# 一阶段目标检测

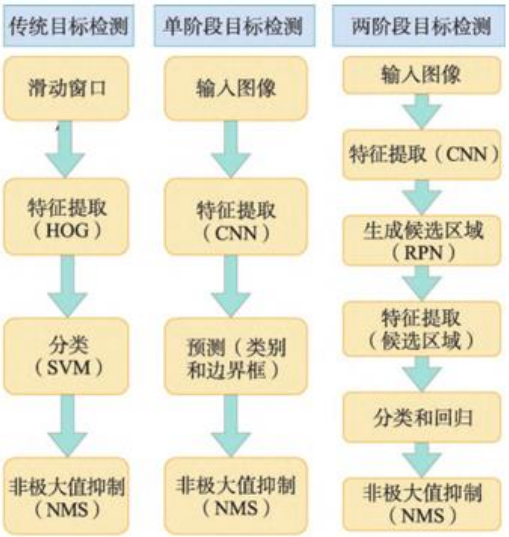
**摘要**　目标检测作为计算机视觉领域的核心任务之一，旨在精准定位图像或视频中的目标对象并完成类别划分，广泛应用于自动驾驶、智能监控、机器人视觉等场景。一阶段目标检测算法摒弃了二阶段算法中候选区域生成的冗余步骤，直接通过单次网络前向传播实现目标的定位与分类，在检测速度上具备显著优势，成为实时检测场景的主流技术方向。
**关键词**　目标检测，一阶段检测，计算机视觉

## Single-stage object detection

**Abstract**　As one of the core tasks in the field of computer vision, object detection aims to accurately locate target objects in images or videos and classify them, and it is widely applied in scenarios such as autonomous driving, intelligent surveillance, and robotic vision. Single-stage object detection algorithms eliminate the redundant steps of region proposal generation found in two-stage algorithms and directly achieve object localization and classification through a single network forward pass. They have significant advantages in detection speed, making them the mainstream technical approach for real-time detection scenarios.
**Key words**　Object detection，Single-stage detection，Computer vision

目标检测是计算机视觉领域的一项基础且核心的任务，旨在从静态图像或动态视频序列中自动识别并精准定位出感兴趣的物体实例，其影响力已渗透至现代社会的众多关键领域。随着深度学习的发展，目标检测算法的性能实现了质的飞跃，不断推动着相关技术应用的边界拓展与落地深化。

早期传统的目标检测方法主要依赖于手工设计的特征提取器与经典的机器学习分类器相结合的多阶段流水线。典型的流程如图 1 (a) 所示：首先在图像上采用多尺度滑动窗口策略密集地生成可能包含目标的候选区域；接着，对每个窗口提取手工设计的特征，如方向梯度直方图(HOG)、尺度不变特征变换(SIFT)等；最后，使用训练好的分类器，如支持向量机(SVM)，对提取的特征进行分类，判断窗口内是否存在目标及其类别。然而，这类方法在特征表达能力与泛化性能上存在局限，难以应对复杂场景与多样化的物体形态。




（a）传统　（b）单阶段　（c）两阶段

图 1 传统检测器、单阶段与两阶段检测器对比

Fig.1　Comparison among traditional detector, one-stage detector and two-stage detector

2012 年以来,以卷积神经网络为代表的深度学习方法彻底改变了目标检测的技术范式。基于深度学习的目标检测算法主要可

分为两大类：两阶段检测器与单阶段检测器

两阶段检测器遵循“先提议，后检测”的思想，其流程如图 1(c)所示。第一阶段负责从图像中生成一系列可能包含目标的候选区域。第二阶段则对这些候选区域进行精细化操作，通常包括特征重采样、精确的分类得分判断以及边界框位置的微调回归。如 Faster R-CNN，通过引入 RPN 实现了候选区域生成与检测网络的端到端联合训练，在准确率上设立了新的标杆。然而，这种分步处理的架构也带来了较高的计算复杂度和较慢的推理速度，在一定程度上限制了其在实时性要求苛刻的应用场景中的直接应用。

单阶段检测器则采用了更为简洁直接的设计理念，其流程如图 1(b)所示。这类方法通常在骨干网络提取的密集特征图上，直接应用一系列预设的锚框或更近期的无锚点机制，通过单个前向传播网络，并行地预测每个位置处目标的类别概率和边界框偏移量。这种高度集成化的设计极大地简化了检测流程，显著提高了推理效率，使其天生具备更优的实时性能。

单阶段检测器因其高效简洁的架构，在实时应用中展现出巨大潜力，但早期版本受限于正负样本极端不平衡等关键问题，其精度曾长期落后于两阶段方法。此后，该领域进入了以多样化创新驱动的高速发展期，涌现出诸如奠定多尺度预测基础的 SSD、RetinaNet、无锚点关键点检测范式的 CornerNet 与 CenterNet、提出全卷积无锚点设计的 FCOS、引入锚框精炼机制的 RefineDet，以及实现网络架构与特征融合协同优化的 EfficientDet 等代表性工作。本文将聚焦于上述七种核心算法，系统阐述其原理、贡献与演进关系，并对未来挑战与方向进行展望。

## 1 目标检测技术评估指标和数据集

### 1.1 精度指标

混淆矩阵是机器学习中用于评估分类模型性能的工具，通过展示模型在不同类别上的预测结果与真实标签之间的关系，帮助研究者了解模型的分类性能。混淆矩阵通常用于二分类问题，但也可以扩展到多分类问题。

1) 真阳性 $TP$ （true positive），实际为正例且被预测为正例的样本数.

2) 假阳性 $FP$ （false positive），实际为负例但被预测为正例的样本数.

3) 假阴性 $FN$ （false negative），实际为正例但被预测为负例的样本数.

4）真阴性 $TN$ （true negative），实际为负例且被预测为负例的样本数.

根据混淆矩阵可以计算出机器学习领域模型的 3 种性能评估指标，精确率、召回率以及 F1 分数。

1) 精确率：表示预测为正类的样本中真正 为正类的比例。

$$\mathrm{Pr}ecision=\frac{TP}{TP+FP} \tag{1}$$

2) 召回率：表示所有真正为正类的样本中被正确预测为正类的比例。

$$\mathrm{Re}call=\frac{TP}{TP+FN} \tag{2}$$

精确率与召回率均为越高越优，根据精确率-召回率（P-R）曲线可知精确率和召回率的变化关系，通常情况下，随着召回率的增加，精确率会下降，即二者存在一种权衡关系，不能同时最大化。为了全面评估模型，引入了 F1 分数，其表示精确率和召回率的调和平均值。

$$F1\text{-}score=\frac{2\times \mathrm{Pr}ecision\times \mathrm{Re}call}{\mathrm{Pr}ecision+\mathrm{Re}call} \tag{3}$$

3) IoU 是目标检测领域最早和最基本的评价指标之一，同样只能评估单个目标的检 测效果。其主要用于衡量预测框和真实框的重叠程度， 特点是简单直观、易计算。

$$IoU=\frac{|PB\cap TB|}{|PB\cup TB|} \tag{4}$$

其中，PB 表示预测框，TB 表示真实框。

多目标检测中常用的评估指标是平均精度（AP），可以定义为各种召回下的平均检测精度，并以某一个特定类别的方式进行评估。AP 通过计算精确率-召回率曲线下的面积来衡量模型的性能。

$$AP=\int_0^1 P(r)dr \qquad (5)$$

为了能够评估模型在所有类别上的平均性能，引入所有对象类别的平均值，即平均精度均值（mAP）。mAP 由所有类别的 AP 值加权平均求得。其成为目标检测和相关领域评估的最终指标。

$$mAP=\frac{1}{N}\sum_{i=1}^{N} AP_i \qquad (6)$$

### 1.2 计算量和速度指标

理论计算复杂度主要通过浮点运算次数（FLOPs）和参数量（Parameters）来度量。FLOPs 指完成一次模型前向推理所需的浮点运算总次数，其单位通常为 GFLOPs（10^9 次），它直接反映了模型的计算负担，与网络深度、宽度及算子复杂度紧密相关。参数量则是模型中所有可学习参数（权重与偏置）的总和，决定了模型的存储占用大小，并在一定程度上影响内存访问开销。这两项指标共同刻画了模型对计算和存储资源的固有需求，是进行模型设计与对比的基础性理论依据。

实际运行效率则通过推理速度来表征，具体体现为每秒帧率（FPS）或单张图像推理时间（Inference Time）。FPS 指在特定的硬件平台（如指定型号的 GPU）和软件环境下，模型每秒能够处理并输出结果的图像数量；推理时间是处理单张图像所需的毫秒数，二者互为倒数。需要注意的是，实际推理速度是一个高度情境依赖的指标，它不仅取决于模型本身的理论复杂度（FLOPs），还深刻受限于硬件计算能力、内存带宽、驱动与框架的优化水平、输入批处理大小以及后处理开销等诸多因素。因此，对算法进行效率对比时，必须在相同的软硬件配置与测试条件下进行，综合理论指标与实际测速结果方能做出客观评估。

