YOLOv4: 最佳目标检测速度和精度

Alexey Bochkovskiy *
alexeyab84@gmail.com

Chien-Yao Wang *
Institute of Information Science
Academia Sinica, Taiwan
kinyiu@iis.sinica.edu.tw

Hong-Yuan Mark Liao
Institute of Information Science
Academia Sinica, Taiwan
liao@iis.sinica.edu.tw

摘要

许多特性据称可以提高卷积神经网络 (CNN) 的精度。为了验证这些特性是否有效，需要在大型数据集上对各种特性组合进行实际测试，并从理论上证明测试结果的合理性。一些特性仅适用于特定模型、特定问题或小型数据集；而另一些特性（例如批量归一化和残差连接）则适用于大多数模型、任务和数据集。我们假设加权残差连接 (WRC)、跨阶段部分连接 (CSP)、跨小型批归一化 (CmBN)、自我对抗训练 (SAT) 和Mish激活函数等属于这种通用特性。我们使用了以下新特性：WRC、CSP、CmBN、SAT、Mish 激活函数、Mosaic 数据增强、CmBN、DropBlock 正则化和 CIoU 损失函数，并通过组合其中的一些特性在 MS COCO 数据集上实现了最先进的结果：43.5% AP（65.7% AP50），在 Tesla V100上的实时速度约为 65 FPS。源代码在 https://github.com/AlexeyAB/darknet 处获得。

1. 引言

目前大多数基于卷积神经网络(CNN)的目标检测器在很大程度上仅适用于推荐系统。例如，通过城市摄像头寻找免费停车位是由速度慢但准确的模型执行的，而防撞预警则依赖于速度快但不太准确的模型。提高实时目标检测器的精度可以使其不仅用于生成提示的推荐系统，还可以用于独立流程管理并减少人工干预。在常规图形处理单元(GPU)上运行实时目标检测器使其能够以实惠的价格大规模使用。然而，最准确的现代神经网络无法实时运行，并且需要大量 GPU 才能使用大批量数据进行训练。我们通过创建一个可以在常规GPU上实时运行的CNN来解决这些问题，并且该网络的训练只需要一台常规 GPU 即可完成。

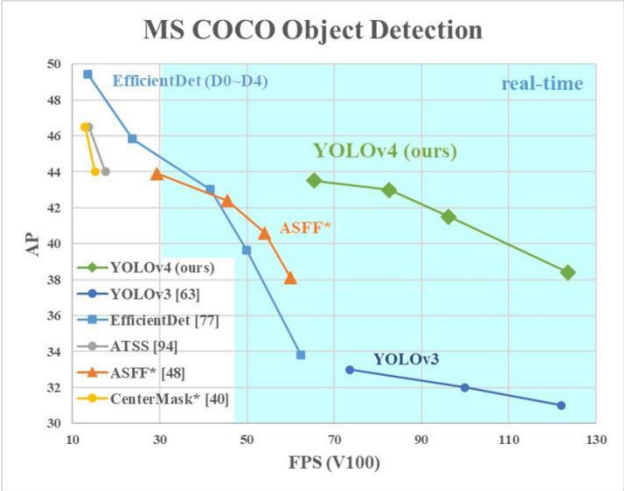


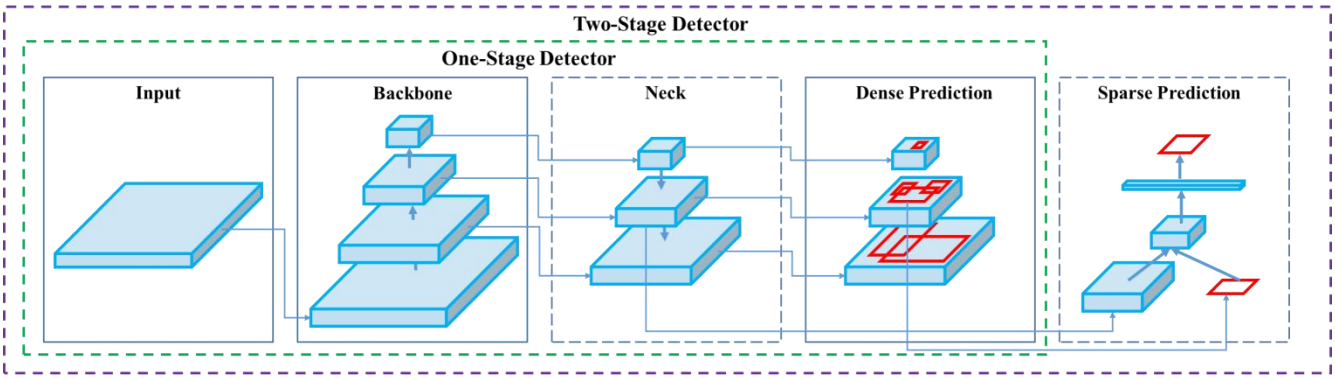
图 1：所建议的YOLOv4与其他最先进的对象检测器之间的比较。在性能相当的情况下，YOLOv4的运行速度比EfficientDet快两倍。分别比YOLOv3的AP和FPS提高了10%和12%。

这项工作的主要目标是设计一个运行速度快的对象检测器，适用于生产系统并针对并行计算进行优化，而非追求低计算量理论指标 (BFLOP)。我们希望设计的对象检测器易于训练和使用。例如，任何使用常规 GPU 训练和测试的人都可以像图1中展示的 YOLOv4 结果那样，实现实时、高质量且令人信服的对象检测。我们做出的贡献可以总结如下：

1. 开发了一个高效强大的对象检测模型。它使每个人都可以在一块 1080 Ti 或 2080 Ti GPU 上训练出一个超快速且准确的对象检测器。

2. 验证了最先进的 “免费赠品包” 和 “特殊赠品包” 方法在对象检测器训练过程中的影响。

3. 修改了最先进的方法，使其更有效且适用于单 GPU 训练，例如 CBN [89]、PAN [49]、SAM [85] 等。



Input: { Image, Patches, Image Pyramid, … }

Backbone: { VGG16 [68], ResNet-50 [26], ResNeXt-101 [86], Darknet53 [63], … }

Neck: { FPN [44], PANet [49], Bi-FPN [77], … }

Head:

Dense Prediction: { RPN [64], YOLO [61, 62, 63], SSD [50], RetinaNet [45], FCOS [78], … }

Sparse Prediction: { Faster R-CNN [64], R-FCN [9], … }

图 2：目标监测器

2. 相关工作

2.1. 目标监测模型

现代目标检测器通常由两部分组成：主干网络和检测头。主干网络用于提取图像特征，通常预训练于 ImageNet 数据集上。在 GPU 平台上，常用的主干网络包括 VGG [68]、ResNet [26]、ResNeXt [86] 或 DenseNet [30] 等。在 CPU 平台上，则有 SqueezeNet [31]、MobileNet [28, 66, 27, 74] 或 ShuffleNet [97, 53] 等。检测头用于预测目标的类别和边界框。检测头通常分为两类：单阶段和两阶段。最具代表性的两阶段目标检测器是 R-CNN [19] 系列，包括 fast R-CNN [18]、faster R-CNN [64]、R-FCN [9] 和 Libra R-CNN [58]。还有一些特殊的两阶段检测器，例如 RepPoints [87]，它摒弃了锚框机制。最具代表性的单阶段目标检测器是 YOLO [61, 62, 63]、SSD [50] 和 RetinaNet [45]。近年来，一些无锚框的单阶段目标检测器也得到发展，例如 CenterNet [13]、CornerNet [37, 38] 和 FCOS [78] 等。近年来开发的目标检测器经常会在主干网络和检测头之间加入一些额外的层，称为颈部 (neck)。这些层通常用于融合来自不同阶段的特征图。常见的颈部结构包括 Feature Pyramid Network (FPN) [44]、Path Aggregation Network (PAN) [49]、BiFPN [77] 和 NAS-FPN [17] 等。除了以上提到的模型，一些研究人员还致力于直接构建新的用于目标检测的主干网络 (DetNet [43]、DetNAS [7]) 或是整个模型 (SpineNet [12]、HitDetector [20])。

综上所述，一个普通的目标检测器由以下几个部分组成：

- **输入**：图像、图像块、图像金字塔
- **主干网络**：VGG16 [68]、ResNet-50 [26]、SpineNet [12]、EfficientNet-B0/B7 [75]、CSPResNeXt50 [81]、CSPDarknet53 [81]
- **颈部**：
  - **附加模块**：SPP [25]、ASPP [5]、RFB [47]、SAM [85]
  - **路径聚合模块**：FPN [44]、PAN [49]、NAS-FPN [17]、全连接 FPN、BiFPN [77]、ASFF [48]、SFAM [98]
- **检测头**：
  - 稠密预测 (单阶段)：
    - RPN [64]、SSD [50]、YOLO [61]、RetinaNet [45] (基于锚框)
    - CornerNet [37]、CenterNet [13]、MatrixNet [60]、FCOS [78] (无锚框)
  - 稀疏预测 (两阶段)：
    - Faster R-CNN [64]、R-FCN [9]、Mask R-CNN [23] (基于锚框)
    - RepPoints [87] (无锚框)

2.2. 免费赠品包

通常情况下，物体检测器采用离线训练的方式。因此，研究人员总是倾向于开发更好的训练方法，这些方法可以在不增加推理成本的情况下提高物体检测器的精度。我们将仅改变训练策略或仅增加训练成本的方法称为“免费赠品包”。数据增强是物体检测方法经常采用且符合免费赠品包定义的一种方法。数据增强的目的是增加输入图像的多样性，从而使设计的物体检测模型对来自不同环境的图像具有更高的鲁棒性。例如，光度畸变和几何畸变是两种常用的数据增强方法，它们肯定能对物体检测任务带来益处。在处理光度畸变时，我们会调整图像的亮度、对比度、色调、饱和度和噪声。对于几何畸变，我们会加入随机缩放、裁剪、翻转和旋转操作。

上面提到的数据增强方法都是逐像素的调整，调整区域的所有原始像素信息都会被保留。此外，一些从事数据增强的研究人员将重点放在模拟物体遮挡问题上。他们在图像分类和物体检测方面取得了良好的效果。例如，随机擦除 [100] 和 CutOut [11] 可以随机选择图像中的矩形区域，并用随机值或补零值填充。Hide-and-Seek [69] 和网格掩码 [6] 则会随机或均匀地选择图像中的多个矩形区域，并替换为全零值。如果将类似的概念应用于特征图，则有 Dropout [71]、DropConnect [80] 和 DropBlock [16] 方法。此外，一些研究人员提出了使用多张图像一起进行数据增强的的方法。例如，MixUp [92] 使用两张图像以不同的系数比率进行乘法叠加，然后根据这些叠加比率调整标签。CutMix [91] 则是将裁剪的图像覆盖到其他图像的矩形区域，并根据混合区域的大小调整标签。除了上述方法外，风格迁移 GAN [15] 也被用于数据增强，这种用法可以有效减少 CNN 学习到的纹理偏差。

与上面提到的各种方法不同，其他一些免费赠品包方法致力于解决数据集中的语义分布可能存在偏差的问题。在处理语义分布偏差问题时，一个非常重要的问题是不同类别之间存在数据不平衡的问题，这个问题通常可以通过两阶段目标检测器中的困难负样本挖掘 [72] 或在线困难样本挖掘 [67] 来解决。但是，示例挖掘方法并不适用于单阶段目标检测器，因为这种检测器属于密集预测架构。因此，Lin 等人 [45] 提出了焦距损失 (focal loss) 来处理不同类别之间存在的数据不平衡问题。另一个非常重要的问题是，使用 one-hot 硬表示很难表达不同类别之间关联程度的关系。这种表示方案在执行标记时经常使用。论文 [73] 中提出的标签平滑化将硬标签转换为软标签进行训练，可以使模型更加鲁棒。为了获得更好的软标签，Islam 等人 [33] 引入了知识蒸馏的概念来设计标签细化网络。

