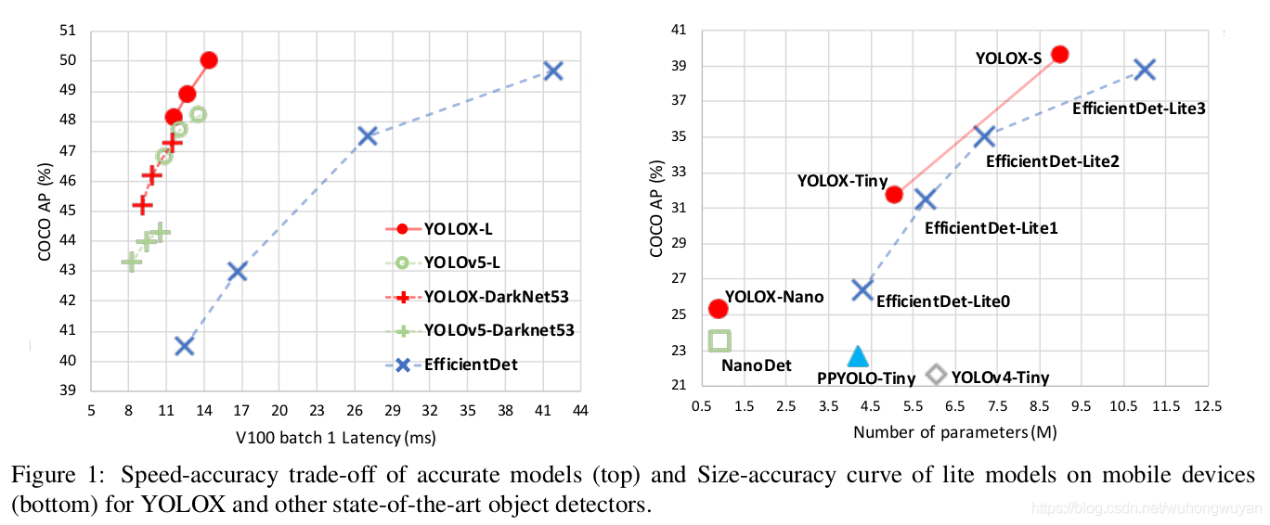
# YOLOX：2021年超越YOLO系列



## 摘要

在本论文中，我们介绍了YOLO系列的一些经验改进，形成了一个新的高性能检测器——YOLOX。我们将YOLO检测器切换为无锚方式并进行其他高级检测技术。即Decoupled head和先进的标签分配策略SimOTA，可以从大规模模型中实现得到最先进的结果。对于仅有0。91M参数的YOLO-Nano和1。08G的FLOPs。我们在COCO数据集上得到25.3%的平均准确率（AP）。超过了NanoDet 1.8%的AP。对于业界使用最广泛的检测器YOLOV3，在COCO数据集上提到了47.3%的AP，输出性能比当前最好的腰高3%个AP。对于和YOLOV4-CSP有相同数量的YOLOX-L，我们使用COCO数据集在Tesla v100上以68.9 FPS的速度得到50%的AP。比YOLOV5-L高1.8%AP。我们使用单个YOLOX-L模型赢得了第一名在Streaming Perception Challenge（Workshop on Autonomous Driving at CVPR 2021）。我们希望这个论文能够为检测研究学者们在实际应用中提供有用的经验，并且我们也提供了支持 ONNX、TensorRT、NCNN 和 Openvino 的部署版本，源代码在[https://github.com/](https://github.com/Megvii-BaseDection/YOLOX)

[Megvii-BaseDetection/YOLOX.](https://github.com/Megvii-BaseDection/YOLOX)

## 引言

随着目标检测的发展，YOLO系列始终追求实时应用的最佳速度与精度平衡。它们提取了当时最先进的检测技术（比如，YOLOv2的anchors，YOLOv3的残差网络）并优化最佳实践的实施。最近，YOLOv5以13.7ms在COCO上48.2%AP保持着最佳平衡性能。