### 1.3 数据集

目标检测算法的训练、验证与性能评估，均依赖于大规模、高质量且具有挑战性的标注数据集。数据集的选择与演进，直接影响着算法发展的方向与性能上限。本文所综述的七个单阶段检测算法，其骨干网络通常在大规模图像分类数据集（如 ImageNet）上进行预训练，以获取强大的通用特征提取能力；而后在目标检测专用数据集上进行微调与评估。其中，PASCAL VOC 与 MS COCO 是两个里程碑式的基准，主导了过去十年目标检测技术的发展与比较。

1）ImageNet 是计算机视觉领域最具影响力的超大规模图像数据库之一。它基于 WordNet 的层次结构进行组织，涵盖了超过 2 万个类别，图像总数超过 1400 万张。其中，约有 100 万张图像提供了精确的目标边界框标注，构成了 ImageNet 检测挑战赛的数据基础。虽然 ImageNet 检测任务本身并非主流评估基准，但其图像分类任务 为几乎所有现代目标检测算法提供了至关重要的骨干网络预训练权重。例如，ResNet、EfficientNet 等骨干网络在 ImageNet-1k 分类数据集上预训练后，其迁移至检测任务的特征提取能力显著优于随机初始化，这是提升检测精度、加速模型收敛的关键步骤。

2）PASCAL VOC 挑战赛（2005-2012 年）是推动目标检测技术早期发展的核心驱动力。其数据集规模适中，标注质量高，至今仍被广泛用于算法原型验证和教学。最常用的版本是 VOC 2007 和 VOC 2012，两者类别一致，均为 20 个常见物体类别（如人、动物、车辆、室内物品等），但图像集合互不重叠，常被合并使用以扩充训练数据。该数据集的特点是场景相对简单，每张图像中的目标实例数量较少。其主要评估指标是 mAP@0.5（即 IoU 阈值为 0.5 时的平均精度均值）。早期具有里程碑意义的单阶段检测器，如 SSD，均在 PASCAL VOC 数据集上报告了其基础性能，证明了多尺度预测架构的有效性。

3）MS COCO 数据集是当前目标检测领域公认的权威基准与主战场。相较于 PASCAL VOC，MS COCO 在规模、复杂性和评估标准上都有质的飞跃。规模更大，场景更复杂：包含超过 30 万张图像，80 个物体类别。图像中目标实例密度更高，且包含

大量非图标式的日常场景，背景干扰更强；小目标占比极高：数据集中包含海量的小尺寸目标，这对模型的多尺度特征融合与细小目标检测能力提出严峻挑战；评估标准更严格：其核心评估指标为 mAP@[.5:.95]，即在 IoU 阈值从 0.5 到 0.95（步长 0.05）的多个等级上计算平均精度并求均值。这要求模型不仅要有高召回率，还必须具备极其精准的定位能力。

本文重点综述的 RetinaNet、FCOS、EfficientDet 等现代单阶段检测器，其核心论文均以在 MS COCO 数据集上取得突破性的 mAP 为主要贡献点。因此，后续章节对各算法性能的横向对比与分析，将主要依据其在 MS COCO 数据集上的精度与效率指标展开。

表 1 目标检测常用数据集概览

Table 1 Overview of the Common Object Detection Datasets

| 名称 | 类别 | 训练集（图像数/目标实例数） | 验证集（图像数/目标实例数） | 训练-验证集（图像数/目标实例数） | 测试集（图像数/目标实例数） |
|---|---|---|---|---|---|
| PASCAL VOC 2007 | 人、鸟、猫、牛、狗、马、羊、飞机、自行车、船、巴士、汽车、摩托车、火车、瓶子、椅子、餐桌、盆栽、沙发和电视/显示器 | 2501/6301 | 2510/6307 | 5011/12608 | 4952/12032 |
| PASCAL VOC 2012 | 与 PASCAL VOC 2007 相同 | 5717/13609 | 5823/13841 | 11540/27450 | - |
| MS COCO 2014 | 80 类：人、自行车、汽车、摩托车、飞机、巴士、火车、卡车、船、交通灯、消防栓、停车标志、停车表、长凳、鸟、猫、狗、马、羊、牛、大象、熊、斑马、长颈鹿、背包、雨伞、手提包、领带、行李箱、飞盘、滑雪板、滑雪板、运动球、风筝、棒球棍、棒球手套、滑板、冲浪板、网球拍、瓶子、红酒杯、杯子、叉子、刀、勺、碗、香蕉、苹果、三明治、橙子、西兰花、胡萝卜、热狗、披萨、甜甜圈、蛋糕、椅子、沙发、盆栽、床、餐桌、厕所、电视、笔记本电脑、鼠标、遥控器、键盘、手机、微波炉、烤箱、烤面包机、水槽、冰箱、书、时钟、花瓶、剪刀、泰迪熊、吹风机和牙刷 | 82783 | 40504 | 123287 | 40775 |
| MS COCO 2015 | 与 MS COCO 2014 相同 | 82783 | 40504 | 123287 | 81434 |
| MS COCO 2017 | 与 MS COCO 2014 相同 | 118287 | 5000 | 123287 | 40670 |

## 2 单阶段目标检测算法

### 2.1 SSD

传统两阶段检测器（如 Faster R-CNN）依赖于独立的区域提议网络（RPN）和分类回归网络，流程复杂，速度受限。SSD（Single Shot MultiBox Detector）的核心贡献在于构建了一个完全卷积的、单阶段前向传播的检测框架，摒弃了耗时的提议生成与特征重采样步骤，首次实现了高精度与高速度（实时）的实质性统一。其设计的精髓在于多尺度特征图的密集预测与先验框（Prior Box/Default Box）系统。

多尺度特征图的密集预测架构：SSD 以修改的 VGG-16 网络（移除了全连接层）为基础骨干。其创新性地在骨干网络后附加了一组逐渐缩减的卷积层（Conv6 到 Conv11），形成了一个特征金字塔。如图 1 所示，SSD 在六个不同尺度的特征图上进行预测：conv4_3 (38x38), conv7 (19x19), conv8_2 (10x10), conv9_2 (5x5), conv10_2 (3x3), conv11_2 (1x1)。浅层特征（如 conv4_3）分辨率高，富含细节，专用于检测小物体；深层特征（如 conv11_2）感受野大，语义信息强，用于检测大物体。

先验框设计与预测机制：在每个特征图的每个空间位置上，SSD 预设了一组具有不同尺度（Scale）和长宽比（Aspect Ratio）的先验框，作为预测的基准锚点。对于特征图$k$ ($k \in [1, m]$)，其先验框的尺度$s_k$由公式线性计算：

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1}(k-1)$$

其中，$s_{min} = 0.2, s_{max} = 0.9$。对于每个尺度，SSD 使用一系列长宽比，$a_r \in \left\{1,2,3,\frac{1}{2},\frac{1}{3}\right\}$。先验框的宽度$w_k^a = s_k\sqrt{a_r}$，高度$h_k^a = s_k/\sqrt{a_r}$。对于长宽比为 1 的情况，额外增加一个尺度为$s_k' = \sqrt{s_k s_{k+1}}$的先验框。因此，每个位置通常有 4 或 6 个先验框。

在预测时，对于尺寸为 $m \times n$、通道数为 $p$ 的特征图，在每个位置应用一个 $3 \times 3 \times p$ 的卷积核。对于每个先验框，该卷积核输出 $(c+4)$个值：$c$个类别置信度分数（包含背景类），以及 4 个边界框偏移量（相对于先验框的中心点 $(cx, cy)$ 和尺寸 $(w, h)$ 的调整值）。因此，对于有$k$个先验框的位置，共需 $(c+4)k$个卷积滤波器。

训练策略与损失函数：训练的核心是将真实边界框（Ground Truth）分配到对应的先验框。SSD 采用两级匹配策略：首先，每个真实框匹配与其具有最高交并比（Jaccard Overlap/IoU）的先验框；其次，任何与真实框 IoU 大于阈值（0.5）的先验框也被视为正样本。这种多对一的匹配方式简化了学习任务，允许网络对同一物体进行多次预测。

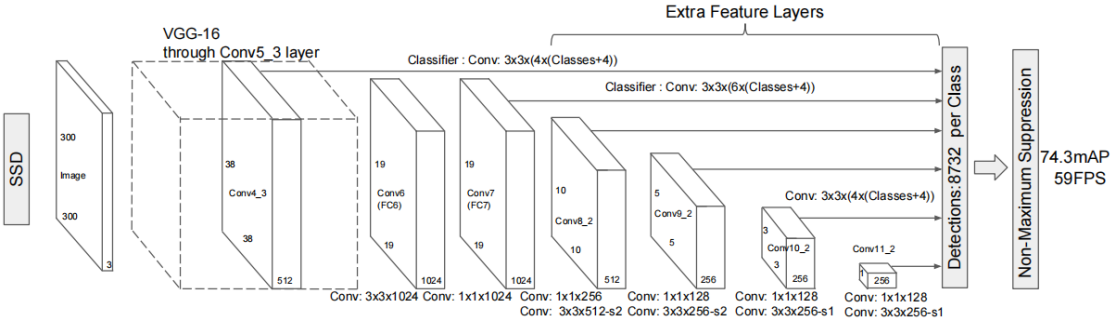



图 2　SSD 结构图

Fig. 2　SSD Network Architecture

SSD 的总体损失函数是定位损失（Loc）和置信度损失（Conf）的加权和：

$$L(x, c, l, g) = \frac{1}{N}(L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g))$$