最后一个免费赠品包是边界框 (BBox) 回归的目标函数。传统的目标检测器通常使用均方误差 (MSE) 直接对 BBox 的中心点坐标和宽高进行回归，即 $\{x_{center}, y_{center}, w, h\}$ 或左上角点和右下角点，即 $\{x_{top\_left}, y_{top\_left}, x_{bottom\_right}, y_{bottom\_right}\}$。对于基于锚框的方法，则是估计对应的偏移量，例如 $\{x_{center_offset}, y_{center_offset}, w_{offset}, h_{offset}\}$和 $\{x_{top\_left\_offset}, y_{top\_left\_offset}, x_{bottom\_right\_offset}, y_{bottom\_right\_offset}\}$。然而，直接估计 BBox 的每个点的坐标值，是将这些点作为独立变量对待，并没有考虑物体本身的完整性。

为了更好地处理这个问题，一些研究人员最近提出了 IoU 损失 [90]，它将预测的 BBox 区域和 ground truth BBox 区域的覆盖率纳入考量。IoU 损失的计算过程会通过与 ground truth 执行 IoU 计算来触发 BBox 的四个坐标点的计算，然后将生成的结果连接成一个整体。由于 IoU 是尺度不变的表示，因此它可以解决传统方法在计算 {x, y, w, h} 的 l1 或 l2 损失时，损失会随着尺度增加的问题。

最近，一些研究人员又对 IoU 损失进行了改进。例如，GIoU 损失 [65] 除了考虑覆盖面积外，还加入了物体的形状和方向。他们提出寻找一个最小的面积 BBox，可以同时覆盖预测的 BBox 和 ground truth BBox，并使用这个 BBox 作为分母来替换 IoU 损失中原本使用的分母。DIoU 损失 [99] 则额外考虑了物体中心的距离，而 CIoU 损失 [99] 则同时考虑了重叠面积、中心点间距和长宽比。CIoU 可以使 BBox 回归问题取得更好的收敛速度和精度。

2.3. 特殊赠品包

对于那些仅略微增加推理成本但可以显著提高目标检测精度的插件模块和后处理方法，我们称之为“特殊赠品包”。一般来说，这些插件模块用于增强模型的某些属性，例如扩大感受野、引入注意力机制或加强特征融合能力等，而后处理是一种筛选模型预测结果的方法。

可用于扩展感受野的常用模块包括 SPP [25]、ASPP [5] 和 RFB [47]。SPP 模块源自空间金字塔匹配 (SPM) [39]，SPM 的原始方法是将特征图分割成几个大小为 d × d 的相等块，其中 d 可以取 {1, 2, 3, ...}，从而形成空间金字塔，然后提取词袋特征。SPP 将 SPM 集成到 CNN 中，并使用最大池化操作代替词袋操作。He 等人 [25] 提出的 SPP 模块会输出一维特征向量，因此无法直接应用于全卷积网络 (FCN)。因此在 YOLOv3 [63] 的设计中，Redmon 和 Farhadi 改进了 SPP 模块，将其变为内核大小为 k × k 的最大池化输出的连接，其

中 k = {1, 5, 9, 13}，步长等于 1。在此设计下，相对较大的 k × k 最大池化有效地增加了主干网络特征的感受野。加入改进版的 SPP 模块后，YOLOv3-608 在 MS COCO 目标检测任务上的 AP50 提升了 2.7%，但额外增加了 0.5% 的计算量。改进后的 SPP 模块与 ASPP [5] 模块之间的操作差异主要在于原始内核大小 k × k，步长为 1 的最大池化改成了多个 3 × 3 的内核大小，空洞率等于 k，步长等于 1 的膨胀卷积操作。RFB 模块则使用多个内核大小为 k × k、空洞率等于 k、步长等于 1 的膨胀卷积来获得比 ASPP 更全面的空间覆盖范围。RFB [47] 只需额外增加 7% 的推理时间，即可将 SSD 在 MS COCO 上的 AP50 提高 5.7%。

目标检测中常用的注意力模块主要分为通道注意力和点注意力，这两种注意力模型的代表分别为 Squeeze-and-Excitation (SE) [29] 和空间注意力模块 (SAM) [85]。虽然 SE 模块可以在 ImageNet 图像分类任务上将 ResNet50 的 top-1 准确率提高 1%，并且仅增加 2% 的计算量，但是在 GPU 上它通常会增加大约 10% 的推理时间，因此更适合用于移动设备。而 SAM 只需增加 0.1% 的额外计算量，就可以在 ImageNet 图像分类任务上将 ResNet50-SE 的 top-1 准确率提高 0.5%，而且它完全不影响在 GPU 上的推理速度。

在特征融合方面，早期做法是使用跳跃连接 [51] 或超列 [22] 将低层次的物理特征与高级语义特征进行融合。随着 FPN 等多尺度预测方法的普及，许多用于整合不同特征金字塔的轻量级模块被提出。这类模块包括 SFAM [98]、ASFF [48] 和 BiFPN [77]。SFAM 的主要思想是使用 SE 模块对多尺度连接的特征图执行通道级的重新加权。对于 ASFF，它使用 softmax 作为点级的重新加权，然后添加不同尺度的特征图。在 BiFPN 中，引入了多输入加权残差连接来执行尺度级的重新加权，然后添加不同尺度的特征图。

在深度学习的研究中，一些人将重点放在寻找良好的激活函数上。良好的激活函数可以使梯度更有效地传播，同时不会带来太大的额外计算成本。2010 年，Nair 和 Hinton [56] 提出了 ReLU 函数，有效地解决了传统 tanh 和 sigmoid 激活函数经常遇到的梯度消失问题。随后，为了解决梯度消失问题，又陆续提出了 LReLU [54]、PReLU [24]、ReLU6 [28]、缩放指数线性单元 (SELU) [35]、Swish [59]、hard-Swish [27] 和 Mish [55] 等激活函数。LReLU 和 PReLU 的主要目的是解决 ReLU 在输出小于零时梯度为零的问题。ReLU6 和 hard-Swish 则专门为量化网络而设计。SELU 激活函数的提出是为了实现神经网络的自归一化。值得注意的是，Swish 和 Mish 都是连续可微的激活函数。

