# 级联 R-CNN：致力于高质量目标检测

蔡兆伟加州
大学圣地亚
哥分校
zwcai@ucsd.edu

Nuno
Vasconcelos UC
圣地亚哥
nuno@ucsd.edu

## 摘 要

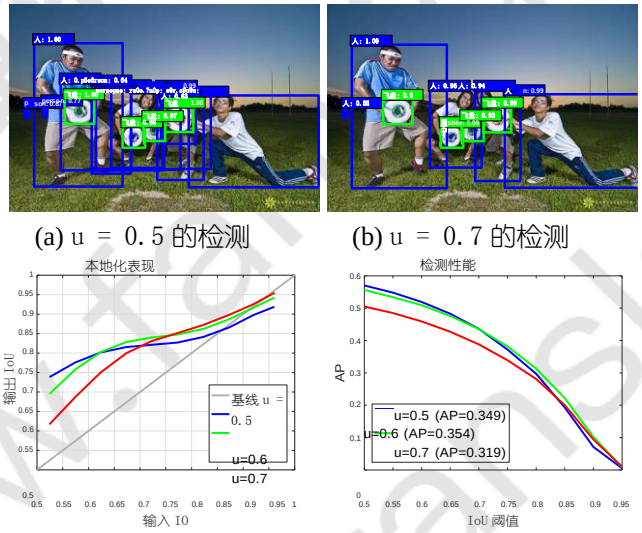
*在对象检测中，需要定义联合正负（IoU）阈值。用低 IoU 阈值（例如 0.5）训练的对象检测器通常会产生嘈杂的检测。但是，随着 IoU 阈值的增加，检测性能趋于下降。造成这种情况的主要原因有两个：1）在训练期间由于正样本呈指数消失而过度拟合，以及 2）探测器最佳的 IoU 与输入假设的 IoU 之间的推理时间不匹配。为了解决这些问题，提出了一种多阶段目标检测架构 Cascade R-CNN。它由一系列经过训练的检测器组成*

*增加 IoU 阈值，以便顺序选择更多
避免误报。对检测器进行逐步培训，利用观察结果，*

*检测器的放置是训练下一个更高质量检测器的良好分布。对经过改进的假设进行重新采样可确保所有检测器都有一组等效尺寸的正例，从而减少了过拟合问题。推理时采用相同的级联过程，从而使假设与每个阶段的检测器质量之间更紧密地匹配。显示了 Cascade R-CNN 的简单实现，可以超越具有挑战性的 COCO 数据集上的所有单模型对象检测器。实验还表明，Cas-cade R-CNN 可广泛应用于检测器架构，获得稳定的增益，而与基线检测器的强度无关。该代码将在 https://github.com/zhaoweicai/cascade-rcnn 提供。*

## 1. 介绍

对象检测是一个复杂的问题，需要解决两个主要任务。首先，检测器必须解决识别问题，以区分前景对象和背景，并为其分配适当的对象类别标签。其次，探测器必须解决定位问题，以便为不同的对象分配准确的边界框。由于检测器面对




(a) u = 0.5 的检测　(b) u = 0.7 的检测

(c) 回归器　(d) 探测器

图 1. 检测输出，定位和检测性能
IoU 阈值 u 增加的物体检测器的数量。

许多“接近”误报，对应于“接近但不正确”的边界框。检测器必须在抑制这些接近的假阳性的同时找到真阳性。许多最近提出的物体检测器都是基于两级 R-CNN 框架[12，11，27，21]，在哪里检测被构造为结合了分类和边界框回归的多任务学习问题。与对象识别不同的是，定义正/负需要交集（IoU）阈值。但是，通常使用的阈值 u（通常为 u = 0.5）对正数建立了相当宽松的要求。生成的检测器经常产生嘈杂的边界框，如图 1 所示。1 (a).大多数人会认为假阳性的假说经常会消失
IoU≥0.5 测试。尽管根据 u = 0.5 准则组装的示例丰富多样，但它们使
很难训练可以有效地拒绝近似误报的检测器。

在这项工作中，我们将假设的质量定义为具有基本事实的 IoU，将检测器的质量定义为用于训练它的 IoU 阈值。目的是调查-

迄今为止，门禁研究了学习高质量目标检测器的研究不足的问题，该目标检测器的输出几乎没有误报，如图 2 所示。1 (b).基本思想是，单个检测器只能针对单个质量级别进行优化。这在对成本敏感的学习文献中是众所周知的[7，24]，其中对接收器工作特性（ROC）的不同点的优化需要不同的损失函数。主要区别在于我们考虑针对给定 IoU 阈值的优化，而不是误报率。

图中说明了这个想法 1 （c）和（d）分别显示了三个训练有 IoU 阈值 u = 0.5、0.6、0.7 的检测器的定位和检测性能。如 COCO [20]所示，根据输入建议的 IoU 评估本地化性能，并根据 IoU 阈值评估检测性能。注意，在图中 1 （c），对于接近训练探测器的阈值的 IoU，每个包围盒回归器表现最佳。这也适用于检测性能，甚至过拟合。数字 1 （d）表明，对于低 IoU 实例，u = 0.5 的检测器性能优于 u = 0.6 的检测器，在较高的 IoU 水平下其性能却不如它。通常，在单个 IoU 级别优化的检测器在其他级别不一定是最佳的。这些观察结果表明，更高质量的检测需要检测器与其处理的假设之间的质量匹配更紧密。通常，只有在提出高质量建议时，检测器才能具有高质量。

然而，为了生产高质量的检测器，仅在训练期间增加 u 是不够的。实际上，如从图的 u = 0.7 的检测器中所见 1 （d），这会降低检测性能。问题在于，从提议检测器发出的假设分布通常在质量低下严重失衡。通常，强迫更大的 IoU 阈值会导致正训练样本数量成倍减少。对于神经网络而言，这尤其成问题，因为神经网络被认为是非常密集的示例，并且使得“高 u”训练策略非常容易过度拟合。另一个困难是推断时检测器质量和测试假设质量之间的不匹配。如图所示 1，高质量检测器仅是针对高质量假设的最佳选择。当要求他们根据其他质量水平的假设进行检测时，检测可能不是最佳的。