其中 $N$ 是匹配成功的正样本先验框数量，$\alpha$ 为平衡权重（通过交叉验证设为 1）。

定位损失采用 Smooth L1 损失，仅对正样本（$x_{ij}^k = 1$）计算。回归目标是预测框 $l$ 与匹配的真实框 $g$ 之间的偏移量：

$$L_{loc}(x, l, g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^k \, \text{smooth}_{L1}(l_i^m - \hat{g}_j^m)$$

其中，$\hat{g}_j^{\text{cx}} = \frac{g_j^{\text{cx}} - d_i^{\text{cx}}}{d_i^w}$，$\hat{g}_j^w = \log\left(\frac{g_j^w}{d_i^w}\right)$，（其余维度同理），$d$ 为先验框参数。

置信度损失采用 Softmax 交叉熵损失，对正样本和负样本均计算：

$$L_{\text{conf}}(x,c) = -\sum_{i\in\text{Pos}}^{N} x_{ij}^{p}\log\left(\hat{c}_i^{p}\right) - \sum_{i\in\text{Neg}}\log\left(\hat{c}_i^{0}\right)$$

其中$\hat{c}_i^{p} = \frac{\exp(c_i^{p})}{\sum_p \exp(c_i^{p})}$。

为了应对正负样本的极端不平衡（正样本极少），SSD 采用了困难负样本挖掘（Hard Negative Mining）：对负样本按其置信度损失排序，仅保留损失最高的那些样本，使得正负样本比例约为 1:3。

数据增强：SSD 使用了非常激进的数据增强策略，远超同期方法。除了标准的随机水平翻转和颜色扭曲外，关键是对输入图像进行随机采样与裁剪（Random Sampling & Cropping）：从原始图像中随机采样一个区域，其与目标的最小 IoU 随机选自 {0.1, 0.3, 0.5, 0.7, 0.9}，采样区域的大小在原始图像的 [0.1, 1] 之间，长宽比在 [1/2, 2] 之间。这种“放大”操作显著增加了训练数据的多样性，尤其提升了对小目标和部分遮挡目标的鲁棒性。在后续改进中，SSD 还引入了“缩小”操作（将图像随机放置在 16 倍大的画布上），进一步提升了小目标检测性能（mAP 提升 2-3%）。

在 PASCAL VOC2007 测试集上，SSD300（输入 300x300）实现了 74.3% mAP @ 59 FPS，SSD512（输入 512x512）实现了 76.9% mAP @ 22 FPS，在精度和速度上均显著超越了当时的 Faster R-CNN 和 YOLO，确立了单阶段检测器的标杆地位。

### 2.2 RetinaNet

在 RetinaNet 之前，以 YOLO 和 SSD 为代表的单阶段检测器虽然在推理速度上显著优于两阶段的 Faster R-CNN 等模型，但其检测精度，尤其是对小目标的检测精度，通常存在一段明显的差距。

这种差距的根源被 RetinaNet 的作者 Lin 等人系统地归结为训练过程中极端的前景-背景类别不平衡。具体而言，在一张典型的图像中，可能只包含数个到数十个待检测的目标（前景），而单阶段检测器在特征图上进行的密集锚框采样会产生成千上万个候选位置，其中绝大部分都是不包含任何目标的简单背景。

在标准的交叉熵损失函数下，这些海量的、容易分类的负样本（背景）虽然每个产生的损失很小，但因其数量巨大，会完全淹没掉数量稀少但至关重要的正样本（前景）产生的损失，导致模型优化方向被背景样本主导，从而无法有效地学习到区分前景目标的关键特征。

RetinaNet 的核心贡献在于创造性地提出了 Focal Loss，一种动态缩放的交叉熵损失函数，它通过重塑标准交叉熵损失的形式，成功解决了上述类别不平衡问题。Focal Loss 通过修改损失函数本身，自动降低那些已经被模型分类得很好（即预测概率很高）的简单样本（无论是背景还是前景）对总损失的贡献权重，从而让训练过程更多地聚焦于那些难以分类的样本。其数学表达式在二分类情况下可表述为：

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (7)$$

其中，$p_t$ 表示模型对真实类别的预测概率。这个公式引入了两个关键的调制因子：首先，权重因子 $\alpha_t$ 用于对正负样本进行基础的平衡，通常可以为正样本设置一个较小值，为负样本设置较大值，以初步缓解数量不均。聚焦因子 $(1-p_t)^{\gamma}$，其中 $\gamma$ 是一个大于 0 的可调节参数（通常设为 2）。当样本被错误分类或分类置信度不高时，聚焦因子接近 1，损失几乎不受影响；当样本被正确分类且置信度很高时，聚焦因子会趋近于 0，从而将该样本的损失大幅降低。这意味着，对于大量能够被轻松判断为背景的锚框，其损失被显著压制，模型将有限的学习容量聚焦于那些模棱两可、难以区分的“困难样本”上，例如部分遮挡的目标、与背景纹理相似的物体，或者尺寸极小的目标。这种“聚焦”机制极大地提升了训练的效率与效果，使得单阶段检测器无需复杂的采样策略也能获得高质量的特征表示。

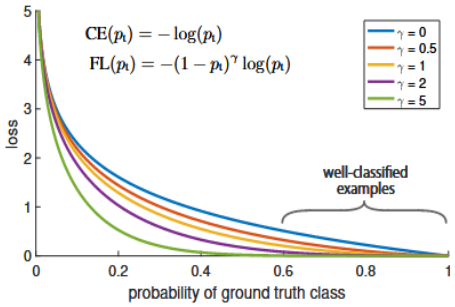



图 3　Focal Loss 对不同分类难度样本的调制效果

Fig.3　Schematic diagram of the modulation effect of Focal Loss on samples with different classification difficulties.

RetinaNet 的网络架构设计整体结构可以被清晰地解构为三个组成部分：一个用于提取多尺度特征的主干网络、一个用于构建高层语义与底层细节融合的特征金字塔网络，以及两个分别负责目标分类和边界框回归的任务特定子网络。

RetinaNet 通常采用在 ImageNet 上预训练过的 ResNet 等网络作为主干网络，从中提取多个不同层级的特征图。然而，深层特征图虽然语义信息丰富，有利于目标类别的判断，但空间分辨率低，对目标的定位不利；浅层特征图则相反，分辨率高、细节丰富，但语义信息较弱。为了兼顾分类与定位，RetinaNet 集成了 FPN 结构。FPN 通过自顶向下的路径和横向连接，将深层的高语义特征与浅层的高分辨率特征进行融合，生成一组具有强语义信息且多尺度的特征金字塔。

在 FPN 提供的每一层特征图上，RetinaNet 并行连接两个结构相同但参数不共享的小型全卷积子网络。分类子网络负责预测每个锚框位置处属于各个类别的概率。它接收 FPN 某一层的特征图，经过若干次卷积后，最终通过一个通道数为 K*A 的 3x3 卷积层输出预测结果，其中 K 是类别数，A 是该层设置的锚框数量。边界框回归子网络则负责预测每个锚框相对于其对应真实目标框的精细位置偏移量。其结构与分类子网络对称，最终输出通道数为 4*A。

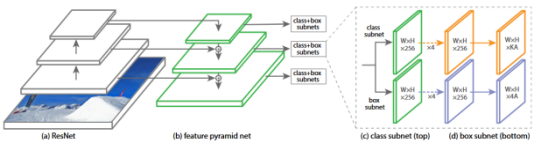



图 4 RetinaNet 网络结构

Fig.4　RetinaNet Network Architecture

RetinaNet 在标准基准上的表现验证了其巨大成功。实验表明，仅使用 Focal Loss 替换标准交叉熵损失，就能在原有单阶段检测器上实现精度提升。完整的 RetinaNet 模型在保持高速的同时，其精度首次全面超越了当时所有最先进的两阶段检测器。

这一成就具有里程碑式的意义，它彻底扭转了人们“单阶段检测器速度优先、精度次之”的固有印象，证明了通过精心设计的损失函数解决核心矛盾，单阶段检测器完全有能力在精度和速度上取得双赢。