深度学习目标检测常用的后处理方法是 NMS，它可以用于过滤那些错误预测相同物体的边界框，只保留具有较高响应的候选边界框。NMS 尝试改进的方式与优化目标函数的方法是一致的。NMS 最初提出的方法没有考虑上下文信息，因此 Girshick 等人 [19] 在 R-CNN 中加入了分类置信度得分作为参考，并根据置信度得分的高低顺序进行贪婪 NMS。Soft NMS [1] 则考虑了在贪婪 NMS 中使用 IoU 得分可能会导致物体遮挡导致置信度得分降低的问题。DIoU NMS [99] 的开发者在软 NMS 的基础上，加入了中心点距离信息到边界框筛选过程中。值得一提的是，由于以上后处理方法都不直接引用捕获的图像特征，因此后续发展的无锚框方法不再需要后处理。

3. 方法

在生产系统中，我们优化的主要目标是神经网络的快速运行速度和并行计算，而不是理论计算量指标 (BFLOP)。下面我们介绍两种实时神经网络的方案：

- 针对 GPU：我们在卷积层中使用少量的组 (1 - 8)，例如 CSPResNeXt50 / CSPDarknet53
- 针对 VPU：我们使用分组卷积，但避免使用 Squeeze-and-Excitement (SE) 模块，具体包括以下模型：EfficientNet-lite / MixNet [76] / GhostNet [21] / MobileNetV3

3.1. 模型结构选择

我们的目标是在输入网络分辨率、卷积层数量、参数数量（滤波器大小 * 过滤器 * 通道 / 组数）和层输出数量（滤波器）之间找到最佳平衡。例如，我们的大量研究表明，在 ILSVRC2012（ImageNet）数据集上进行对象分类时，CSPResNext50 的性能明显优于 CSPDarknet53 [10]。然而，另一方面，在 MS COCO 数据集 [46] 上进行对象检测时，CSPDarknet53 的性能优

表1：用于图像分类的神经网络参数大小

骨干模型	网络输入分辨率	感受野大小	参数量	每层输出的平均大小(WxHxC)	BFLOPs (512x512网络输入分辨率)	FPS (GPU RTX 2070)
CSPResNext50	512x512	425x425	20.6 M	**1058 K**	31 (15.5 FMA)	62
CSPDarknet53	512x512	725x725	**27.6 M**	950 K	52 (26.0 FMA)	**66**
EfficientNet-B3 (ours)	512x512	**1311x1311**	12.0 M	668 K	11 (5.5 FMA)	26

于 CSPResNext50。

下一个目标是选择用于增加感受野的附加模块，以及用于不同检测器层级来自不同主干网络层级参数聚合的最佳方法：例如 FPN、PAN、ASFF、BiFPN。

仅仅在分类任务上表现最佳的模型并不总是适用于检测任务。与分类器相比，检测器需要以下几点：

- 更高的输入网络尺寸（分辨率）- 用于检测多个小型物体
- 更多的层 - 用于更大的感受野以覆盖增加的输入网络尺寸
- 更多的参数 - 用于模型更大的容量，以便在单个图像中检测不同大小的多个物体

从理论上讲，我们可以假设应该选择具有更大感受野大小（具有更多个 3x3 的卷积层）和更多参数的模型作为主干网络。表 1 显示了 CSPResNeXt50、CSPDarknet53和 EfficientNet B3 的信息。CSPResNext50 仅包含 16 个 3x3 的卷积层、425 × 425 的感受野和 20.6 M 参数，而 CSPDarknet53 包含 29 个 3x3 的卷积层、725 × 725 的感受野和 27.6 M 参数。这种理论依据与我们的大量实验一起表明，CSPDarknet53 神经网络作为检测器的主干网络比两者更优。

不同大小的感受野的影响总结如下：

- 小于等于物体大小 - 允许查看整个物体
- 小于等于网络大小 - 允许查看物体周围的上下文
- 超过网络大小 - 增加图像点和最终激活之间连接的数量

我们在 CSPDarknet53 上添加了 SPP 模块，因为它可以显着增加感受野，分离出最重要的上下文特征，并且几乎不会降低网络运行速度。我们使用 PANet 作为来自不同主干网络层级到不同检测器层级的参数聚合方法，而不是 YOLOv3 中使用的 FPN。

最后，我们选择 CSPDarknet53 主干网络、SPP 附加模块、PANet 路径聚合颈部和 YOLOv3（基于锚框）头作为 YOLOv4 的结构。

未来我们计划大幅扩展免费赠品包 (BoF) 的内容，理论上可以解决一些问题并提高检测器精度，并依次以实验方式检查每个特征的影响。

我们不使用跨 GPU 批归一化 (CGBN 或 SyncBN) 或昂贵的专用设备。这允许任何人在常规图形处理器（例如 GTX 1080Ti 或 RTX 2080Ti）上重现我们最先进的成果。

3.2. BoF和BoS的选择

CNN 通常使用以下组件以改善目标检测训练效果：

- **激活函数**：ReLU、leaky-ReLU、parametric-ReLU、ReLU6、SELU、Swish 或 Mish
- **边框回归损失**：MSE、IoU、GIoU、CIoU、DIoU
- **数据增强**：CutOut、MixUp、CutMix
- **正则化方法**：DropOut、DropPath [36]、Spatial DropOut [79] 或 DropBlock
- **网络激活值的归一化**：Batch Normalization (BN) [32]、Cross-GPU Batch Normalization (CGBN 或 SyncBN) [93]、Filter Response Normalization (FRN) [70] 或 Cross-Iteration Batch Normalization (CBN) [89]
- **跳跃连接**：残差连接、加权残差连接、多输入加权残差连接或 Cross stage partial connections (CSP)

由于 PReLU 和 SELU 训练难度更大，ReLU6 专为量化网络设计，因此我们从候选列表中删除了这些激活函数。对于正则化方法，DropBlock 的发布者详细对比了他们的方法与其他方法，并且他们的正则化方法在对比中表现良好。因此，我们毫不犹豫地选择 DropBlock 作为正则化方法。由于我们专注于仅使用单个 GPU 的训练策略，因此在归一化方法的选择上不考虑 SyncBN。

