Anchor-free）的特性。即，不需要事先定义锚框或依赖特定大小的锚框来检测物体，而是由模型自主学习和推断出目标的大小和位置。

然而, DETR 的这种端到端和无锚框的设计，虽然在处理一些传统目标检测问题上降低了复杂性，但在实际应用中还存在一些挑战。尤其是对于小目标的检测，以及处理图像中对象数量动态变化的情况。

总的来说，尽管存在一些局限性，DETR 的提出仍然为目标检测领域提供了一个全新的视角和可能性。它的出现推动了计算机视觉领域的发展，并在一系列目标检测任务中显示出潜力和价值。后续基于 DETR 的改进，如 DINO 和 RT-DETR 在各领域取得了一定的成果，充分的展现了它的价值。

3. 目标检测关键模块

3.1. 可变形卷积

一般来说，标准卷积操作中的卷积核是要在输入的特征图中在空间位置严格与预设的卷积核对齐的。而可变性卷积（Deformable Convolution）[13] 是一种允许卷积核位置发生变形的机制，具体来讲，针对卷积核的每一个空间位置，都有一个可学习的偏移量，将其加到原本位置上，完成位置的偏移。这样一来，可以适应原图的形变，从而强化了模型对于空间形变的建模能力。

举一个简单的例子，假设我们的任务是识别一只猫，但猫可能会蜷缩或者伸展，这些都是形变。传统的卷积核在处理这些形变上可能会比较困难，但如果使用可变性卷积，可以更好地处理这种形变。

3.2. Focal Loss

Focal Loss 是一种专门为解决目标检测任务中正负样本不均衡问题设计的损失函数。它在标准的交叉熵损失基础上，引入了一个调整因子，使得模型在训练过程中对难分类（易出错）的样本给予更多关注。

Focal Loss 的定义如下:

对于二分类问题，输入 x 的标签为 $y \in \{+1, -1\}$，模型的输出为 $p \in [0, 1]$，其中 p 是模型预测正样本的概率。定义 p_t 如下:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases}$$

Focal Loss 可表示为:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$$

这里的 $\gamma \geq 0$ 是所谓的调节系数，当 γ 增大时，模型对于那些易于正确分类的样本所关注的程度就会降低。

3.3. 非极大值抑制

非极大值抑制（Non-Maximal Suppression,NMS）是物体检测中用来清楚冗余和重叠较高预测框子的常用技术。

工作原理首先是设定一个经验性的阈值，例如 0.3。然后，找到所有预测框中得分最高的，即置信度最高的，并将其添加到最内的物体位置列表中。之后的步骤是删除所有与当前得分最高的预测框交并比 (IoU) 值大于这个阈值的框。也就是说，删除所有重叠部分超过设定阈值的框。随后，对剩下的预测框重复这个过程，直到所有的框都被检查过。这样，最后保留下来的只有得分较高，并且重叠不大的预测框，有效地降低了冗余和重复检测的可能性。

3.4. 锚框

在目标检测领域中，例如 R-CNN 系列（包括 Fast R-CNN, Faster R-CNN）和 YOLO 等模型中，锚框（Anchor）是一个非常关键的概念 [14]。在这些模型中，锚框通常被定义为一个预设的固定框（可以使用多种尺寸和比例），它在整个输入图像上进行滑动窗口操作。

Anchor 的主要目的是为每个可能的物体位置提供一个初步的预测框 [14]。我们可以理解为在目标检测任务中，Anchor 提供了起始点或“种子”框，用于确定可能存在物体的区域。

模型在训练过程中，对这些 Anchor 进行调整（包括宽度、高度和中心点位置），以此逼近真实的物体框，从而实现对物体的准确检测 Anchor 在许多先进的目标检测模型中发挥实质性作用，无论是精度提升还是检测速度加快，都离不开 Anchor 的贡献。

4. 困难与挑战

4.1. 复杂视点下的检测

在目标检测中，复杂视点是一个重要的挑战。现实世界中的物体可以从各种视点或角度出现，比如一辆车可能会正面、侧面或者背部朝向摄像头，甚至可能部分被其他物体遮挡。另外一个更形象的例子是，我们可能只能看到一个人的侧脸或背影，而面部特征是实现准确识别的关键部分。

这些都意味着，对于同一个物体，如果没有优秀的算法，我们的模型需要学习其在各种视点下的各种外观，这将大大增加了模型的复杂性。因为目标检测算法必须具备足够的灵活性来处理这些视觉差异。

此外，由于不同视点可能会导致物体的形状、尺寸和外观变化，这给物体的边界框定位和分割也带来了额外的挑战。例如，当一个物体从侧面看时，它的形状可能会与当其从正面看时完全不同。对于模型来说，预测这种情况下物体的准确边界框就相对困难。另外，当物体的视点变化引发遮挡问题时，这个问题就更加突出了。

4.2. 复杂光照下的检测

光照变化下的目标检测也是一个极具挑战性的问题。无论是自然环境下的日夜变化、室内外的光线变化，还是人为因素如手电筒光源的变化，都会引发光照条件的变化，这些变化都会对物体的外观产生深远影响。