## 2.3 RefineDet

Refine Det 的核心创新在于融合两阶段检测器的精度优势与单阶段检测器的效率优势，通过“取长补短”的设计思想，在保持实时性的同时显著提升检测精度。

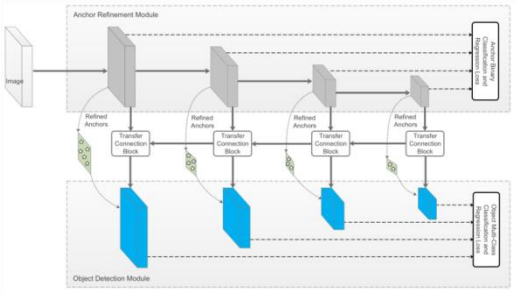



图 5　RefineDet 结构

Fig. 5　RefineDet Network Architecture

Refine Det 算法结构如图 5 所示。

ARM：为常规 SSD 多尺度预测阶段，做预测所提取的特征图分别为：conv4_3，conv5_3，fc7，conv6_2。每一个特征图都会有两个子网络分支，为别为预测锚框位置的子网络和预测是否为背景类别的子网络，剔除背景置信度较高的样本。

ODM：和 SSD 基本相同，融合不同层的特征。主要的不同点一方面在于这部分的输入 anchors 是 ARM 部分得到的粗调和筛选后的锚框，类似RPN网络输出的proposal。另一方面，Refine Det 的浅层特征图通过 TCB 模块融合了高层特征图的信息，将各层融合的结果再整合到一起进行回归预测。

TCB：

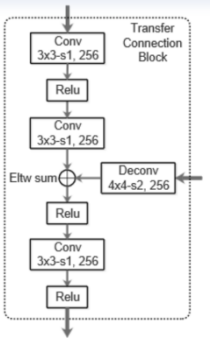



图 6　TCB 结构

Fig. 6　TCB Architecture

TCB 模块通过反卷积，实现高层语义信息向浅层传播，实现更高精度的回归预测。

## 2.4 CornerNet

先前的方法依赖大量的锚框（Anchor Boxes）引发诸多问题。一方面，锚框数量通常高达数万甚至数十万，导致正负样本极度不均衡，严重影响训练效率；另一方面，锚框的尺寸、长宽比等设计依赖人工经验，在多尺度检测场景中配置复杂，且不同尺度的预测结果映射回原图会产生误差。

基于此，CornerNet 将目标的边界框检测问题转化为目标左上角点和右下角点的预测。其创新点主要包括：一是提出角点池化（Corner Pooling），解决角点缺乏局部视觉特征支撑的定位难题；二是采用关联嵌入（Associative Embedding）策略，通过嵌入向量实现角点的精准配对；三是设计适配角点检测的损失函数与网络结构，在简化检测流程的同时提升检测精度，最终在 MS COCO 数据集上实现 42.2%的 AP，超越当时所有单阶段检测器。

CornerNet 的网络结构如图 7（a）所示，其以堆叠沙漏网络（Hourglass Network）为骨干网络，该网络通过下采样聚合全局特征、上采样恢复分辨率，并借助跳跃连接保留细节信息，能有效捕捉目标的多尺度特征，适配不同大小目标的角点检测需求。

网络整体框架包含两大预测模块，分别对应左上角和右下角角点的检测，预测模块的结构如图 7（b）所示。每个预测模块由修改后的残差块（嵌入角点池化层）、卷积层及多分支输出层组成，最终输出三类关键信息：角点热力图（Heatmaps）、位置偏移（Offsets）和嵌入向量（Embeddings）。

角点池化是 CornerNet 的核心组件，旨在解决边界框角点常位于目标外部、缺乏局部视觉证据的定位难题。其核心思想是通过聚合特定方向的全局特征，为角点判断提供依据。

对于左上角角点检测，角点池化层接收两个特征图，分别沿水平向右和垂直向下方向进行最大池化，再将池化结果相加。水平方向聚合目标顶部边界特征，垂直方向聚合目标左侧边界特征，二者结合实现左上角角点的精准定位。右下角角点池化则沿水平向左和垂直向上方向进行最大池化，原理与左上角池化对称，如图 8 所示。

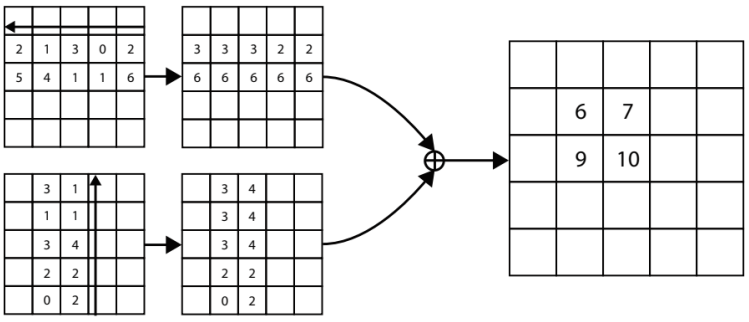
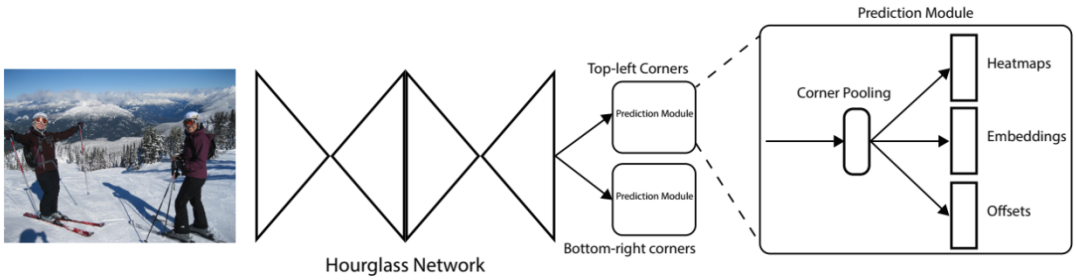
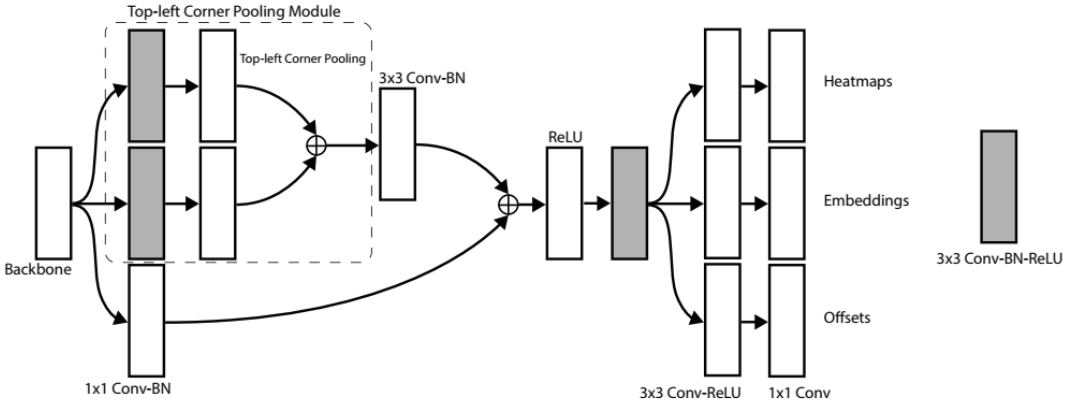



图 8　角点池化示意图

Fig. 8　Corner Pooling




(a)




(b)

图 7　CornerNet 整体结构图(a)和预测模块结构图(b)

Fig. 7　CornerNet Overall Architectural Diagram(a) and Structure diagram of the prediction module(b)

热力图用于表征不同类别角点的位置概率，分为左上角和右下角两组，每组含 C 个通道（C 为目标类别数），无背景通道。训练时，每个角点的真实位置为正样本，其余为负样本，但为避免过度惩罚真实角点附近的负样本，引入未归一化的 2D 高斯函数 $e^{-\frac{x^2+y^2}{2\sigma^2}}$ 降低半径范围内的惩罚，$\sigma$为 1/3 的半径，半径范围内两点与真实标注至少具有 0.3 的交并比。损失函数采用改进的焦点损失（Focal Loss），通过$(1-y_{cij})^4$项降低真实角点附近负样本的损失权重，如公式（8）所示。

$$L_{det}=\frac{-1}{N}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\begin{cases}(1-p_{cij})^2\log(p_{cij}) & if\ y_{cij}=1\\ (1-y_{cij})^4(p_{cij})^2\log(1-p_{cij}) & otherwise\end{cases} \quad (8)$$

由于网络下采样操作，特征图上的角点坐标映射回原图时会产生整数离散误差，尤其影响小目标检测精度。为此，CornerNet 预测位置偏移量$o_k$，对离散化后的角点坐标进行微调，训练时仅在真实角点位置计算 Smooth L1 损失，如公式（9）、（10）、（11）所示。

$$o_k=\left(\frac{x_k}{n}-\left\lfloor\frac{x_k}{n}\right\rfloor,\frac{y_k}{n}-\left\lfloor\frac{y_k}{n}\right\rfloor\right) \quad (9)$$

$$L_{off}=\frac{1}{N}SmoothL1Loss(o_k,\widehat{o_k}) \quad (10)$$

$$SmoothL1Loss(x)=\begin{cases}0.5x^2 & if\ |x|<1\\ |x|-0.5 & otherwise\end{cases} \quad (11)$$