3.3. 额外的改进

为了使设计的检测器更适合于单个 GPU 训练，我们进行了以下额外的设计和改进：

- 我们引入了一种新的数据增强方法 Mosaic 和对抗生成训练 (SAT)
- 我们在应用遗传算法时选择最佳的超参数
- 我们修改了一些现有方法以使我们的设计适用于高效训练和检测 - 修改后的 SAM、修改后的 PAN 和跨迷你批归一化 (CmBN)

Mosaic 是一种新的数据增强方法，它混合了 4 张训练图像。因此，可以混合 4 种不同的上下文，而 CutMix 只混合 2 张输入图像。这允许检测超出其正常上下文的对象。此外，批量归一化会计算来自每个层上 4 张不同图像的激活统计信息。这大大减少了对大批量大小的需求。

对抗生成训练 (SAT) 也代表了一种新的数据增强技术，它分两个前向后向阶段进行操作。在第一阶段，神经网络不是改变网络权重，而是改变原始图像。这样，神经网络对自己执行对抗攻击，改变原始图像，使其看

起来图像上不存在目标物体。在第二阶段，神经网络以正常方式训练该修改后的图像来检测物体。

图 3：Mosaic 代表了一种新的数据增强方法

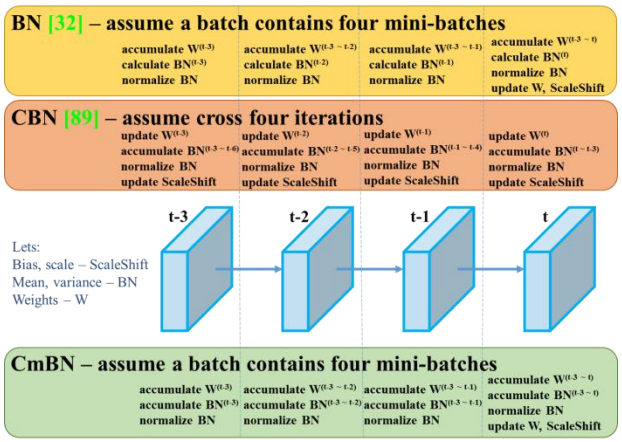


图 4: 跨迷你批归一化

CmBN 表示如图 4 所示的 CBN 修改版本，称为跨迷你批归一化 (CmBN)。它仅收集单个批次内迷你批次之间的统计信息。

我们从空间注意力修改为点注意力修改 SAM，并分别如图 5 和图 6 所示，将 PAN 的捷径连接替换为连接操作。

3.4. YOLOv4

本章节我们将详细介绍 YOLOv4 的组成部分：

- 主干网络 (Backbone): CSPDarknet53 [81]
- 颈部 (Neck): SPP [25]、PAN [49]
- 头部 (Head): YOLOv3 [63]

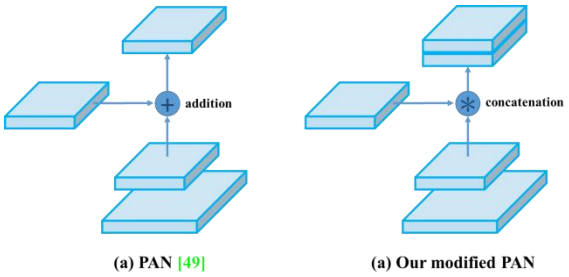


图 5: 修改的 SAM.

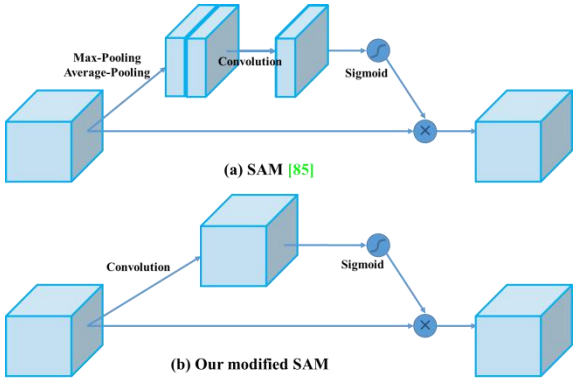


图 6: 修改的 PAN.

YOLOv4 使用了以下方法：

- 主干网络的免费赠品包 (BoF): CutMix 和 Mosaic 数据增强、DropBlock 正则化、类别标签平滑化
- 主干网络的特殊赠品包 (BoS): Mish 激活函数、跨阶段部分连接 (CSP)、多输入加权残差连接 (MiWRC)
- 检测器的免费赠品包 (BoF): CIoU 损失函数、CmBN、DropBlock 正则化、Mosaic 数据增强、对抗生成训练、消除网格敏感性、使用单个真实框匹配多个锚框、Cosine 退火调度器 [52]、最佳超参数、随机训练形状
- 检测器的特殊赠品包 (BoS): Mish 激活函数、SPP 模块、SAM 模块、PAN 路径聚合模块、DIoU-NMS

4. 方法

我们在 ImageNet (ILSVRC 2012 val) 数据集上测试了不同训练改进技术对分类器精度的影响，然后在 MS COCO (test-dev 2017) 数据集上测试了对检测器精度的影响。

4.1. 实验设置

在 ImageNet 图像分类实验中，默认超参数如下：训练步数为 8,000,000；批量大小和迷你批处理大小分别为 128 和 32；采用多项式衰减学习率调度策略，初始学习率为 0.1；热身步数为 1000；动量和权重衰减分

别设置为 0.9 和 0.005。我们所有 BoS 实验都使用与默认设置相同的超参数，在 BoF 实验中，我们额外增加了 50% 的训练步数。在 BoF 实验中，我们验证了 MixUp、CutMix、Mosaic、模糊数据增强和标签平滑正则化方法。在 BoS 实验中，我们比较了 LReLU、Swish 和 Mish 激活函数的效果。所有实验均使用 1080 Ti 或 2080 Ti GPU 训练。