在本文中，我们提出了一种解决这些问题的新型检测器架构 Cascade R-CNN。它是 R-CNN 的多级扩展，其中级联更深的检测器级依次对接近的假阳性更具选择性。依次训练 R-CNN 阶段的级联，使用一个阶段的输出来训练下一个阶段。这是由于观察者的动机，即回归器的输出 IoU 几乎总是好于输入 IoU。这个观察可以在图 1 （c），所有地块均在灰线上方。这表明以一定 IoU 阈值训练的检测器的输出是良好的分布，可以训练下一个更高的 IoU 阈值的检测器。这类似于对象检测文献[31，8]中通常用于组合数据集的增强方法。主要区别在于，Cascade R-CNN 的重采样过程并非旨在挖掘硬底片。相反，通过调整边界框，每个阶段的目的是找到一组良好的近似误报，以训练下一阶段。当以这种方式操作时，一系列适用于越来越高的 IoU 的检测器可以克服过拟合的问题，因此可以有效地进行训练。推断时，将应用相同的级联过程。逐步改进的假设在每个阶段都与不断提高的检测器质量更好地匹配。如图所示，这可以实现更高的检测精度。1 （c）和（d）。

Cascade R-CNN 的实施和端到端培训非常简单。我们的结果表明，在具有挑战性的 COCO 检测任务[20]方面，没有任何风吹草动的香草实现比以前的所有现有技术先进的单模式检测器大了很多。 -特别是在较高质量的评估指标下。此外，Cascade R-CNN 可以用任何两阶段构建
基于 R-CNN 框架的目标检测器。我们已经观察到了一致的收益（2-4 点），计算量略有增加。该增益与基线物体检测器的强度无关。因此，我们认为，这种简单有效的检测体系结构可能会引起许多对象检测研究的兴趣。

## 2. 相关工作

由于 R-CNN [12]体系结构的成功，通过结合提议检测器和区域分类器，检测问题的两阶段表述在最近变得很重要。为了减少 R-CNN 中多余的 CNN 计算，SPP-Net [15]和 Fast-RCNN [11]引入了区域特征提取的思想，显着加快了整体检测器的速度。后来，Faster-RCNN [27]通过引入区域提议网络（RPN）进一步提高了速度。这种架构已成为领先的目标检测框架。最近的一些工作已将其扩展为解决各种细节问题。例如，R-FCN [4]提出了有效的区域级全卷积，而没有精度损失，以避免 Faster-RCNN 的繁重的区域级 CNN 计算。而 MS-CNN [1]和 FPN [21]在多个输出层检测提议，以减轻 RPN 接收字段与实际对象大小之间的比例失配，以进行高召回率的提议检测。

另外，一级目标检测架构也已变得很流行，这主要是由于它们的计算能力。

国民效率。这些架构接近经典的滑动窗口策略[31，8]。YOLO [26]通过转发输入图像一次，输出非常稀疏的检测结果。当通过高效的骨干网络实现时，它可以以合理的性能实现实时对象检测。

SSD [23]以类似于 RPN [27]的方式检测物体，但使用分辨率不同的多个要素贴图来覆盖各种比例的对象。这些架构的主要局限在于其精度通常低于两级检测器的精度。最近，RetinaNet [22 为了解决密集物体检测中极端的前景-背景类别不平衡问题，建议使用[]来达到比最新的两级物体检测器更好的结果。

多阶段目标检测的一些探索有也提出了。多区域检测器[9]引入了迭代边界框回归，其中多次应用了 R-CNN，以生成更好的边界框。CRAFT [33]和 AttractioNet [10]使用了多阶段程序来生成准确的建议，并将其转发给 Fast-RCNN。[19，25]在对象检测网络中嵌入了[31]的经典级联架构。[3]交替进行检测和分割任务，例如分割。

## 3. 物体检测

在本文中，我们扩展了 Faster-RCNN 的两阶段架构[27， 21]，如图所示 3 (a).第一阶段是提议子网（“ H0”），应用于整个图像，以产生初步的检测假设，称为目标提议。在第二阶段，这些假设随后由感兴趣区域检测子网（“ H1”）进行处理，表示为检测头。最终分类得分（“ C”）和边界框（“ B”）被分配给每个假设。我们专注于对多级检测子网建模，并采用但不限于 RPN [27]进行提议检测。

### 3.1. 边界框回归

边界框 b = (bx，by，bw，bh）包含图像块 x 的四个坐标。边界框回归的任务是使用回归器 f (x，b）将候选边界框 b 回归到目标边界框 g。这是从训练样本{$g_i$，bi}中学到的，以便将边界框风险降至最低

$$R_{loc}[f] = \sum_{i=1}^{N} Lloc(f(xi, bi), g_i), \quad (1)$$

其中 Lloc 是 R-CNN 中的 L2 损失函数[12]，但已更新为 Fast-RCNN 中的平滑 L1 损失函数[11]。为了鼓励回归不变尺度和位置，Lloc 对距离矢量Δ = (δx，δy， δw， δh）进行运算

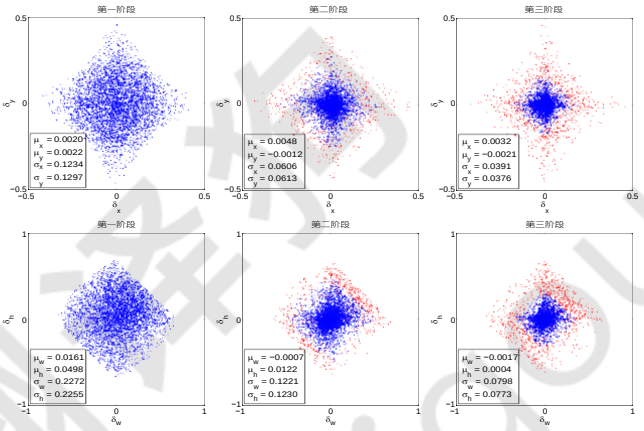



图 2. 不同级联阶段的顺序Δ分布（未归一化）。使用增加的 IoU 阈值时，红点是异常值，并且在去除异常值后会获得统计信息。

被定义为

$$\delta x = (gx - bx) / bw, \qquad \delta_y = (g_y - b_y)/b_h$$
$$\delta_w = \log(g_w / b_w), \qquad \delta_h = \log(g_h / b_h). \quad (2)$$

由于包围盒回归通常会对 b 执行较小的调整，因此（2）可以很小。因此，（1）通常比分类风险小得多。为了提高多任务学习的有效性，通常将Δ的均值归一化
方差，即 δx 被 δ′ = (δx-μx) /σx 代替。这在文献中被广泛使用[27，1，4，21，14].