嵌入向量用于实现角点配对，每个角点对应 1 维嵌入向量，同一目标的左上角和右下角角点嵌入向量距离应最小化，不同目标的角点嵌入向量距离应大于阈值。训练时通过拉拽损失（Pull Loss）使同一目标的两角点嵌入向量趋近于平均值，通过推送损失（Push Loss）使不同目标的角点嵌入向量距离大于阈值。测试时计算左上角与右下角角点嵌入向量的 L1 距离，仅保留距离小于 0.5 且类别一致的角点对，生成边界框。损失函数如公式（12）、（13）所示，$e_{t_k}$是目标$k$左上角点对应的嵌入向量，$e_{b_k}$是目标$k$右下角点对应的嵌入向量，$e_k$是$e_{t_k}$和$e_{b_k}$的均值。

$$L_{pull}=\frac{1}{N}\sum_{k=1}^{N}\left[(e_{t_k}-e_k)^2+(e_{b_k}-e_k)^2\right] \quad (12)$$

$$L_{push} = \frac{1}{N(N-1)} \sum_{k=1}^{N} \sum_{\substack{j=1 \\ j \neq k}}^{N} \max(0, 1 - |e_k - e_j|) \quad (13)$$

训练时采用 Adam 优化器，总损失函数融合检测损失、拉拽损失、推送损失和偏移损失，如公式（14）所示，权重分别设为 1、0.1、0.1 和 1；数据增强采用随机水平翻转、缩放、裁剪、颜色抖动及 PCA 处理，有效缓解过拟合。

$$L = L_{det} + 0.1L_{pull} + 0.1L_{push} + L_{off} \quad (14)$$

推理阶段：对热力图应用 3×3 非极大值抑制，筛选前 100 个左上角和右下角角点；通过偏移量修正角点坐标，计算嵌入向量距离实现角点配对；结合原始图像与水平翻转图像的检测结果，应用 Soft-NMS 去除冗余边界框，保留前 100 个检测结果。

## 2.5 CenterNet

CenterNet是另一种不依赖锚框的算法，与 CornerNet 不同，其通过预测目标的中心点位置来确定目标位置。它的核心创新点主要体现在三个方面，一是通过峰值提取替代 NMS 后处理，提升了检测效率；二是构建了灵活的多任务扩展体系，基于中心点估计的核心框架，通过新增独立输出头即可实现三维检测、姿态估计等任务的适配，各任务共享骨干网络特征，无需重新设计网络结构，实现了通用框架的构建；三是优化了网络结构与训练策略，通过引入可变形卷积层增强模型对目标形变的适应性，提升定位精度；

CenterNet 的核心思路是将目标检测转化为关键点估计问题，通过预测目标中心点及相关属性实现目标定位与识别，整体方法可分为基础检测框架、多任务扩展及网络结构设计三部分。

在基础目标检测框架中，模型以全卷积骨干网络为基础，核心目标是通过预测三类关键特征完成目标检测，各特征的定义、训练方式及整体优化逻辑如下：

第一类核心特征是目标中心点热力图。该热力图用于精准定位图像中各目标的中心位置，其维度为$\hat{Y} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$，其中 W、R 为输入图像分辨率，R 为输出步幅，C 为目标类别数。训练时，模型先将真实目标的中心点映射到低分辨率热力图上，以映射后的中心点为中心生成 2D 高斯分布作为正样本标签，其余位置为负样本。为解决正负样本数量不均衡问题，模型采用焦点损失（Focal Loss）训练该分支，损失函数定义如公式（15）所示。

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1 - \hat{Y}_{xyc})^{\alpha} \log(\hat{Y}_{xyc}) & if\ Y_{xyc} = 1 \\ (1 - \hat{Y}_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} \log(1 - \hat{Y}_{xyc}) & other \end{cases} \quad (15)$$

其中Y为热力图标签，N为图像中目标数量，(α、β)为调节参数，用于进一步平衡正负样本权重。

第二类核心特征是局部偏移量预测。由于模型输出步幅 R 的存在，将原始图像坐标映射到低分辨率热力图时会产生离散化误差，导致中心点定位精度下降。为此，模型设计偏移量预测分支，输出$\hat{O} \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R} \times C}$，其中 2 个通道分别对应 x 轴和 y 轴方向的偏移量。该分支采用 L1 损失训练，损失函数如公式（16）所示。

$$L_{off} = \frac{1}{N} \sum_{p} \left| \hat{O}_{\tilde{p}} - \left( \frac{p}{R} - \tilde{p} \right) \right| \quad (16)$$

其中 p 原始图像中的真实中心点坐标，$\tilde{p}$为映射到热力图后的整数坐标，$\hat{O}_{\tilde{p}}$为模型在 p 位置预测的偏移量。

第三类核心特征是目标尺寸回归。模型直接预测每个目标边界框的宽高，输出$\hat{S} \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R} \times C}$，2 个通道分别对应边界框的宽度和高度。尺寸回归同样采用 L1 损失训练，为控制计算量，该分支为所有类别共享同一预测头，损失函数定义如公式（17）所示。

$$L_{size} = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{S}_{p_k} - s_k \right| \quad (17)$$

其中$s_k = (x_2^{(k)} - x_1^{(k)}, y_2^{(k)} - y_1^{(k)})$为第 k 个目标的真实边界框坐标，$\hat{S}_{p_k}$为模型在中心点$p_k$位置预测的尺寸。

基础检测框架的整体训练目标为上述三类损失的加权和如公式（18）所示。

$$L_{det} = L_k + \lambda_{size} L_{size} + \lambda_{off} L_{off} \quad (18)$$

其中$\lambda_{size}$和$\lambda_{off}$为损失权重参数，用于平衡各分支损失的贡献度。在推理阶段，模型先通过提取热力图峰值获取目标中心点位置

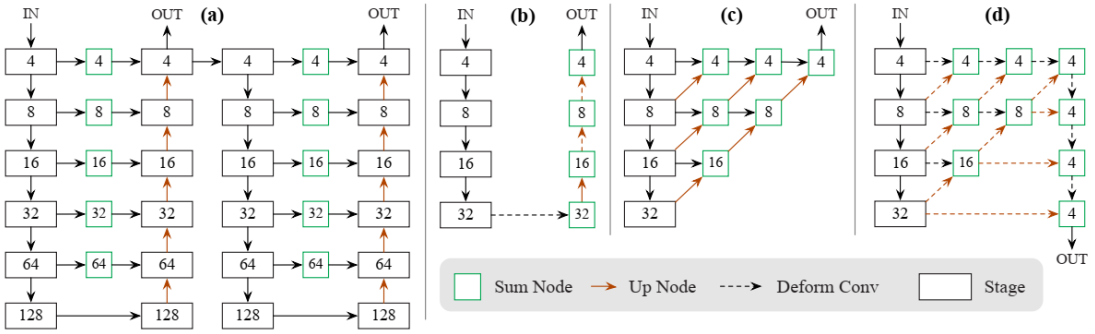



图 9　模型图示。方框内的数字表示图像的步幅。（a）沙漏网络，直接使用 CornerNet 的网络。（b）带有转置卷积的 ResNet，在每一个上采样层之前添加一个 3×3 可变卷积层。(c)用于语义分割的原始 DLA-34。（d）改进后的 DLA-34，添加了更多的跳跃连接，并将上采样阶段中的每个卷积层升级为可变卷积层。

Fig. 9　Model diagrams. The numbers in the boxes represent the stride to the image. (a): Hourglass Network, We use it as is in CornerNet. (b): ResNet with transpose convolutions. We add one 3 × 3 deformable convolutional layer before each up-sampling layer. (c): The original DLA-34 for semantic segmentation. (d): Our modified DLA-34. We add more skip connections from the bottom layers and upgrade every convolutional layer in upsampling stages to deformable convolutional layer.