例如，强烈的日光可能会在物体上产生高对比度的阴影，有时甚至可能将物体的一部分过度曝光导致细节丢失。同样，弱光或暗光环境下，物体的可视信息可能大大减少，甚至可能使物体完全在黑暗中消失。这些都对目标检测任务带来了极大的挑战。

此外，光照变化也可能影响到目标的颜色、纹理等特性，使得同一个物体在不同光照下展现出截然不同的外观。因此，即使模型已经学习了物体在特定光照条件下的外观，也可能无法准确识别出在其它光照条件下的同一物体。这就需要目标检测模型具有更强的光照变化处理能力和更高的泛化性。

4.3. 小目标检测

小目标检测是目标检测的一个重要难点。在许多实际应用中，例如无人驾驶、智能监控以及医学图像处理等场景，都需要对图像中非常小的物体进行识别和检测。

一个典型的例子是在进行卫星遥感图像分析时，需要从高空拍摄的图像中识别出地面上的小尺寸物体，如小汽车、单人行走等。在这种情况下，小目标在整个图像中仅占据很小的一部分区域，且往往缺乏足够的细节和信息，使得传统的目标检测方法难以准确识别和定位这些目标。其次，由于小目标在图像中占据的像素少，容易受到周围环境的干扰。例如，在道路监控中，小的

汽车或行人可能会与周围的树木、建筑物等环境元素混淆，使得小目标的检测变得非常困难。

最后，在深度学习模型中，由于卷积操作和池化操作的使用，精细的小目标信息可能在特征提取过程中丢失，这对小目标的识别和检测也构成了一定的挑战。

综上所述，小目标检测的主要难点源于小目标所提供的信息量少、易受周围环境干扰以及可能在特征提取过程中信息丢失等因素。

4.4. 重叠目标的处理

当涉及到重叠或遮挡的目标检测时，难度也会大大增加。相比只涉及单一目标的检测，重叠目标常常需要对更复杂的场景进行分析和理解。在实际应用中，重叠目标的检测场景非常常见，比如人群密集的场所，多辆车并行的交通路口，甚至是拥挤的超市货架等都是典型的例子。在这些场景中，多个目标可能会同时出现在同一块区域内，或者一个目标被其它目标遮挡。

在这种情况下，一个关键的挑战是如何准确地定位并分割出每个目标。例如，当两个人重叠在一起时，如果不能准确地分割出每个人的边界，那么可能就会误检测为一个人。

另一个挑战是如何正确识别被遮挡的部分。例如，当一个物体被另一个物体部分遮挡时，可能只能看到物体的一部分。这就需要模型具备一定的推理能力，能够根据可见的部分推断出整个物体的性质。

最后，当遮挡程度极高时，如果遮挡物的大小、形状和颜色与被遮挡物相近，那么分割和识别任务就更加困难。因此，重叠目标的检测还需要模型具备丰富的上下文理解能力，能够对整个场景进行全局的分析和理解。

综合以上因素，重叠目标的检测无疑是一个技术难点，并需要深度学习模型具备较强的推理和全局理解能力。

5. 总结与展望

通过本文的讨论，我们可以看到目标检测在计算机视觉中的重要性，以及它的演变和进化。从早期的 SIFT 特征，再到后来基于梯度的 HOG 特征，然后是卷积神经网络的采用和发展，我们看到了目标检测接受怎样的挑战，并且如何突破这些挑战，达到更高的灵敏度和准确性。

最新的深度学习技术，特别是 R-CNN、YOLO 和 DETR 等算法，已经大大提高了目标检测的能力。然而，尽管这些技术取得了一些突破，但仍然充满了复杂和困难的问题，例如复杂视点、光照变化、小目标检测和遮挡目标的问题。

对于更复杂且挑战性的问题，我们仍然需要持续努力并改进现有的方法。技术的爆炸进化从未停止，从采用机器学习方法的“冷兵器时代”，到基于卷积神经网络的深度学习“热兵器时代”，再到现代已有不俗表现的多模态大模型如 GPT-4 Turbo with Vision，技术的发展从来都是一个从探索、到涌现、再理解简化、最后完全驾驭的循环过程。如今已经步入大参数大模型时代，我们需要理清缰绳，驯服野马的勇士，也需要在技术前线继续前行的开拓者，希望你我不论是选择尝试利用大模型解决已有问题，还是继续研发新的技术，都能保持一颗追求创新的心。目标检测仍然是一个活跃且充满机会的领域，让我们拭目以待。

参考文献

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 05)*, Jul 2005.

[2] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2008.

[3] 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*, page 1627–1645, Sep 2010.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic

segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2014.

[5] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1137–1149, Jun 2017.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[8] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Apr 2018.

[9] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-YuanMark Liao. Yolov4: Optimal speed and accuracy of object detection. *Cornell University - arXiv,Cornell University - arXiv*, Apr 2020.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. *SSD: Single Shot MultiBox Detector*, page 21–37. Jan 2016.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems,Neural Information Processing Systems*, Jun 2017.

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. *End-to-End Object Detection with Transformers*, page 213–229. Jan 2020.

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[14] Ross Girshick Jian Sun Shaoqing Ren, Kaiming He. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NeurIPS)*, 2015.