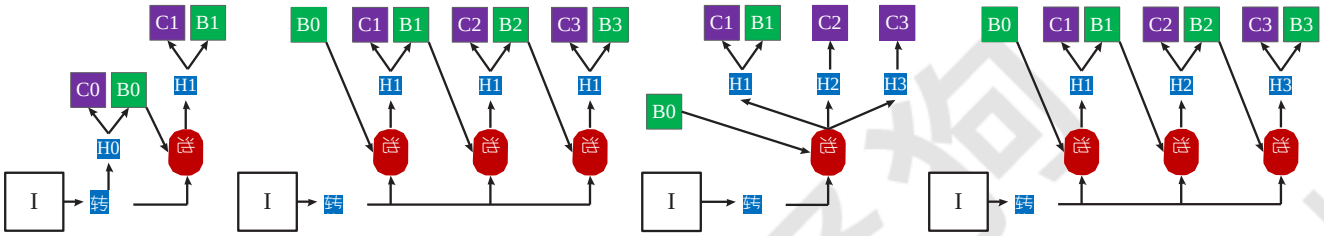
一些著作[9，10，16]认为 f 的单个回归步骤不足以进行精确定位。而是将 f 迭代地应用，作为后处理步骤

$$f'(x,b) = f \circ f \circ \cdots \circ f(x,b), \quad (3)$$

完善边界框 b。这称为迭代边界框回归，表示为迭代 BBox。可以用图的推理架构来实现 3 (b）所有首长均相同的地方。但是，这个想法忽略了两个问题。首先，如图 1，在 u = 0.5 时训练的回归因子 f 对于较高 IoU 的假设是次优的。实际上，它会使大于 0.85 的 IoU 边界框降级。二，如图 2，边界框的分布在每次迭代后都会发生显着变化。虽然回归变量对于初始分布是最佳的，但是在那之后它可能是次优的。由于这些问题，迭代的 BBox 需要大量的人工工程，包括提案累积，投票表决等形式[9，10，16]，并且收益有些不可靠。通常，除了两次应用 f 之外没有其他好处。

### 3.2. 分类

分类器是一个函数 h (x)，它将图像补丁 x 分配给 M + 1 个类之一，其中类 0 包含




(a) 更快的 R-CNN (b) 推理中的迭代 BBox (c) 积分损失 (d) 级联 R-CNN

图 3. 不同框架的架构。“ I”是输入图像，“ conv”主干卷积，“ pool”区域特征提取，“ H”网络头，“ B”边界框和“ C”分类。“ B0”是所有体系结构中的建议。

背景和其余要检测的对象。h（x）是
后分布的 M +1 维估计
以上类，即 hk（x）= p（y = k | x），其中 y 是
类标签。给定一个训练集（xi，yi），它是由
最小化分类风险

$$R_{cls[h]} = \sum_{i=1}^{N} Lcls(h(xi), yi), \quad (4)$$

其中 Lcls 是经典的交叉熵损失。

### 3.3. 检测质量

由于边界框通常包含一个对象和一定数量的背景，因此很难确定检测结果是阳性还是阴性。这通常通过 IoU 指标解决。如果 IoU 高于阈值 u，则将补丁视为该类的示例。因此，假设 x 的类别标签是 u 的函数，

$$y = \begin{cases} gy, & IoU(x,g) \geq u \\ 0, & \text{除此以外} \end{cases} \quad (5)$$

其中 gy 是地面真实物体 g 的类别标签。此 IoU 阈值 u 定义了检测器的质量。

对象检测具有挑战性，因为无论阈值如何，检测设置都具有很高的对抗性。当 u 高时，积极因素包含较少的背景知识，但是很难收集足够的积极培训示例。当 u 低时，可以使用更丰富，更多样化的阳性训练集，但是训练有素的检测器几乎没有动力拒绝接近的假阳性。通常，很难要求单个分类器在所有 IoU 级别上均一地表现良好。根据推断，由于提议检测器（例如 RPN [27]或选择性搜索[30]）产生的大多数假设的质量较低，因此对于较低质量的假设，检测器必须更具判别力。这些相互矛盾的要求之间的标准折衷方案是使 u = 0.5。但是，这是一个相对较低的阈值，从而导致低质量的检测，大多数人认为这些检测是假阳性，如图所示。1 (a).

朴素的解决方案是使用 Figure 的体系结构开发一组分类器。3 （c），有损失地进行了优化

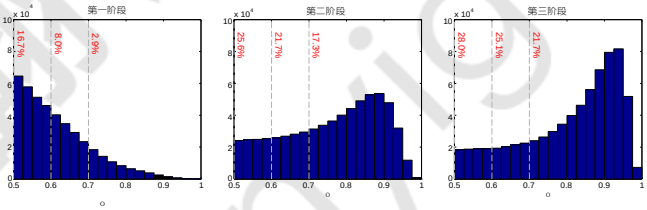



图 4. 训练样本的 IoU 直方图。分布
第一阶段是 RPN 的输出。红色数字比相应的 IoU 阈值高正百分比。

针对各种质量水平，

$$Lcls(h(x), y) = \sum_{u \in U} Lcls(hu(x), yu), \quad (6)$$

其中 U 是一组 IoU 阈值。 这与[34]的积分损失密切相关，其中 U = {0.5，0.55，●●●，0.75}，旨在适应 COCO 挑战的评估指标。根据定义，分类器需要
归纳为推理。该解决方案无法解决以下问题：6）对不同数量的正数进行运算。如图第一图所示 4，一组正样本随 u 迅速减少。这尤其成问题，因为高质量的分类器容易过拟合。另外，那些高质量的分类器被要求在推论中处理压倒性的低质量的建议，而这些建议并未得到优化。由于所有这些，所以（6）在大多数质量水平上都无法获得更高的精度，并且该结构与图 1 相比几乎没有收益。3 (a).