（筛选出数值大于等于 8 邻域像素的响应，保留前 100 个峰值），再结合对应位置的偏移量$\hat{O}$修正中心点坐标，得到精准的原始图像坐标$(\hat{x}_i + \delta\hat{x}_i, \hat{y}_i + \delta\hat{y}_i)$（其中$(\hat{x}_i, \hat{y}_i)$为热力图峰值坐标，$(\delta\hat{x}_i, \delta\hat{y}_i)$为预测偏移量）；最后利用预测尺寸$(\hat{w}_i, \hat{h}_i)$生成目标边界框，坐标为$\left(\hat{x}_i + \delta\hat{x}_i - \frac{\hat{w}_i}{2}, \hat{y}_i + \delta\hat{y}_i - \frac{\hat{h}_i}{2}, \hat{x}_i + \delta\hat{x}_i + \frac{\hat{w}_i}{2}, \hat{y}_i + \delta\hat{y}_i + \frac{\hat{h}_i}{2}\right)$。整个推理过程无需非极大值抑制（NMS）等后处理步骤，仅通过热力图峰值提取即可完成冗余框过滤，大幅提升了检测效率。

在多任务扩展方面，针对三维目标检测任务，模型新增三个独立输出头：深度预测头、三维尺寸回归头和朝向预测头。其中，深度预测采用 Eigen 等人提出的输出变换方法，通过 sigmoid 函数将深度值映射到[0,1]区间后再反向转换，避免直接回归深度的难度；朝向预测借鉴 Mousavian 等人的策略，采用双区间+区间内回归编码，通过 8 个标量实现朝向估计（每个区间 4 个标量，含 2 个分类标量和 2 个角度回归标量）；三维尺寸则直接回归以米为单位的绝对值。针对人体姿态估计任务，模型将姿态视为目标中心点的 k×2 维属性（k 为关节点数量，COCO 数据集 k=17），回归关节点相对于中心点的偏移量，同时引入自下而上的关节热力图优化策略：先通过热力图提取高置信度关节点，再以中心点偏移量为分组依据，将回归的关节点与热力图检测到的关节点匹配，实现多人姿态估计。

在网络结构设计上，本文采用四种骨干网络进行实验，包括 ResNet-18、ResNet-101、DLA-34 和 Hourglass-104，如图 9 所示。其中，ResNet 系列和 DLA-34 均引入可变形卷积层进行改进：ResNet 在三个上采样层前添加 3×3 可变形卷积层，调整上采样层通道数以节省计算量；DLA-34 则新增底层到输出层的跳跃连接，并将上采样阶段的所有卷积层替换为可变形卷积，增强特征融合能力；Hourglass-104 则直接沿用原始结构，该网络含两个串联的沙漏模块，通过对称的下采样与上采样结构及跳跃连接，保留丰富的细节信息。

## 2.6 FCOS

在深度学习目标检测的发展历程中，长期以来，单阶段检测器为了在速度与精度间取得平衡，普遍依赖于预定义锚框（Anchor Boxes）的设计。锚框作为一组密集铺设在图像上的先验框，通过预测其与真实框的偏移量（如中心点偏移、宽高缩放）来完成检测。SSD、RetinaNet、EfficientDet 等主流算法均属此类。

尽管 Anchor-Based 方法取得了巨大成功，但其固有缺陷也日益凸显：首先，锚框的数量、尺寸、宽高比等超参数需要根据特定数据集精心设计和调整，泛化性较差且引

入先验偏差。其次，为了覆盖各种尺度和形状的目标，往往需要在每个空间位置部署海量锚框（如 RetinaNet 在特征金字塔上每个位置部署 9 个锚框），导致巨大的正负样本不平衡和计算负担。最后，锚框机制与后续的 IoU 计算、非极大值抑制（NMS）等后处理步骤耦合紧密，整个流程不够简洁。

在此背景下，一种回归目标检测本质的思路——“Anchor-Free”开始受到重视。其核心思想是摒弃复杂的锚框设计，直接预测目标本身的关键属性（如边界点、中心点等）。Tian 等人于 2019 年提出的 FCOS 算法（Fully Convolutional One-Stage Object Detection）正是在这一思潮下诞生的里程碑式工作。它首次证明了在一个完全卷积的、无锚框的框架下，单阶段检测器可以达到甚至超越当时主流 Anchor-Based 方法的精度，同时设计更简单、更灵活。FCOS 将目标检测任务重构为逐像素（per-pixel）的预测问题，为后续众多 Anchor-Free 方法（如后来的 CenterNet、CornerNet 的改进版等）开辟了道路。

FCOS 的核心创新在于其“逐像素预测”的范式。它完全移除了锚框，将特征图上的每个空间位置直接视为一个训练样本。对于一个给定位置，模型直接预测其到所属目标边界框四条边的距离，即一个四元组（l, t, r, b），分别代表到左侧、上侧、右侧和下侧的距离。这种设计使得检测过程变得极为直观。在正负样本匹配上，FCOS 也更为简洁；如果一个位置落在任何真实框内，它就被视为正样本并负责预测该框，这避免了基于锚框方法中复杂的 IoU 阈值计算。为了处理不同尺度的目标并解决多个目标框重叠区域带来的歧义，FCOS 巧妙地引入了特征金字塔网络（FPN）进行多尺度预测。具体而言，不同层级的特征图负责预测不同尺度范围的目标，小目标由高分辨率特征层负责，大目标由低分辨率特征层负责。例如，在 P4 特征图上，只有当目标的最大边界距离（max(l, t, r, b)）处于预设范围（如 64 到 128 像素之间）时，该位置才被视为该层的正样本。这种分层策略有效分配了不同大小目标的预测任务，显著缓解了重叠区域的歧义。

FCOS 的网络架构清晰高效。它通常采用 ResNet 等网络作为骨干，并构建一个包含 P3 至 P7 层的 FPN 来提取多尺度特征。一个共享的检测头被应用于 FPN 的每一层，该检测头包含三个并行分支：一个分类分支预测该位置所属的物体类别；一个回归分支预测一个 4 维向量，即从该位置到真实边界框左、上、右、下四条边的距离；此外，还有一个额外的中心度（Centerness）分支。其中，中心度分支是 FCOS 的一个关键设计，它专门预测一个标量，用于度量当前位置与目标几何中心的偏离程度，其真实值由公式定义，在目标中心处为 1，在边界处趋近于 0。在训练时，分类任务使用 Focal Loss 来解决样本不平衡，回归任务使用 IoU Loss 或 GIoU Loss 以获得更好的定位精度，中心度预测则使用二元交叉熵损失。在推理阶段，模型直接输出每个位置的类别置信度、边界距离和中心度得分。最终的检测置信度由类别分数与中心度相乘得到，这一操作能有效抑制那些偏离目标中心的低质量预测框，然后通过非极大值抑制（NMS）输出最终结果。

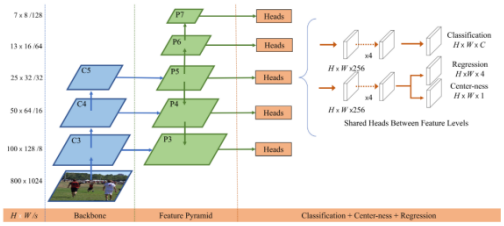



图 10　FCOS 网络结构

Fig. 10　FCOS Network Architecture

FCOS 算法的优势是多方面且显著的。首先，其最大的贡献在于彻底移除了预定义锚框这一复杂组件，使算法设计变得异常简洁。这不仅减少了对数据集敏感的众多超参数（如锚框尺寸、比例和数量），提高了模型的泛化能力，也避免了训练中计算海量锚框与真实框之间 IoU 的巨大开销。其次，在性能上，FCOS 展现了强大的竞争力。由于其逐像素的预测方式，理论上可以产生更密集和灵活的正样本位置，从而经常在召回率（Recall）上优于锚框方法，在 COCO 等标准基准上达到了当时顶尖的精度。其三，中心度机制是一个精巧而有效的设计。它是一个可学习的模块，而非启发式规则，能让网络自主评估每个预测点的质量，并在 NMS

前对低质量预测进行抑制，显著提升了最终的平均精度（AP）。其四，FCOS 的结构呈现出高度的统一性和灵活性。其全卷积、逐像素的形式与语义分割任务天然契合，便于进行多任务学习与扩展。同时，简洁的框架使其能够轻松集成最新的骨干网络、训练技巧和模块改进，例如后续的 ATSS（自适应训练样本选择）算法进一步提升了其性能，证明了该框架强大的可扩展潜力。总之，FCOS 不仅仅是一个高效的检测器，它更代表了一种更清晰、更本质的检测范式，成功推动了目标检测领域从锚框依赖向锚框自由的深刻转变，其思想持续影响着后续众多先进检测模型的发展。

## 2.7 EfficientDet

随着模型规模膨胀，检测器的效率问题日益突出。EfficientDet 的核心目标是在极致的计算效率约束下实现最高精度。它并非简单改进某个组件，而是对检测器的骨干网络（Backbone）、特征融合网络（Neck）和缩放策略（Scaling）进行了系统性协同优化。其两大支柱是 BiFPN（加权双向特征金字塔网络）和 Compound Scaling（复合缩放方法）。

传统多尺度特征融合方法（如 FPN 的自上而下，PANet 的自上而下+自下而上）通常对输入特征进行简单相加，忽视了不同分辨率特征对输出贡献的差异性。此篇论文提出的 BiFPN （结构如图 11 所示）进行了三项关键优化。

结构简化与效率提升：移除那些只有一个输入边的节点（因为它们不进行特征融合），并增加从原始输入到输出的快捷连接（如果它们在相同层级），形成了一个更精简、更高效的双向（自顶向下+自底向上）连接拓扑。