然而，在过去的两年里，目标检测学术界的主要进展集中在anchor-free检测器，先进的标签分配策略，以及端到端（无NMS）检测器。这些还并未整合进YOLO系列中，YOLOv4和YOLOv5仍然是基于anchor的检测器，带有人工指定的训练分配规则。

这是本文存在的含义，将有经验的优化方式带给YOLO系列。考虑到YOLOv4和YOLOv5对于基于anchor的基线可能存在过度优化，我们选择YOLOv3作为起点（默认YOLOv3为YOLOv3-SPP）。事实上，由于有限计算资源与各种实际应用中不够充足的软件支持，YOLOv3仍然是工业上最广泛使用检测器之一。

## YOLOX

### YOLOX-DarkNet53

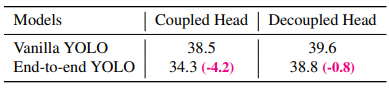


表1 带解耦头端到端YOLO在COCO上AP（%）的影响

我们选择带有Darknet53的YOLOv3作为baseline。在接下来的部分，我们将逐步介绍整个设计流程。

实施细节 最终模型的训练设置基本上延续baseline。模型训练总共300epochs，其中5epochs在COCO train2017热身。使用随机梯度下降（SGD）训练，学习速率为lr×BatchSize/64 （线性缩放），初始 lr = 0.01，余弦lr schedule。权重衰减为0.0005，SGD的动量为0.9。在默认典型8-GPU设备时Batch size=128。包括单个GPU训练在内的其他batch size也运行良好。输入大小以32步长从448到832均匀绘制。本汇报中的FPS与延迟均在单Tesla V100以FP16精度和batch=1基础上测量。

YOLOv3 baseline 我们的baseline采用DarkNet53框架和一个SPP层，在一些论文中被叫做YOLOv3-SPP。与原本实施相比，我们微小改动了一些训练策略，加入了EMA权重更新，余弦lr schedule，IoU损失和IoU感知分支。我们使用BCE损失来衡量训练cls和obj分支，IoU损失衡量训练reg分支。这些常规训练技巧与YOLOX的关键改进是正交的，因此将其放在baseline中。并且我们仅进行了随机水平翻转，颜色抖动和多尺度数据增强，抛弃了随机缩放裁剪策略，因为我们发现随机缩放裁剪与计划的Mosaic增强方式有些许重合。有了这些增强手段，baseline在COCO val上达到了38.5% AP，如表2所示。

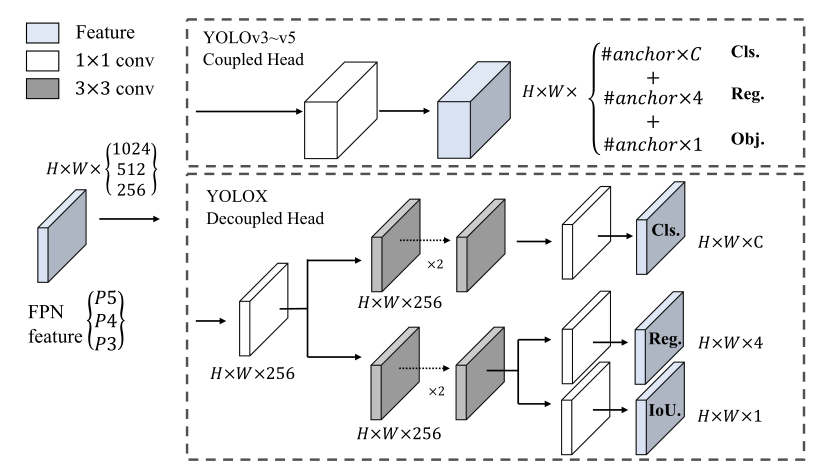


图2 YOLOv3和提出的解耦头之间的区别图注。对FPN每层特征，我们首先采用1×1卷积层来减少特征通道至256，然后加带有两个3×3卷积层（一个用来分类，一个用来回归）的两个并行分支。IoU分支加在回归任务上