## 4. 级联 R-CNN

在本节中，我们介绍图的建议的 Cascade R-CNN 对象检测架构。3(d).

### 4.1. 级联边界框回归

如图所示 1 （c），很难要求单个回归器在所有质量水平上都能完美地均匀地执行。受 Cas-cade 姿势回归[6]和人脸对齐[2]的启发，可以将困难的回归任务分解为一系列更简单的步骤 32]. 在里面

级联 R-CNN，它被构造为级联回归问题，其结构如图 1 所示。3 (d).这依赖于一连串的专业回归器

$$f(x, b) = fT \circ fT-1 \circ \cdots \circ f1(x, b), \quad (7)$$

其中 T 是级联总数。注意

级联中的每个回归变量 ft 均会通过 sam-
ple 分布{b^t}到达相应的阶段，在-

代替{b}的初始分布。此级联 IM[1]
逐步证明假设。

它与图的迭代 BBox 体系结构不同 3 （b）有几种方式。首先，虽然迭代 BBox 是用于改善边界框的后处理过程，但是级联回归是一种重采样过程，可以更改要在不同阶段处理的假设的分布。其次，由于它既用于训练又用于推理，因此在训练和推理分布之间没有差异。第三，多重专业规则-
sors {fT, fT -1, ● ● ●, f1}针对不同阶段的重采样分布进行了优化。这与
（的单个 f3），这仅适用于初始分布。这些差异可以实现比迭代 BBox 更精确的定位，而无需进行进一步的人工操作。

如本节所述 3.1，$\Delta = (\delta x, \delta y, \delta w, \delta h)$ 在 (2 有效的多任务学习需要通过其均值和方差进行归一化。在每个回归阶段之后，这些统计信息将顺序发展，如图所示。2.在训练中，相应的统计数据用于标准化每个阶段的Δ。

### 4.2. 级联检测

如图左图 4，初始假设（例如 RPN 提案）的分布严重倾向于低质量。这不可避免地导致对高质量分类器的无效学习。Cascade R-CNN 通过将级联回归作为重采样机制来解决该问题。这是由于以下事实引起的：1 （c）所有曲线均在对角灰线上方，
也就是说，针对某个 u 进行训练的包围盒回归器往往会产生更高 IoU 的包围盒。因此，从一系列示例（xi, bi）开始，级联回归成功
对更高的示例分布（x′, b′）进行重新采样

IoU.通过这种方式，可以保持
即使当检测器质量（IoU 阈值）提高时，连续阶段的示例也大致保持不变。如图所示 4，在每个重采样步骤之后，分布会更倾向于高质量的示例。随后有两个后果。首先，没有过度拟合，因为各个级别的示例很多。其次，针对较高的 IoU 阈值优化了较深阶段的检测器。请注意，通过增加 IoU 阈值顺序地去除了一些离群值，如图所示

在图 2，从而可以更好地训练专门的探测器序列。

在每个阶段 t，R-CNN 包括分类器 ht 和针对 IoU 阈值 $u^t$ 优化的回归器 ft，其中 $u^t > u^{t-1}$。通过最小化损失来保证

$$L(x^t, g^t) = Lcls(h^t(x^t), y^t) + \lambda[y^t \geqslant 1] Lloc(f^t(x^t(b, b), g), \quad (8)$$

其中 $b^t = f^{t-1}(x^{t-1}, b^{t-1})$，g 是地面真实物体
对于 $x^t$，$\lambda = 1$ 的折衷系数，[●]指标函数，并且 $y^t$ 是 $x^t$ 赋予 $u^t$ 的标签，由 (5).不像
(的积分损失 6)，这样可以保证一系列经过有效训练的检测器质量不断提高。推论上，通过应用相同的级联过程，假设的质量得到了依次改善，并且仅需要更高质量的检测器即可对更高质量的假设进行操作。如图所示，这可以实现高质量的目标检测。1 （c）和（d）。

## 5. 实验结果

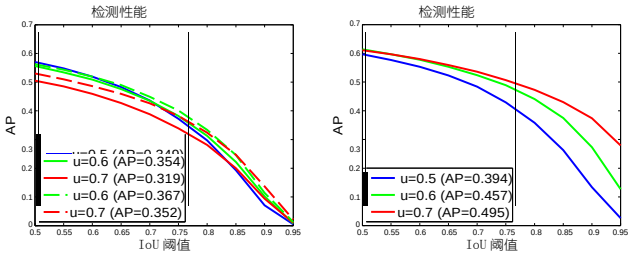
在 MS-COCO 2017 [20]上对 Cascade R-CNN 进行了评估，其中包含约 118k 的图像进行训练，5k 的验证（val）和约 20k 的测试（未提供注释）（test-dev）。COCO 风格的平均精度
（AP）跨 IoU 阈值的平均 AP 从 0.5 到 0.95，间隔为 0.05。这些评估指标可衡量各种质量的检测性能。所有模型都在 COCO 训练集上进行训练，并在 val 集上进行评估。最终结果也报告在测试开发集上。

### 5.1. 实施细节

为了简单起见，所有回归器均与类无关。级联 R-CNN 中的所有级联检测阶段都具有相同的架构，这是基线检测网络的头。Cascade R-CNN 共有四个阶段，一个 RPN
除非另有说明，否则三个用于 U = {0.5, 0.6, 0.7}进行检测。第一检测阶段的采样如下-
最低点[11, 27]。在接下来的阶段中，只需使用先前阶段中的回归输出即可实现重采样。4.2.除了标准水平图像翻转之外，没有使用任何数据增强。推理是在单个图像比例上进行的，没有其他的风吹草动。重新实现了所有基线检测器
与 Caffe [18]，在同一代码库上进行公平比较。

#### 5.1.1 基准网络

为了测试 Cascade R-CNN 的多功能性，我们使用三种流行的基线检测器进行了实验：带有主干 VGG-Net 的 Faster-RCNN [29]，R-FCN [4]和 FPN [21ResNet 骨干[16]。这些基线具有广泛的检测性能。除非另有说明，否则使用其默认设置。使用端到端培训代替了多步培训。




(a) (b)