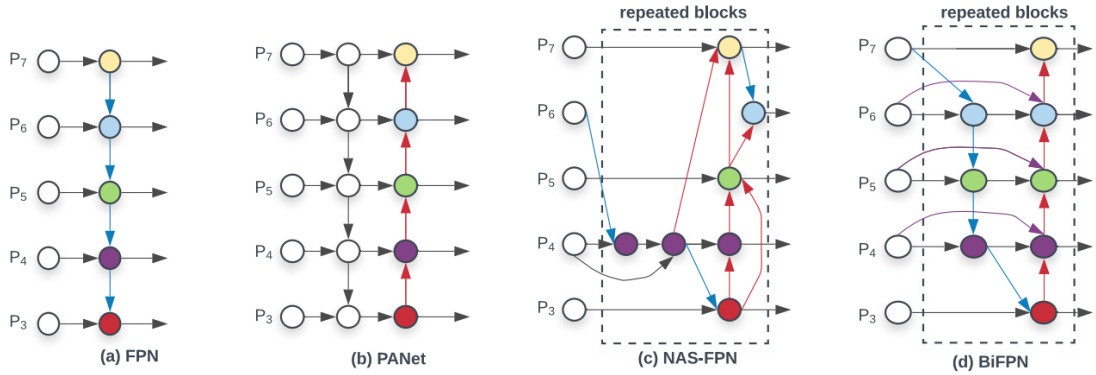



图 11　FPN、PANet、NAS-FPN、BiFPN 结构

Fig. 11　FPN、PANet、NAS-FPN、BiFPN Architecture

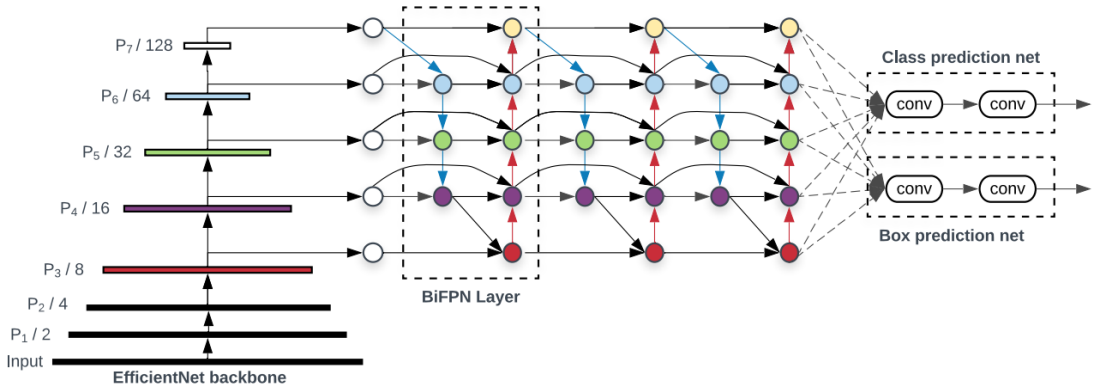



图 12　EfficientDet 架构

Fig. 12　EfficientDet Network Architecture

可学习的加权特征融合：提出了快速标准化融合（Fast Normalized Fusion）。对于一组要融合的输入特征 $I_i$，输出 $O$ 计算如下：

$$O = \sum_i \frac{w_i}{\epsilon + \sum_j w_j} \cdot I_i$$

其中，$w_i$ 是可学习的标量权重（每个特征对应一个），通过 ReLU 激活确保非负，$\epsilon = 0.0001$ 用于数值稳定。该操作赋予网络根据输入特征的重要性动态调整融合权重的能力。论文对比了无界融合（$O = \sum_i w_i I_i$）和基于 Softmax 的融合（$O = \sum_i \frac{e^{w_i}}{\sum_j e^{w_j}} I_i$），发现快速标准化融合在取得与 Softmax 相近精度的同时，GPU 推理速度快了约 30%。

深度可分离卷积：在 BiFPN 的所有特征融合卷积操作中，均使用深度可分离卷积（Depthwise Separable Convolution）以大幅减少计算量。

EfficientDet的单阶段检测器架构如图3所示，它在网络结构中的贡献在于提出了一个统一的复合缩放方法，使用一个单一的复合系数 $\phi$ 来联合放大网络的所有维度（骨干、BiFPN、预测头、分辨率），而非像之前工作那样只缩放其中一两个。

缩放规则基于启发式与网格搜索确定。骨干网络直接复用 EfficientNet-B0 至 B6 的预定义宽度与深度系数，确保能够充分利用其 ImageNet 预训练权重。对于核心的特征融合模块 BiFPN，其宽度（通道数）按照如下公式进行指数增长，以捕获更丰富的特征表示：

$$W_{bifpn} = 64 \cdot (1.35^{\phi})$$

而其深度（层数）则通过 $D_{bifpn} = 3 + \phi$ 进行线性增加，从而增强特征融合能力。分类与边界框预测网络的宽度设置为与 BiFPN 保持一致，即 $W_{pred} = W_{bifpn}$，以实现特征维度对齐；其深度则依据公式 $D_{box} = D_{class} = 3 + \lfloor \phi/3 \rfloor$ 线性增加。在输入分辨率方面，由于 BiFPN 使用了从 1/8 下采样（level 3）至 1/128 下采样（level 7）的多级特征，输入尺寸必须是 128 的整数倍以确保特征图尺寸规整，因此分辨率按照 $R_{input} = 512 + \phi \cdot 128$ 线性提升。

通过设定复合系数 φ 从 0 至 7 的整数值，我们构建了从 EfficientDet-D0 到 D7 的完整模型谱系，具体配置详见表 1。以两个端点模型为例：EfficientDet-D0（φ=0）采用 EfficientNet-B0 作为骨干网络，输入分辨率为 512x512，BiFPN 宽度为 64 通道且深度为 3 层，预测网络深度同样为 3 层；而 EfficientDet-D7（φ=7）则采用更强大的 EfficientNet-B6 骨干网络，将输入分辨率提升至 1536x1536，BiFPN 扩展至 384 通道与 8 层深度，预测网络深度也相应增加至 5 层。这一系列模型通过统一的缩放规则，能够灵活覆盖从移动端到数据中心等广泛场景下的不同计算与精度需求。

在训练细节与性能方面，训练采用了多项优化：使用 SGD with momentum (0.9)、权重衰减 (4e-5)、余弦学习率衰减、SiLU (Swish-1) 激活函数、指数移动平均 (EMA)。损失函数为 Focal Loss ($\alpha = 0.25, \gamma = 1.5$) 用于分类，L1 损失用于边界框回归。数据增强采用了大幅度的尺度抖动[0.1, 2.0]，即在训练时随机将图像缩放到原图的 0.1 倍到 2.0 倍之间，然后再裁剪。论文发现，在较长训练周期（如 300 轮）下，这种强力的尺度抖动能起到有效的正则化作用，显著提升模型性能。

在 COCO 2017 test-dev 上的结果令人瞩目：EfficientDet-D7 在单模型单尺度下取得了 55.1% AP 的当时最优性能，而参数量仅为 77M，计算量 (FLOPs) 为 410B。相比之前的先进检测器（如基于 AmoebaNet 和 AutoAugmentation 的 NAS-FPN），D7 在 AP 上高出 4.0 个百分点，同时参数量减少 2.7 倍，计算量减少 7.4 倍。EfficientDet 不仅在精度上领先，在实际硬件（GPU/CPU）上的推理速度也比同期同类精度的模型快数倍，符合其高效率的设计宗旨。

## 3 总结与展望

在以经典卷积为架构的单阶段检测器中，SSD 预测多个边界框，但它使用一组不同比例和纵横比的预定义默认框（或锚点），即通过在网络的不同层检测不同尺度的对象，融合了多尺度特征将每层的网络都参与到目标检测中，使得模型显著提高了对不同比例的物体检测精度，弥补了对于一些小物体检测不敏感的缺陷，但是其卷积层数较少，

缺少对深层语义信息的提取。RetinaNet 针对小目标的检测精度低问题，引入了 FocalLoss 损失函数和 FPN 特征提取层，虽然提高了对困难分类样本的关注度，但是模型复杂度很高，其与 SSD 相比推理速度较慢，进而在边缘设备上的性能表现有限。而 FCOS 是一种完全卷积的单阶段目标检测算法，采用无锚点和中心点预测，简化了模型结构并提高了精度，使得整体的参数量下降，但与上述模型相比其推理速度相对较慢。EfficientDet 通过复合缩放可以同时统一缩放所有骨干网络等，以及 BiFPN 在多尺度特征图之间进行双向特征融合提高了特征的丰富性和层次性，进而实现了高性能和低计算成本的平衡，但模型训练时间长、收敛慢。此外，CenterNet 和 CornerNet 均采用无锚点方法，分别通过中心点和角点检测提高精度，适合更高精度的检测任务且二者均保持速度与精度的平衡，但是前者对中心点定位精度需求较高使得需要更多的计算资源，而后者容易缺失对象内部特征信息。最后，RefineDet 的核心思想是融合单阶段检测的速度与双阶段检测的精度。它通过锚框优化模块（ARM）初步筛选与调整锚框，再经目标检测模块（ODM）进行精细分类与回归，中间由转移连接块（TCB）传递特征。该设计有效缓解了类别不平衡问题，提升了小目标检测精度，在当时实现了速度与准确度的较好平衡，并启发了后续的级联检测思路，但是相比纯单阶段方法，其双模块设计增加了计算和调试成本。