在 MS COCO 目标检测实验中，默认超参数如下：训练步数为 500,500；采用 step decay 学习率调度策略，初始学习率为 0.01，分别在 400,000 步和 450,000 步时乘以因子 0.1；动量和权重衰减分别设置为 0.9 和 0.0005。所有架构都使用单个 GPU 在批量大小为 64 的情况下执行多尺度训练，而迷你批处理大小为 8 或 4，具体取决于架构和 GPU 内存限制。除了使用遗传算法进行超参数搜索实验之外，所有其他实验都使用默认设置。遗传算法使用带 GIoU 损失的 YOLOv3-SPP 进行训练，搜索 300 个 epoch 以获得最小验证集 5k。对于遗传算法实验，我们采用搜索到的学习率 0.00261、动量 0.949、分配 ground truth 的 IoU 阈值 0.213 和损失归一化因子 0.07。我们验证了大量 BoF 技术，包括消除网格敏感性、Mosaic 数据增强、IoU 阈值、遗传算法、类别标签平滑化、跨迷你批归一化、对抗生成训练、cosine 退火调度器、动态迷你批处理大小、DropBlock、优化锚框、不同类型的 IoU 损失函数。我们还对各种 BoS 进行了实验，包括 Mish、SPP、SAM、RFB、BiFPN 和高斯 YOLO [8]。对于所有实验，我们只使用单个 GPU 进行训练，因此不使用诸如 syncBN 之类的优化多个 GPU 的技术。

4.2. 分类器训练中不同特征的影响

首先，我们研究了不同特征对分类器训练的影响；具体而言，研究了类别标签平滑化、不同数据增强技术（双边模糊、MixUp、CutMix 和 Mosaic，如图7所示）以及不同激活函数（默认 Leaky-ReLU、Swish 和 Mish）的影响。

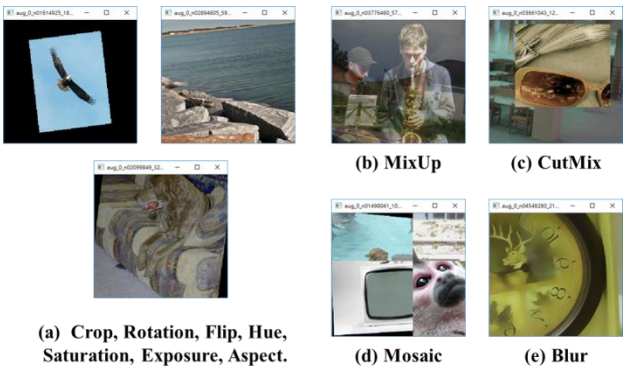


图 7：各种数据增强方法

实验结果如表2所示，引入 CutMix 和 Mosaic 数据增强、类别标签平滑化以及 Mish 激活函数等特征可以提高分类器的精度。因此，我们的分类器训练用 BoF（免费赠品包）主干网络包含以下内容：CutMix 和 Mosaic 数据增强以及类别标签平滑化。此外，我们还可以选择使用 Mish 激活函数，如表2和表3所示。

表 2：BoF 和 Mish 激活函数对 CSPResNeXt-50 分类器精度的影响。

MixUp	CutMix	Mosaic	Bluring	Label Smoothing	Swish	Mish	Top-1	Top-5
							77.9%	94.0%
✓							77.2%	**94.0%**
	✓						**78.0%**	**94.3%**
		✓					**78.1%**	**94.5%**
			✓				77.5%	93.8%
				✓			**78.1%**	**94.4%**
					✓		64.5%	86.0%
						✓	**78.9%**	**94.5%**
	✓	✓		✓			**78.5%**	**94.8%**
	✓	✓		✓		✓	**79.8%**	**95.2%**

表 3：BoF 和 Mish 激活函数对 CSPDarknet-53 分类器精度的影响。

MixUp	CutMix	Mosaic	Bluring	Label Smoothing	Swish	Mish	Top-1	Top-5
							77.2%	93.6%
	✓	✓		✓			**77.8%**	**94.4%**
	✓	✓		✓		✓	**78.7%**	**94.8%**

4.3. 检测器训练中不同特征的影响

接下来我们研究了不同免费赠品包 (BoF-detector) 对检测器训练精度的影响，如表 4 所示。我们通过研究可以提高检测精度而不影响 FPS 的不同特征，显著地扩展了 BoF 列表：

- S: 消除网格敏感性 - YOLOv3 中用于评估目标坐标的方程 $b_x = \sigma(t_x) + c_x$ 和 $b_y = \sigma(t_y) + c_y$ 要求 c_x 和 c_y 始终为整数，因此对于 b_x 值接近 c_x 或 $c_x + 1$ 则需要非常高的 t_x 绝对值。我们通过将 sigmoid 乘以大于 1.0 的因子来解决这个问题，从而消除网格对物体检测的影响
- M: Mosaic 数据增强 - 在训练过程中使用 4 张图像的 Mosaic 数据增强而不是单张图像
- IT: IoU 阈值 - 使用多个锚框匹配单个真实框，要求 IoU (真实框，锚框) > IoU 阈值
- GA: 遗传算法 - 在网络训练开始的 10% 时间内使用遗传算法选择最佳超参数
- LS: 类别标签平滑化 - 使用类别标签平滑化用于

sigmoid 激活函数

- CBN: CmBN - 使用跨迷你批归一化来收集整个批次中的统计信息，而不是收集单个迷你批次中的统计信
- CA: Cosine 退火调度器 - 在正弦训练过程中改变学习率
- DM: 动态迷你批处理大小 - 通过使用随机训练形状在我们进行小分辨率训练期间自动增加迷你批处理大小
- OA: 优化锚框 - 使用针对 512x512 网络分辨率训练的优化锚框
- GIoU、CIoU、DIoU、MSE - 使用不同的损失算法进行边界框回归

我们进一步研究了不同特殊赠品包 (BoS-detector) 对检测器训练精度的影响，例如 PAN、RFB、SAM、高斯 YOLO (G) 和 ASFF，如表5 所示。实验结果表明，当使用 SPP、PAN 和 SAM 时，检测器可以获得最佳性能。