图 5. (a) 是经过单独训练的检测器的检测性能，具有自己的建议（实线）或 Cascade R-CNN 阶段建议（虚线），并且（b）是通过将基本事实添加到建议集中。

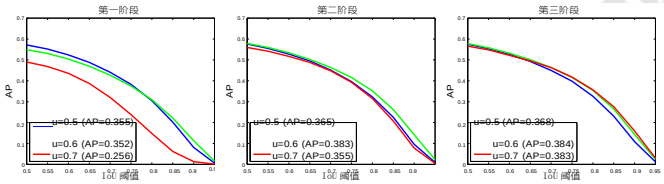



图 6. 在所有级联级的所有 Cascade R-CNN 检测器的检测性能。

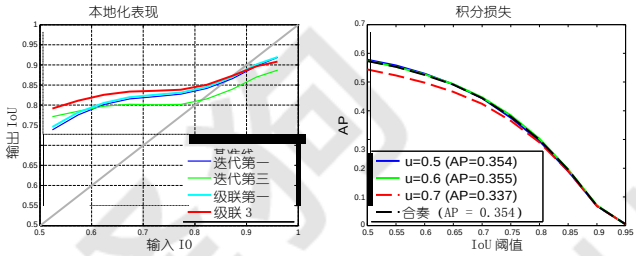
Faster-RCNN：网络头具有两个完全连接的层。为了减少参数，我们使用[13]修剪不太重要的连接。每个完全连接的层保留 2048 个单元，并删除掉落层。在 2 个同步的 GPU 上，训练以 0.002 的学习率开始，在 60k 和 90k 迭代中减少了 10 倍，并在 100k 迭代中停止，每个 GPU 每次迭代包含 4 张图像。每个图像使用 128 个 RoI。

R-FCN：R-FCN 向 ResNet 添加了卷积，边界框回归和分类层。级联 R-CNN 的所有头都具有此结构。没有使用在线硬负挖矿[28]。训练以 0.003 的学习率开始，在 4 个同步的 GPU 上，在 160k 和 240k 迭代中，学习率降低了 10 倍，在 280k 迭代中停止了学习，每个 GPU 每次迭代都保留一张图像。每个图像使用 256 个 RoI。

FPN：由于尚未公开可用于 FPN 的源代码，因此我们的实现细节可能会有所不同。RoIAlign [14]用于获得更强的基准。这被表示为 FPN +，并用于所有消融研究中。与往常一样，ResNet-50 用于消融研究，ResNet-101 用于最终检测。培训使用的学习率是
在 8 个同步的 GPU 上，每 120k 迭代 0.005，接下来的 60k 迭代 0.0005，每个 GPU 每次迭代保存一张图像。每个图像使用 256 个 RoI。

## 5.2. 质量不匹配

数字 5 (a) 显示了三个单独训练的检测器的 AP 曲线，这些检测器的 IoU 阈值提高了 U = {0.5, 0.6, 0.7}. 在低 IoU 水平下，u = 0.5 的检测器性能优于 u = 0.6 的检测器，但在较高水平时，性能却不佳。但是，u = 0.7 的检测器性能低于形成其他两个。要了解为什么会发生这种情况，我们推论改变了提案的质量。数字 5 (b) 显示当地面真理受约束时所获得的结果—盒子被添加到提案集中。虽然所有检波器性能提高，u = 0.7 的检波器具有最大增益，几乎在所有 IoU 电平下均达到最佳性能。这些结果表明了两个结论。首先，对于精确检测而言，u = 0.5 并不是一个好的选择，只是对于低质量的建议而言，它更容易被淘汰。其次，高精度检测需要符合检测器质量的假设。接下来，将原始检测器建议替换为质量更高的 Cascade R-CNN 建议（分别使用第二和第三阶段建议的 u = 0.6 和 u = 0.7）。数字 5 (a) 还表明，当测试提案更接近探测器质量时，两个探测器的性能将得到显着改善。




(a) (b)

图 7. (a) 是定位比较，(b) 是积分损耗检测器中各个分类器的检测性能。

在所有级联阶段对所有 Cascade R-CNN 检测器进行测试，得出的观察结果相似。数字 6 结果表明，当使用更精确的假设时，每个检测器都得到了改进，而质量更高的检测器具有更大的增益。例如，对于第一阶段的低质量建议，u = 0.7 的检测器性能较差，而对于更深的级联阶段可用的更精确的假设，则性能要好得多。此外，图的联合训练探测器 6 胜过图 1 的单独训练的探测器 5 (a)，即使使用了相同的建议。这表明在 Cascade R-CNN 框架内对探测器进行了更好的训练。

## 5.3. 与*迭代 BBox 和积分损失的比较*

在本节中，我们将 Cascade R-CNN 与迭代 BBox 和积分损耗检测器进行比较。迭代 BBox 是通过反复应用 FPN +基线三次来实现的。积分损耗检测器的分类头数与 Cascade R-CNN 相同，具有 $U = \{0.5, 0.6, 0.7\}$.

| | AP | AP50 | AP60 | AP70 | AP80 | AP90 |
|---|---|---|---|---|---|---|
| FPN +基线 | 34.9 | 57.0 | 51.9 | 43.6 | 29.7 | 7.1 |
| *迭代 BBox* | 35.4 | 57.2 | 52.1 | 44.2 | 30.4 | 8.1 |
| *积分损失* | 35.4 | 57.3 | 52.5 | 44.4 | 29.9 | 6.9 |
| 级联 R-CNN | 38.9 | 57.8 | 53.4 | 46.9 | 35.8 | 15.8 |

表 *1. 与迭代 BBox 和积分损失的比较。*

| 测试阶段 | AP | AP50 | AP60 | AP70 | AP80 | AP90 |
|---|---|---|---|---|---|---|
| 1 | 35.5 | 57.2 | 52.4 | 44.1 | 30.5 | 8.1 |
| 2 | 38.3 | 57.9 | 53.4 | 46.4 | 35.2 | 14.2 |
| 3 | 38.3 | 56.6 | 52.2 | 46.3 | 35.7 | 15.9 |
| $\overline{1 \sim 2}$ | 38.5 | 58.2 | 53.8 | 46.7 | 35.0 | 14.0 |
| $\overline{1 \sim 3}$ | 38.9 | 57.8 | 53.4 | 46.9 | 35.8 | 15.8 |
| FPN +基准 34.9 | | 57.0 | 51.9 | 43.6 | 29.7 | 7.1 |