解耦头 目标检测中，分类和边框回归任务的冲突是众所周知的问题。因此分类和定位的解耦头在一阶段和二阶段检测器中广泛使用。然而，随着YOLO系列的骨干网和特征金字塔（如，FPN，PAN）不断地演变，它们的检测头保持耦合，图2示。

我们两个分析实验表示耦合的检测头可能对性能有不好影响。1）将YOLO的头用一个解耦替代，极大的改善了收敛速度，如图3所示。2）解耦头对应端到端版本的YOLO至关重要（接下来会描述）。从表1可以看出，端到端性能在耦合头时下降了4.2% AP，而在解耦头时下降了0.8% AP。我们因此将YOLO检测头用一个轻量解耦头替代，图2示。具体地，它包含一个1×1卷积层来减少通道维度，后接着两个带有两层3×3卷积层的并行分支。表2中的推理时间是在V100上batch=1，轻量解耦头带来额外的1.1ms（11.6ms v.s. 10.5ms）。

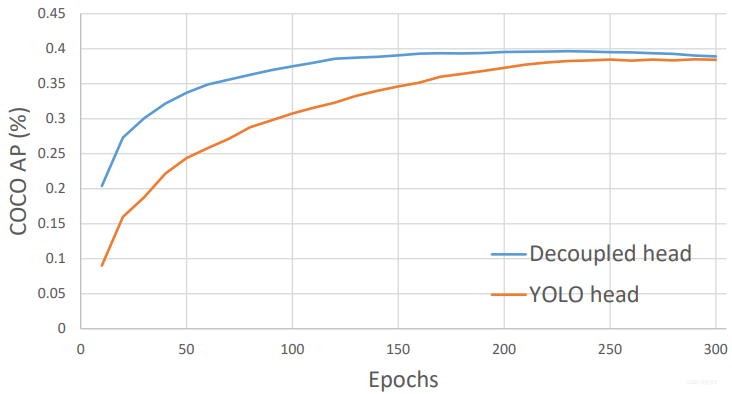


图3 检测器带YOLOv3头/解耦头的训练曲线。我们在COCO val上每10 epochs评估了AP。显然解耦头收敛比YOLOv3头快很多，并且最终达到了更好的结果

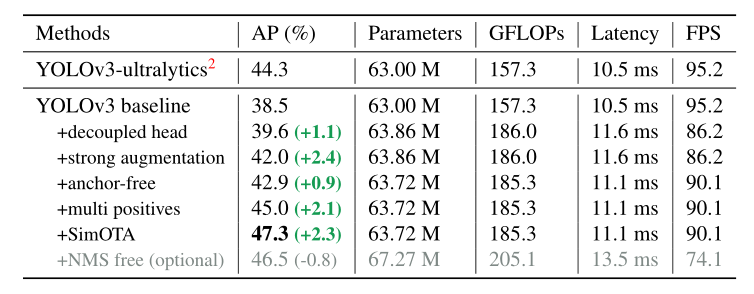


表2 YOLOX-Darknet53在COCO val上AP（%）的路线图。所有的模型均以 640 × 640 640×640 640×640分辨率，Tesla V100 FP16精度和batch=1测试。表中的延迟与FPS测量没有后处理

强数据增强我们将增加Mosaic和MixUp到增强策略中，来提高YOLOX性能。Mosaic（将4张训练图像合并成一张的数据增强方法）是一个有效的增强策略，它由ultralystics-YOLOv3提出。在YOLOv4，YOLOv5和其他检测器中广泛使用。MixUp （将不同类别图像进行通道方向上的合并）最初是为图像分类任务而设计的，但后为了目标检测训练，在BoF中调整（Bag of freebies：针对训练过程进行一些优化，比如数据增强、类别不平衡、成本函数、软标注…… 这些改进不会影响推理速度）。模型采用了MixUp和Mosiac，并在最后15epoch关闭掉，达到了42%AP，见表2。使用强数据增强后，我们发现ImageNet预训练不再具有优势，因此从头开始训练所有模型。