4.4. 不同主干网络和预训练权重对检测器训练的影响

接下来我们研究了不同主干网络模型对检测器精度的影响，如表6所示。我们注意到，具有最佳分类精度的模型并不总是具有最佳检测精度的模型。

首先，虽然使用不同特征训练的 CSPResNeXt-50 模型的分类精度高于 CSPDarknet53 模型，但是在目标检测方面，CSPDarknet53 模型表现出更高的精度。

其次，使用 BoF 和 Mish 训练 CSPResNeXt50 分类器可以提高其分类精度，但是进一步将这些预训练权重用于检测器训练会降低检测器精度。然而，使用 BoF 和 Mish 训练 CSPDarknet53 分类器可以提高分类器和使用该分类器预训练权重的检测器的精度。最终结果是，主干网络 CSPDarknet53 比 CSPResNeXt50 更适合检测器。

4.5. 不同迷你批处理大小对检测器训练的影响

最后，我们分析了使用不同迷你批处理大小训练的模型的结果，结果如图 7 所示。从表 7 的结果可以看出，加入了 BoF 和 BoS 训练策略后，迷你批处理大小对检测器的性能几乎没有影响。这个结果表明，加入了 BoF 和 BoS 之后，就不再需要使用昂贵的 GPU 进行训练了。换句话说，任何人都可以使用普通的 GPU 来训练出优秀的检测器。

5. 结果

图 8 展示了与其他最先进目标检测器的对比结果。我们的 YOLOv4 位于帕累托最优曲线 (Pareto optimality curve) 上，在速度和精度方面都优于速度和精度最快的检测器。

6. 总结

我们提供了一个最先进的目标检测器，其速度 (FPS) 和精度 (MS COCO AP50...95 和 AP50) 均优于所有现有的替代检测器。所述的检测器可以在具有 8-16 GB 显存的常规 GPU 上训练和使用，这使其能够广泛应用。单阶段基于锚框的检测器这一原始概念已经证明了其可行性。我们验证了大量特征，并从中挑选了一些可以同时提高分类器和检测器精度的特征。这些特征可以作为未来研究和开发的最佳实践。

7. 致谢

作者感谢 Glenn Jocher 提供了 Mosaic 数据增强、使用遗传算法选择超参数以及解决网格敏感性问题的想法 https://github.com/ultralytics/yolov3。

表 5：特殊赠品包的消融实验 (Size 512x512).

模型	AP	AP_{50}	AP_{75}
CSPResNeXt50-PANet-SPP	42.4%	64.4%	45.9%
CSPResNeXt50-PANet-SPP-RFB	41.8%	62.7%	45.1%
CSPResNeXt50-PANet-SPP-SAM	**42.7%**	**64.6%**	**46.3%**
CSPResNeXt50-PANet-SPP-SAM-G	41.6%	62.7%	45.0%
CSPResNeXt50-PANet-SPP-ASFF-RFB	41.1%	62.6%	44.4%

表 6：使用不同的分类器预训练权重进行检测器训练（所有其他训练参数在所有模型中都类似）

模型（最优配置）	Size	AP	AP_{50}	AP_{75}
CSPResNeXt50-PANet-SPP	512x512	42.4	64.4	45.9
CSPResNeXt50-PANet-SPP (BoF-backbone)	512x512	42.3	64.3	45.7
CSPResNeXt50-PANet-SPP (BoF-backbone + Mish)	512x512	42.3	64.2	45.8
CSPDarknet53-PANet-SPP (BoF-backbone)	512x512	42.4	64.5	46.0
CSPDarknet53-PANet-SPP (BoF-backbone + Mish)	512x512	43.0	64.9	46.5

表 7：使用不同迷你批处理大小进行检测器训练

模型（没有 OA）	Size	AP	AP_{50}	AP_{75}
CSPResNeXt50-PANet-SPP (without BoF/BoS, mini-batch 4)	608	37.1	59.2	39.9
CSPResNeXt50-PANet-SPP (without BoF/BoS, mini-batch 8)	608	38.4	60.6	41.6
CSPDarknet53-PANet-SPP (with BoF/BoS, mini-batch 4)	512	41.6	64.1	45.0
CSPDarknet53-PANet-SPP (with BoF/BoS, mini-batch 8)	512	41.7	64.2	45.2

表 4：免费赠品包的消融实验 (CSPResNeXt50-PANet-SPP, 512x512).

S	M	IT	GA	LS	CBN	CA	DM	OA	loss	AP	AP_{50}	AP_{75}
									MSE	38.0%	60.0%	40.8%
✓									MSE	37.7%	59.9%	40.5%
	✓								MSE	**39.1%**	**61.8%**	**42.0%**
		✓							MSE	36.9%	59.7%	39.4%
			✓						MSE	**38.9%**	**61.7%**	**41.9%**
				✓					MSE	33.0%	55.4%	35.4%
					✓				MSE	**38.4%**	**60.7%**	**41.3%**
						✓			MSE	**38.7%**	**60.7%**	**41.9%**
							✓		MSE	35.3%	57.2%	38.0%
✓									GIoU	**39.4%**	59.4%	**42.5%**
✓									DIoU	**39.1%**	58.8%	**42.1%**
✓									CIoU	**39.6%**	59.2%	**42.6%**
✓	✓	✓	✓						CIoU	**41.5%**	**64.0%**	**44.8%**
	✓		✓					✓	CIoU	36.1%	56.5%	38.4%
✓	✓	✓	✓					✓	MSE	**40.3%**	**64.0%**	**43.1%**
✓	✓	✓	✓					✓	GIoU	**42.4%**	**64.4%**	**45.9%**
✓	✓	✓	✓					✓	CIoU	**42.4%**	**64.4%**	**45.9%**