表 2. Cascade R-CNN 的舞台性能。1∽3 表示第三阶段提案中三个分类器的集合。

本地化：图中比较了级联回归和迭代 BBox 的本地化性能 7 (a).使用单个回归变量会降低高 IoU 假设的局域性。当像迭代 BBox 那样迭代应用回归器时，此效果会累积，而性能实际上会下降。请注意，经过 3 次迭代后，迭代 BBox 的性能非常差。相反，级联回归器在后期阶段具有更好的性能，在几乎所有 IoU 级别上都优于迭代 BBox。

积分损耗：积分损耗检测器中所有分类器的检测性能（共享一个回归变量）如图 1 所示。7 (b).在所有 IoU 级别上，u = 0.6 的分类器是最佳的，而 u = 0.7 的分类器是最差的。所有分类器的集合均未显示可见增益。表 1 如图所示，迭代 BBox 和积分损耗检测器均会稍微改善基线检测器。级联 R-CNN 在所有评估指标上均具有最佳性能。对于较低的 IoU 阈值，增益是温和的，但是对于较高者来说很重要。

## 5.4. 消融实验

还进行了消融实验。

阶段比较：表格 2 总结舞台表演。由于多阶段多任务学习的优势，第一阶段的性能已经超过了基线检测器。第二阶段实质上提高了性能，第三阶段与第二阶段等效。这与整体损耗检测器不同，后者较高的 IOU 分类器相对较弱。虽然在较低（较高）IoU 指标下，前一个（后期）阶段比较好，但所有分类器的整体效果最好。

IoU 阈值：对于所有头部，均使用相同的 IoU 阈值 u = 0.5 训练了初步的 Cascade R-CNN。在这种情况下，各个阶段的区别仅在于它们的假设

| IoU↑ | *统计* | AP | AP50 | AP60 | AP70 | AP80 | AP90 |
|---|---|---|---|---|---|---|---|
| | | 36.8 | 57.8 | 52.9 | 45.4 | 32.0 | 10.7 |
| ✓ | | 38.5 | 58.4 | 54.1 | 47.1 | 35.0 | 13.1 |
| | ✓ | 37.5 | 57.8 | 53.1 | 45.5 | 33.3 | 13.1 |
| ✓ | ✓ | 38.9 | 57.8 | 53.4 | 46.9 | 35.8 | 15.8 |

表 3。消融实验。“ IoU↑”表示增加 IoU 阈值，“ stat”表示采用顺序回归统计。

| #阶段 | 测试阶段 | AP | AP50 | AP60 | AP70 | AP80 | AP90 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 34.9 | 57.0 | 51.9 | 43.6 | 29.7 | 7.1 |
| 2 | $\overline{1 \sim 2}$ | 38.2 | 58.0 | 53.6 | 46.7 | 34.6 | 13.6 |
| 3 | $\overline{1 \sim 3}$ | 38.9 | 57.8 | 53.4 | 46.9 | 35.8 | 15.8 |
| 4 | $\overline{1 \sim 3}$ | 38.9 | 57.4 | 53.2 | 46.8 | 36.0 | 16.0 |
| 4 | $\overline{1 \sim 4}$ | 38.6 | 57.2 | 52.8 | 46.2 | 35.5 | 16.3 |

表 4. 级联 R-CNN 中阶段数的影响。

接收。每个阶段都用相应的假设进行训练，即考虑到图的分布 2.表的第一行 3 显示了在基线检测器上级联得到改善。这表明对于相应的样本分布，优化阶段的重要性。第二行显示，通过增加阶段阈值 u，可以使检测器针对接近的假阳性更具选择性，并专门用于更精确的假设，从而带来额外的收益。这支持本节的结论 4.2.

回归统计：利用图的逐步更新的回归统计 2，有助于有效的多任务分类和回归学习。通过在 Ta-ble 中比较有/没有模型，可以注意到其优点 3.学习对这些统计数据不敏感。

阶段数：表中汇总了阶段数的影响 4.添加第二个检测阶段会大大改善基线检测器。三个检测阶段仍可带来不小的改进，但增加第四个阶段（u = 0.75）会导致性能略有下降。但是请注意，虽然 AP 的整体性能下降，但四阶段级联对于高 IoU 级别具有最佳性能。三级级联实现了最佳折衷。

## 5.5. 与最新技术的比较

在表中将基于 FPN +和 ResNet-101 主干的 Cascade R-CNN 与最先进的单模型对象检测器进行了比较 5.设置如章节中所述 5.1.1，但总共进行了 280k 训练迭代，学习率下降到 160k 和 240k 迭代。RoI 的数量也增加到 512。表中的第一组检测器 5 是一级检测器，第二组是两级检测器，最后一组是多级检测器（对于 Cascade R-CNN，是三级检测器+ RPN）。所有比较过的最新探测器都经过 u = 0.5 训练。它是

| | 骨干 | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|
| YOLOv2 [26] | DarkNet- | 21.6 | 44.0 | 19.2 | 5.0 | 22.4 | 35.5 |
| SSD513 [23] | ResNet-101 | 31.2 | 50.4 | 33.3 | 10.2 | 34.5 | 49.8 |
| 视网膜网[22] | ResNet-101 | 39.1 | 59.1 | 42.3 | 21.8 | 42.7 | 50.2 |
| 更快的 R-CNN +++ [16] * | ResNet-101 | 34.9 | 55.7 | 37.4 | 15.6 | 38.7 | 50.9 |
| 带 FPN 的 R-CNN 更快[21] | ResNet-101 | 36.2 | 59.1 | 39.0 | 18.2 | 39.0 | 48.2 |
| 更快的 R-CNN 和 FPN + (我 | ResNet-101 | 38.8 | 61.1 | 41.9 | 21.3 | 41.8 | 49.8 |
| 通过 G-RMI 更快地进行 R- | Inception-ResNet-v2 | 34.7 | 55.5 | 36.7 | 13.5 | 38.1 | 52.0 |
| 可变形的 R-FCN [5]* | 对齐初始资源网 | 37.5 | 58.0 | 40.8 | 19.4 | 40.1 | 52.5 |
| 遮罩 R-CNN [14] | ResNet-101 | 38.2 | 60.3 | 41.7 | 20.1 | 41.1 | 50.2 |
| AttractioNet [10]* | VGG16 +宽 ResNet | 35.7 | 53.4 | 39.3 | 15.6 | 38.0 | 52.7 |
| 级联 R-CNN | ResNet-101 | 42.8 | 62.1 | 46.3 | 23.7 | 45.5 | 55.2 |