Anchor-free框 YOLOv4和YOLOv5均遵循原始基于anchor的 YOLOv3基线。然而，锚框机制有许多已知的问题。首先，为了实现最佳检测性能，我们需要使用聚合分析来决定训练前的一系列最优锚框。这些聚合anchors是特定域的，不通用。第二，锚框机制增加了检测头的复杂性，以及每张图的预测数量。在一些边缘AI系统，在设备间（如从NPU到CPU）传输如此大量的预测就整体延迟来说，可能成为一个潜在瓶颈。

anchor-free检测器在过去两年中飞速发展。这些工作已经说明anchor-free检测器的性能能够与基于锚框检测器平分秋色。为了使检测器获得良好的性能，尤其是其训练和解码阶段变得简单。anchor-free机制显著减少了需要启发式调整的设计参数数量和涉及的许多技巧(例如，锚聚类，网格敏感)。

将YOLO切换成一个anchor-free方式非常简单。我们减少每个位置的预测数量从3到1，并使它们直接预测4个参数，即网格左上角的两个偏置，预测框的高和宽。我们分配每个对象的中心位置为正样本并预定义一个比例范围，如[29]，以指定每个对象的FPN级。这样的修改减少了检测器的参数和GFLOPs，使它更快，且获得了更好的性能—42.9%AP，如表2示。

多重正级（Multi positives） 与YOLOv3的分配原则一致，以上anchor-free版本每个物体选择仅1个正样本（中心位置），同时忽视其他高质量的预测。然而，优化这些高质量预测可能带来有益梯度，缓解训练期间的正/负样本极端不均匀。我们简单地将中心3×3区域分为正，同FCOS中命名“中心采样”。检测器的性能改善至45.0%AP如表2，已经超过当前最佳ultralystics-YOLOv3（44.3%AP）。

SimOTA 先进标签分配是近年来目标检测另一种重要进展。基于我们自己的研究OTA，我们总结了先进标签分配的四个关键见解：1）损失/质量 感知，2）中心先验，3）每个GT的正锚框的动态数量（缩写为dynamic top-k）4）全局视图。OTA满足了以上所有四个条件，因此我们选择它作为一种候选标签分配策略。

具体地，OTA从全局角度分析了标签分配，并将分配过程表示为优化传输问题（Optimal Transport），在当前分配策略中产生了SOTA性能。但是，实际上我们发现用Sinkhorn-Knopp算法解决OT问题带来25%的额外训练时间，对于训练30epochs来说有些费时。所以简化其成为动态top-k策略，叫SimOTA，以得到一个合适的解决方式。

我们这里简要介绍了SimOTA。SimOTA首先计算了成对匹配程度，每个预测-gt对由cost或quality表示。例如，在SimOTA，GT gi 和预测pj 的cost 计算如下：

其中 λ 是平衡系数。 L i j^cls 和 L i j^reg 是GT gi和预测pj之间的分类损失和定位回归损失。对于GT gi，我们选择固定中心区域最少cost的前k个预测作为它的正样本。最后，这些正预测的对应网格被分配为正，其他为负。值得注意的是不同GT的 k 值不同。更多详情请参考OTA动态k估计策略。

SimOTA不仅减少了训练时间，也避免了Sinkhorn-Knopp算法中额外解决器的参数。如表2所示，SimOTA将AP从45.0%提高至47.3%，比SOTA ultralytics-YOLOv3高3%，表示先进分配策略的有效性。

端到端YOLO 我们按照[39]添加两个另外卷积层，一对一标签分配，和停止梯度。这些使得检测器能够以端到端的方式进行，但稍微降低了新歌能以及推理速度，列在表2中。因此，将其作为一个可选模块，并不包含在最终模型中。