### 3.1 对比分析

**表 2 主要单阶段检测器优缺点分析**

Table 2 Analysis of the Advantages and Disadvantages of Primary One-Stage Detectors

| 模型 | 发布时间 | 特点优势 | 局限性 | 应用场景 |
|---|---|---|---|---|
| SSD | 2016 | 多参考检测与多分辨率检测、高效性、速度快、泛化能力强、适应边缘设备部署 | 卷积层数较少，缺少对深层语义信息的提取，模型检测精度较低 | 交通检测、农业检测、实时视频分析等 |
| RetinaNet | 2017 | 引入 Focal Loss 损失函数，增加 FPN 结构，提高模型精度 | 推理速度很慢，模型复杂度高 | 小目标检测、工业检测、生物个体检测等 |
| RefineDet | 2018 | 融合单阶段检测器的速度和双阶段检测器的精度优势，有效解决了单阶段方法的类别不平衡和特征利用不充分问题 | 相比纯单阶段方法其双模块设计增加了计算和调试成本，其性能受预设锚框的尺寸、比例影响，不够灵活 | 工业检测、基础设施检测等 |
| CornerNet | 2018 | 无锚点，将检测转变为角点检测（成对的对角关键点），多尺度检测，精度高 | 容易缺失对象内部特征信息，需要对属于同一对象的角关键点进行分组使得模型计算复杂度增高 | 交通检测、生物个体检测、农业检测等 |
| CenterNet | 2019 | 无锚点，将检测任务视为三元组，使用中心点预测，速度与精度达到平衡 | 对小目标的中心点定位不够准确，训练时间长，需要较多的计算资源，对部署在边缘设备很不友好 | 3D 目标检测、生物个体检测、交通检测、农业检测等 |

| | | | | |
|---|---|---|---|---|
| FCOS | 2019 | 无锚框，使用中心点预测替换锚框机制，降低计算复杂度提高精度 | 推理速度慢，模型计算复杂度高，对资源受限设备不友好 | 安防监控、工业检测、高精度检测等 |
| EfficientDet | 2019 | 引入复合缩放、加权双向特征金字塔（BiFPN）网络，模型轻量灵活性强 | 模型复杂度高，训练时间长，对小目标检测不敏感 | 移动设备、嵌入式系统、 UAVs 检测等 |

### 3.2 未来发展方向

近年来，单阶段目标检测器凭借其结构简洁、推断快速的显著优势，得到了广泛应用与持续发展。然而，该技术路径仍面临诸多挑战：高质量、有代表性的公开数据集仍然稀缺，尤其在特定专业领域，数据采集与标注成本高昂；模型普遍存在泛化能力有限、严重依赖海量标注数据进行训练等问题，同时在计算资源受限的移动端等边缘设备上部署困难。此外，尽管研究不断深入，当前模型的感知速度与效率同人眼视觉系统相比仍有巨大差距。特别是在复杂背景中检测小目标时，模型易受干扰，导致误检和漏检率较高。为此，本文梳理了以下潜在研究方向，以期为单阶段检测器的进一步发展提供参考。

（1）生成式大模型增强

生成式大模型技术蓬勃发展，给各个领域带来了新的研究方向。针对目标检测领域数据集采集困难以及数据集多样性和代表性不足，模型过拟合到特定类型的输入上，从而降低模型泛化能力，而使用生成式网络大模型可以很好地解决这个问题，其原理是通过学习真实数据的分布特征，生成新的数据样本，这些样本在特征上与真实数据相似，但又具有一定的变化和多样性。在数据稀缺的情况下，生成式模型可以生成大量合成数据，减少对真实数据的依赖，从而降低数据标注的成本。

（2）探索开发弱监督检测

由于基于深度学习的单阶段检测器的训练通常依赖于大量注释良好的图像，注释过程耗时、昂贵且效率低下。而弱监督检测（weakly supervised object detection，WSOD）仅利用有限的标注信息来训练模型，使其能够识别和定位图像中的对象。这些标注信息可以是图像级别的标签，而无需具体的边界框。如多实例学习（multiple instance learning，MIL）策略是经典的弱监督检测策略，其中训练数据以“包”（bags）的形式提供，每个包内包含多个未标注的实例（instances），而整个包则有一个整体标签。在单阶段检测器中，图像被视为包，图像中的基于锚点或网格单元的候选框作为实例。训练过程试图找出那些能够最好解释图像标签的候选框，即最有可能包含目标物体的检测框。

（3）小目标检测

长期以来，面对如野外、开放区域等背景复杂的大型场景，单阶段检测器在检测小目标方面一直是一项挑战。该研究方向的一些潜在应用包括利用遥感影像统计野生动物的种类数量和军事目标检测等。解决该问题可以引入多尺度特征融合（如 FPN、BiFPN）、高分辨率输入来增强检测能力，引入注意力机制利用空间注意力或通道注意力模块突出重要区域，帮助模型聚焦于小目标周围的背景信息。如 Focus-DETR 在编码器中引入双重注意力机制和基于多尺度特征图的 token 评分系统，专注于更有小目标信息量的 token。此外还可以通过加权损失函数，对小目标赋予更大的权重，使得它们在训练过程中得到足够的关注进而提升检测精度，如上文提到的 FCOS 通过 Focal Loss 损失函数降低了易分类样本的权重，使模型更加关注难分类样本，进而提升对小目标检测精度。

（4）多模态信息融合

单阶段检测器在不同领域的应用面临不同数据模式之间切换的问题，如对于自动驾驶和无人机应用，探索如何将训练好的检测器迁移到不同的数据模式，如何进行信息融合以提高检测能力。可以使用多模态的数据（如 RGB-D 图像、3D 点云、LIDAR、文本、音频等）进行单阶段检测器训练，通过设置多分支分别处理不同的数据模式进行特征提取，之后通过注意力等融合机制，进一步提高模型的鲁棒性和泛化能力。如上文提到的 Complex-YOLO 就是结合点云三维数据进行的 3D 目标检测。

## 4 结论

本文从基于深度学习的单阶段通用目标检测算法出发，系统地梳理了单阶段通用目标检测器的进展。具体来讲，首先，本文回顾了早期目标检测发展历程，对现有目标检测通用数据集 PASCAL VOC 与 MS COCO 等进行了归纳整理，分析了目标检测领域各类评价指标的作用。其次，详细介绍了 SSD 算法、RetinaNet 算法、RefineDet 算法、CornerNet 算法、CenterNet 算法、FCOS 算法、EfficientDet 算法等七个核心算法，分析了各个检测器的特点优势、局限性及其应用场景。最后，提出了未来可研究方向。

## References


[1] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. SSD: Single Shot MultiBox Detector. In: Proceedings of the European Conference on Computer Vision (Conference Proceedings style), Amsterdam, Netherlands: Springer International Publishing, 2016. 21—37.

[2] Tan M, Pang R, Le Q V. EfficientDet: Scalable and Efficient Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (Conference Proceedings style), Seattle, USA: IEEE Press, 2020. 10778—10787.

[3] Ran B, Boyce D E. *Modeling Dynamic Transportation Network* (Book style). Berlin: Springer-Verlag, 1996, 69--83

[4] Wang F Y, Fundamental Issues in Research of Computing with Words and Linguistic Dynamic Systems. *Acta Automatica Sinica* (Periodical style), 2005, **31**(6): 844--852

[5] Roychoudhury R, Bandyopadhyay S, Paul K. Adistributed mechanism for topology discovery in ad hoc wireless networks using mobile agents. In: *Proceeding of IEEE First Annual Workshop on Mobile and Ad hoc Networking and Computing* (Conference Proceedings style), Piscataway, USA: IEEE Press, 2000. 145--146

[6] Hryniewicz O. An evaluation of the reliability of complex systems using shadowed sets and fuzzy lifetime data. *International Journal of Automation and Computing* (Periodical style—Accepted for publication), to be published.

[7] Zhang W. *Reinforcement Learning for Job-Shop Scheduling*. [Ph.D.Dissertation], Oregon State University, 1996

[8] *IEEE Criteria for Class IE Electric Systems* (Standards style), IEEE Standard 308, 1969.

[9] Jones J. *Networks*, 2nd ed. (Online Sources style). [Online], available: http://www.atm.com , May 10, 1991.

[10] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 734-750.

[11] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv preprint arXiv:1904.07850, 2019.

[12] TIAN Z, SHEN C, CHEN H, et al. FCOS: fully convolu tional one-stage object detection[EB/OL]. [2024-10-13]. https://arxiv.org/abs/1904.01355.