表 5. 与 COCO test-dev 上最新的单模型检测器的比较。用 “ *”表示的条目在推断时使用了钟声。

| | 骨干级联 | | 培养试 | 测 | 参数 | 值 | | | | | | 测试开发 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
| 更快的 R-CNN | VGG | ✗ | 0.12s | 0.075s | 278M | 23.6 | 43.9 | 23.0 | 8.0 | 26.2 | 35.5 | 23.5 | 43.9 | 22.6 | 8.1 | 25.1 | 34.7 |
| | | ✓ | 0.14s | 0.115s | 704M | 27.0 | 44.2 | 27.7 | 8.6 | 29.1 | 42.2 | 26.9 | 44.3 | 27.8 | 8.3 | 28.2 | 41.1 |
| – | ResNet-50 | ✗ | 0.19s | 0.07s | 133M | 27.0 | 48.7 | 26.9 | 9.8 | 30.9 | 40.3 | 27.1 | 49.0 | 26.9 | 10.4 | 29.7 | 39.2 |
| | | ✓ | 0.24s | 0.075s | 184M | 31.1 | 49.8 | 32.8 | 10.4 | 34.4 | 48.5 | 30.9 | 49.9 | 32.6 | 10.5 | 33.1 | 46.9 |
| – | ResNet-101 | ✗ | 0.23s | 0.075s | 206M | 30.3 | 52.2 | 30.8 | 12.0 | 34.7 | 44.3 | 30.5 | 52.9 | 31.2 | 12.0 | 33.9 | 43.8 |
| | | ✓ | 0.29s | 0.083s | 256M | 33.3 | 52.0 | 35.2 | 11.8 | 37.2 | 51.1 | 33.3 | 52.6 | 35.2 | 12.1 | 36.2 | 49.3 |
| FPN+ | ResNet-50 | ✗ | 0.30s | 0.095s | 165M | 36.5 | 58.6 | 39.2 | 20.8 | 40.0 | 47.8 | 36.5 | 59.0 | 39.2 | 20.3 | 38.8 | 46.4 |
| | | ✓ | 0.33s | 0.115s | 272M | 40.3 | 59.4 | 43.7 | 22.9 | 43.7 | 54.1 | 40.6 | 59.9 | 44.0 | 22.6 | 42.7 | 52.1 |
| FPN+ | ResNet-101 | ✗ | 0.38s | 0.115s | 238M | 38.5 | 60.6 | 41.7 | 22.1 | 41.9 | 51.1 | 38.8 | 61.1 | 41.9 | 21.3 | 41.8 | 49.8 |
| | | ✓ | 0.41s | 0.14s | 345M | 42.7 | 61.6 | 46.6 | 23.8 | 46.2 | 57.4 | 42.8 | 62.1 | 46.3 | 23.7 | 45.5 | 55.2 |

表 6. 多个流行的基线目标检测器的详细比较。在单个 Titan Xp GPU 上每个图像报告所有速度。

注意到我们的 FPN +实施比原始 FPN 更好[21]，提供非常强的基准。此外，
从 FPN +扩展到 Cascade R-CNN 的性能提高了约 4 点。Cascade R-CNN 的性能也大大优于所有单模检测器，评估所有评估指标。这包括 2015 年和 2016 年 COCO 挑战赛获奖者的单模型参赛作品（更快的 R-CNN +++ [16]和 G-RMI [17]），以及最近的可变形 R-FCN [5]，RetinaNet [22]和 Mask R-CNN [14]。在 COCO 上最好的多级检测器 At-tractioNet [10]使用迭代 BBox 生成建议。尽管 Attrac-tioNet 使用了许多增强功能，但香草 Cascade R-CNN 的性能仍优于
7.1 分请注意，与 Mask R-CNN 不同，在 Cascade R-CNN 中没有利用分段信息。最后，香草单模 Cascade R-CNN 还超越了在 2015 年和 2016 年赢得 COCO 挑战的严格设计的集成探测器（分别为 AP 37.4 和 41.6）[1]。

### 5.6. 泛化能力

表中比较了所有三个基线检测器的三级级联 R-CNN6. 所有设置均与上述相同，更改了 Section5.5 对于 FPN +。

检测性能：同样，我们的实现优于原始检测器[27，4，21]。尽管如此，Cas-cade R-CNN 在这些基线上始终可以提高 2-4 点，而与它们的强度无关。这些增益在 val 和 test-dev 上也一致。这些结果表明，Cascade R-CNN 在检测器体系结构中广泛适用。

参数和时序：级联 R-CNN 参数的数量随级联级数的增加而增加。基线检测头的参数数量呈线性增加。另外，由于与 RPN 相比检测头的计算成本通常较小，因此在训练和测试方面，Cascade R-CNN 的计算开销都较小。

## 6. 结论

在本文中，我们提出了一种多级目标检测框架 Cascade R-CNN，用于设计高质量的目标检测器。事实证明，该体系结构避免了训练时过度拟合和推理时质量不匹配的问题。在具有挑战性的 COCO 数据集上对 Cascade R-CNN 进行了坚实而一致的检测改进，这表明需要对各种并发因素进行建模和理解才能推进物体

[1] http://cocodataset.org/#detections-leaderboard

检测。Cascade R-CNN 被证明可用于许多物体检测架构。我们相信，它对于将来的许多物体检测研究工作很有用。

致谢我们要感谢何凯明的宝贵讨论。

## 参考文献


[1] Z. Cai，Q。Fan，RS Feris 和 N. Vasconcelos。用于快速物体检测的统一多尺度深度卷积神经网络。在 ECCV 中，第 354-370 页，2016 年。2，3