### 其他骨干网

除了DarkNet53外，我们也测试了不同尺寸的其他骨干网的YOLOX，YOLOX也达到了一致的改进效果相对于对应产品。

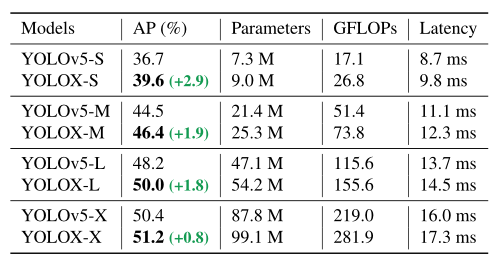


表3 YOLOX和YOLOv5在COCO val上AP（%）的比较。所有模型以 640 × 640 640×640 640×640分辨率，Tesla V100 FP16精度，batch=1测试。

修改YOLOv5的CSPNet 为了有一个公平比较，我们采用精确的YOLOv5的主干包括修改的CSPNet，SiLU激活，和PAN头。我们也遵循它的缩放规则来产生YOLOX-S，YOLOX-M，YOLOX-L，和YOLOX-X模型。与YOLOv5比较，见表3，我们的模型一致提升了3.0%到1.0%AP，只有边缘时间增加（来自于解耦头）。

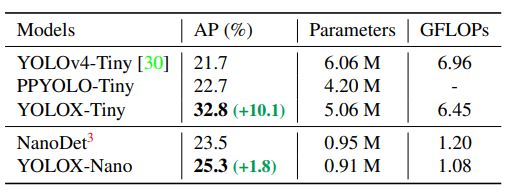


表4 YOLOX-Tiny和YOLOX-Nano与对应产品在COCO val上AP（%）之间的比较。所有的模型均在 416 × 416 416×416 416×416分辨率上测试。

Tiny和Nano检测器 我们进一步缩小了模型成为YOLOX-Tiny来与YOLOv4-Tin

y比较。对于移动设备，我们采用深度方向卷积来构建YOLOX-Nano，仅0.91M参数和1.08FLOPs。如表4所示，YOLOX在更小模型尺寸下比对应模型表现更好。

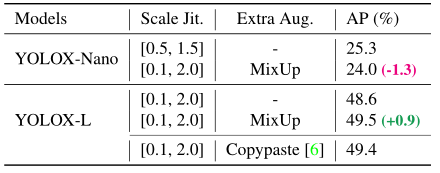


表5 不同模型尺寸下数据增强的影响。“Scale Jit.”表示马赛克图像一定范围的缩放抖动。采取Copypaste时使用COCO trainval中的实例掩膜标注

模型尺寸和数据增强 我们的实验中，所有模型保证几乎相同的学习策略和优化参数，如2.1所描述。但是，我们发现合适的增强策略在不同尺寸模型上也不同。如表5，当应用MixUp在YOLOX-L上可以提高AP0.9%，但对于像YOLOX-Nano的小模型，最好还是减弱这种增强。具体地，当训练小模型即YOLOX-S，YOLOX-Tiny，YOLOX-Nano时，我们移除mix up增强并减弱mosaic（减少缩放范围从[0.1, 2.0]到[0.5, 1.5]）。这样的修改将YOLOX-Nano的AP从24.0%提升到25.3%。

对于大模型，我们同样发现更强的增强方式更有帮助。确实，我们的MixUp措施比原始版本有一部分更重。受Copypaste启发，在将两幅图像混合之前，我们通过随机采样的比例因子对它们进行了抖动。为了理解带有缩放抖动的MixUp，我们在YOLOX-L上比较其与Copypaste。值得注意的是Copypaste需要额外实例掩膜标注但MixUp不用。但如表5显示，这两个方法达到了可竞争性能，表明当没有实例掩膜标注可用时，带有缩放抖动的MixUp是Copypaste的带限替换。