[2]曹 X，魏 Y，温 F，和孙 J。通过示例性形状回归进行面部对齐。在 CVPR 中，第 2887-2894 页，2012 年。5

[3]戴志坚，何 K 和孙坚。通过多任务网络级联的实例感知语义分割。在 CVPR 中，第 3150 至 3158 页，2016 年。3

[4]戴建勋，李玉 He，何凯和孙建勋。R-FCN：通过基于区域的全卷积网络进行对象检测。在 NIPS 中，第 379-387 页，2016 年。2，3，5，8

[5]戴杰，齐海，熊勇，李立，张庚，胡海，魏伟。可变形卷积网络。在 ICCV 中，2017 年。8

[6] P. Dolla´r，P。Welinder 和 P. Perona。级联姿势回归。在 CVPR 中，第 1078–1085 页，2010 年。5

[7] C. Elkan.成本敏感型学习的基础。在 IJ-CAI，第 973–978 页，2001 年。2

[8] PF Felzenszwalb，RB Girshick，DA McAllester 和 D.拉玛南具有区分性训练的基于零件的模型进行对象检测。IEEE Trans.模式肛门。马赫 Intell。，32（9）：1627–1645，2010 年。2，3

[9] S. Gidaris 和 N. Komodakis。通过多区域和语义分段感知的 CNN 模型进行对象检测。在 ICCV 中，第 1134–1142 页，2015 年。3

[10] S. Gidaris 和 N. Komodakis。参加优化重复：通过内到外本地化主动生成提案。在 BMVC 中，2016 年。3，8

[11] RB Girshick.快速 R-CNN。在 ICCV 中，第 1440-1448 页，2015 年。1，2，3，5

[12] RB Girshick，J。Donahue，T。Darrell 和 J. Malik。丰富的功能层次结构，用于准确的对象检测和语义分割。在 CVPR 中，第 580-587 页，2014 年。1，2，3

[13] S. Han，J。Pool，J。Tran 和 WJ Dally。学习权重和连接以建立有效的神经网络。在 NIPS 中，第 1135–1143 页，2015 年。6

[14] K. He，G。Gkioxari，P。Dolla´r 和 R. Girshick。遮罩 r-cnn。在 ICCV 中，2017 年。3，6，8

[15]何凯，张旭，任圣和孙捷。深度卷积网络中的空间金字塔池，用于视觉识别。在 ECCV 中，第 346–361 页，2014 年。2

[16] K. He，X。Zhang，S。Ren 和 J.Sun。深度残差学习，用于图像识别。在 CVPR 中，第 770–778 页，2016 年。3，5，8

[17] J. Huang，V. Rathod，C. Sun，M. Zhu，A. Korattikara，A. Fathi，I。Fischer，Z。Wojna，Y。Song，S。Guadarrama 和 K.墨菲。现代卷积目标检测器的速度/精度权衡。CoRR，abs / 1611.10012，2016 年。8

[18]贾伊，谢尔哈默，多纳休，萨克拉耶夫，朗，RB·吉尔希克，瓜达拉马和塔雷尔。Caffe：用于快速功能嵌入的卷积体系结构。MM，第 675-678 页，2014 年。5

[19] H. Li，Z. Lin，X. Shen，J. Brandt 和 G. Hua。卷积神经网络级联用于面部检测。在 CVPR 中，第 5325-5334 页，2015 年。3

[20] T. Lin，M。Maire，SJ Belongie，J。Hays，P。Perona，D。Ra-manan，P。Dolla´r 和 CL Zitnick。Microsoft COCO：上下文中的公共对象。在 ECCV 中，第 740-755 页，2014 年。2，5

[21] T.-Y.Lin，P.Dolla´r，R.Girshick，K.He，B.Hariharan 和 S.贝隆吉。特征金字塔网络用于物体检测。在 CVPR 中，2017 年。1，2，3，5，8

[22] T.-Y.Lin，P。Goyal，R。Girshick，K。He 和 P. Dolla´r。焦点丢失，用于密集物体检测。在 ICCV 中，2017 年。3，8

[23] W. Liu，D. Anguelov，D. Erhan，C. Szegedy，SE Reed，C. Fu 和 AC Berg。SSD：单发多盒检测器。在 ECCV 中，第 21-37 页，2016 年。3，8

[24] H. Masnadi-Shirazi 和 N. Vasconcelos。成本敏感的提升。IEEE Trans.模式肛门。马赫 Intell。，33（2）：294-309，2011 年。2

[25] W. Ouyang，K。Wang，X。Zhu 和 X. Wang。学习链接的深度特征和分类器，以级联对象检测。CoRR，abs / 1702.07054，2017 年。3

[26] J. Redmon，SK Divvala，RB Girshick 和 A. Farhadi。您只需要看一次：统一的实时对象检测。在 CVPR 中，第 779–788 页，2016 年。3，8

[27] S. Ren，K。He，RB Girshick 和 J.Sun。更快的 R-CNN：通过区域建议网络实现实时目标检测。在 NIPS 中，第 91–99 页，2015 年。1，2，3，4，5，8

[28] A. Shrivastava，A。Gupta 和 RB Girshick。通过在线艰苦的示例挖掘训练基于区域的对象检测器。在 CVPR 中，第 761–769 页，2016 年。6

[29] K. Simonyan 和 A. Zisserman。用于大规模图像识别的超深度卷积网络。CoRR，abs / 1409.1556，2014 年。5

[30] JRR Uijlings，KEA van de Sande，T。Gevers 和 AWM Smeulders.选择性搜索对象识别。国际计算机视觉杂志，104（2）：154-171，2013 年。4

[31] PA Viola 和 MJ Jones。强大的实时面部检测功能。国际计算机视觉杂志，57（2）：137–154，2004 年。2，3

[32]严 J.雷 Z.雷 D.易，和 S.李。学习结合多种假设以实现准确的人脸对齐。在 ICCV 研讨会上，第 392–396 页，2013 年。5

[33] B. Yang，J。Yan，Z。Lei 和 SZ Li。从图像中制作对象。在 CVPR 中，第 6043–6051 页，2016 年。3

[34] S. Zagoruyko，A. Lerer，T. Lin，PO Pinheiro，S. Gross，S. Chintala 和 P. Dolla´r。用于对象检测的多路径网络。在 BMVC 中，2016 年。4