## 与SOTA比较

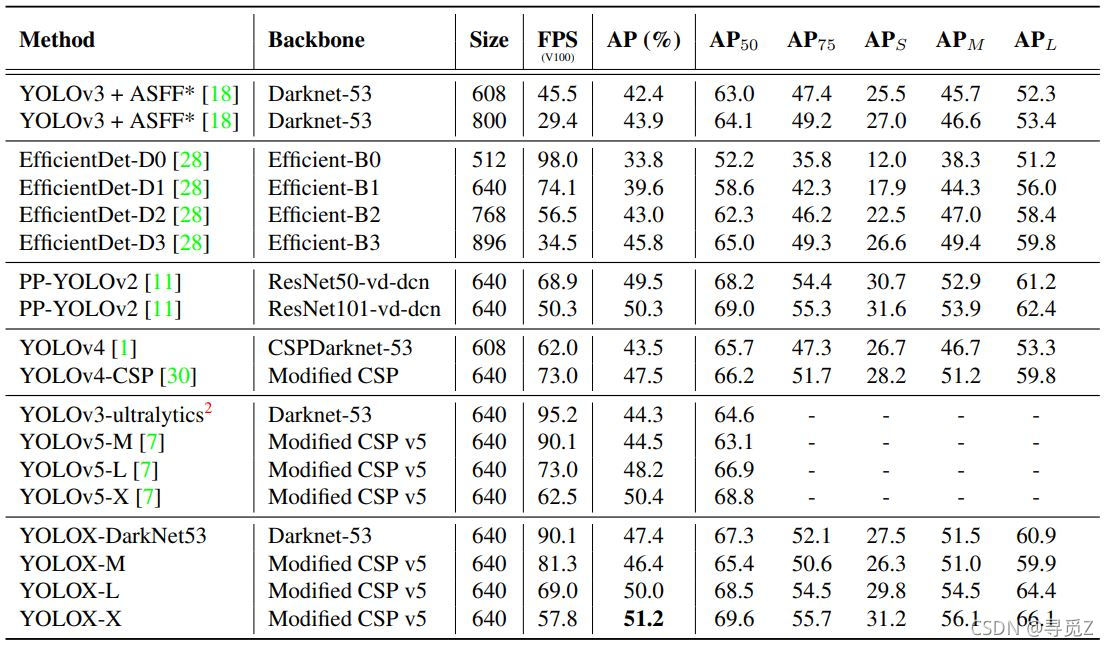


表6 不同目标检测器在COCO 2017 test-dev上的速度与精度比较。为了公平对比，选择的所有模型均训练 300epochs

表6为传统比较SOTA显示表。但是，记住表中的模型推理速度通常不可控，不同的软件与硬件速度不同。因此，我们对图1中的所有YOLO系列使用相同的硬件和代码库，绘制了稍微受控的速度/精度曲线。

我们注意到有些高性能YOLO系列具有更大模型尺寸比如Scale-YOLOv4和YOLOv5-P6。当前基于Transformer的检测器将SOTA的准确度提升至~60AP。由于时间和资源限制，本文中我们没有探讨这些重要特征。但它们已经在考虑中。

## 流媒体感知挑战（WAD at CVPR 2021）

2021WAD的流媒体感知挑战是通过最近提出的指标对准确性和延迟进行的联合评估：流媒体准确性。该指标背后的意义是联合评估每时刻整个感知栈的输出，强制堆栈考虑计算时被忽略的流数据量。我们发现在30 FPS数据流上度量的最佳平衡点是推理时间≤ 33ms的强大模型。所以我们采用TensorRT的YOLOX-L模型来产生最终模型赢得第一名。详情请见挑战网页。

## 结论

本汇报中，我们在YOLO系列上展示了一些有经验的更新，构成了高性能anchor-free检测器叫YOLOX。装配一些最近先进检测技术，即解耦头，anchor-free，先进标签分配策略。YOLOX在所有模型尺寸比其他产品达到了一个更好的速度和准确率的平衡点。由于广泛的兼容性，YOLOv3是工业上最广泛使用的检测器之一，令人惊讶的是提升了其架构使在COCO上达到了47.3%AP，超出当前最佳实践3.0%AP。希望本汇报可以帮助开发者和研究者们在实际场景中得到更好的经验。