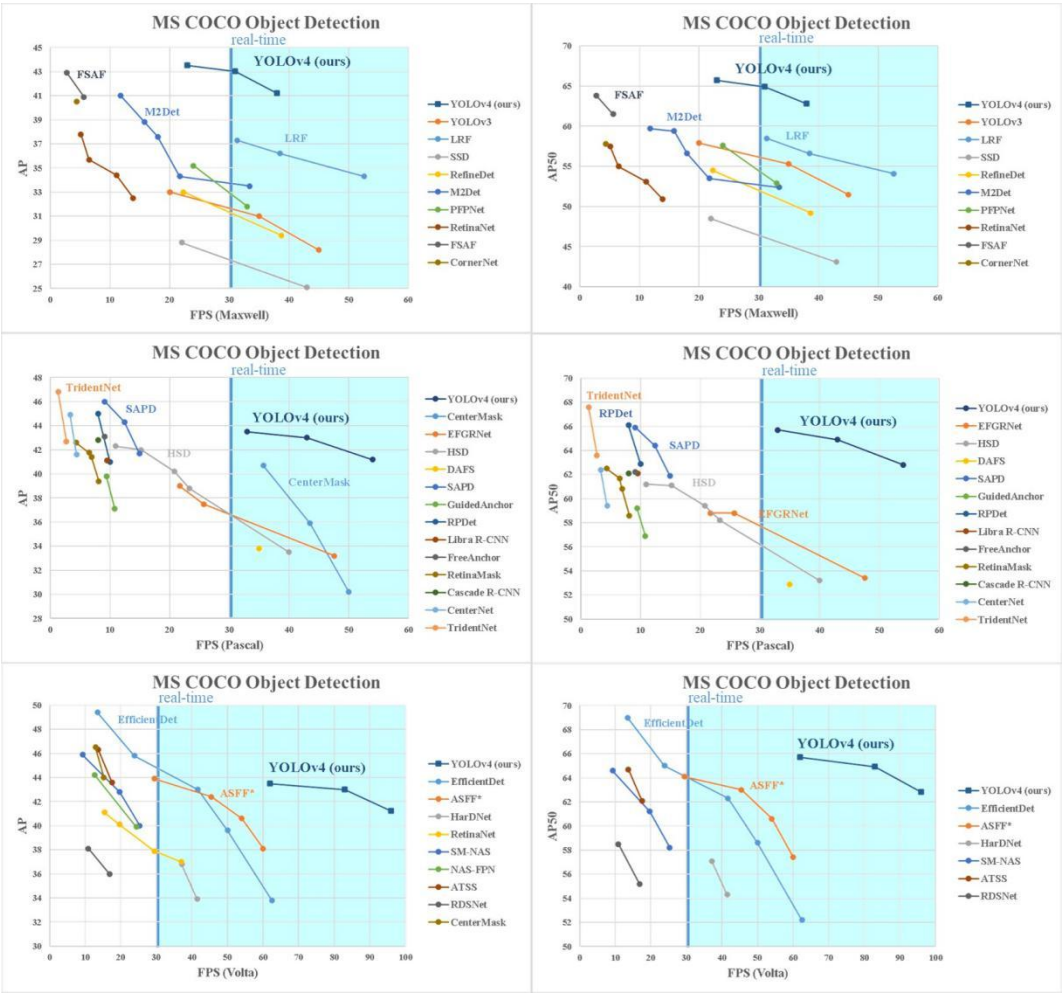


图 8：不同目标检测器的速度和精度对比（一些文章只报告了其检测器在 Maxwell/Pascal/Volta 等特定 GPU 上的 FPS）

参考文献

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5561–5569, 2017. 4

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6154–6162, 2018. 12

[3] Jiale Cao, Yanwei Pang, Jungong Han, and Xuelong Li. Hierarchical shot detector. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9705–9714, 2019. 12

[4] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. HarDNet: A low memory traffic network. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 13

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2017. 2, 4

[6] Pengguang Chen. GridMask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 3

[7] Yukang Chen, Tong Yang, Xiangyu Zhang, Gaofeng Meng, Xinyu Xiao, and Jian Sun. DetNAS: Backbone search for object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6638–6648, 2019. 2

[8] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 502–511, 2019. 7

[9] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Sys-tems (NIPS)*, pages 379–387, 2016. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with CutOut. *arXiv preprint arXiv:1708.04552*, 2017. 3

[12] Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V Le, and Xiaodan Song. SpineNet: Learning scale-permuted backbone for recognition and localization. *arXiv preprint arXiv:1912.05027*, 2019. 2

[13] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. CenterNet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6569–6578, 2019. 2, 12

[14] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C Berg. RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free. *arXiv preprint arXiv:1901.03353*, 2019. 12

[15] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[16] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. DropBlock: A regularization method for convolutional networks. In *Ad- vances in Neural Information Processing Systems (NIPS)*, pages 10727–10737, 2018. 3

[17] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7036–7045, 2019. 2, 13

[18] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014. 2, 4

[20] Jianyuan Guo, Kai Han, Yunhe Wang, Chao Zhang, Zhaohui Yang, Han Wu, Xinghao Chen, and Chang Xu. Hit-Detector: Hierarchical trinity architecture search for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[21] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. GhostNet: More features from cheap operations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[22] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 447–456, 2015. 4

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 4

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(9):1904–1916, 2015. 2, 4, 7

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. 2

[27] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetV3. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 4

[28] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2, 4

[29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 4

[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 2

[31] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size. *arXiv preprint arXiv:1602.07360*, 2016. 2

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 6

[33] Md Amirul Islam, Shujon Naha, Mrigank Rochan, Neil Bruce, and Yang Wang. Label refinement network for coarse-to-fine semantic segmentation. *arXiv preprint arXiv:1703.00551*, 2017. 3

[34] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018. 11

[35] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 971–980, 2017. 4

[36] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. FractalNet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016. 6

[37] Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 2, 11

[38] Hei Law, Yun Teng, Olga Russakovsky, and Jia Deng. CornerNet-Lite: Efficient keypoint based object detection. *arXiv preprint arXiv:1904.08900*, 2019. 2

[39] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178. IEEE, 2006. 4

[40] Youngwan Lee and Jongyoul Park. CenterMask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 12, 13

[41] Shuai Li, Lingxiao Yang, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Dynamic anchor feature selection for single-shot object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6609–6618, 2019. 12

[42] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6054–6063, 2019. 12

[43] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. DetNet: Design backbone for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–350, 2018. 2

[44] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. 2

[45] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2, 3, 11, 13

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 5

[47] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018. 2, 4, 11

[48] Songtao Liu, Di Huang, and Yunhong Wang. Learning spatial fusion for single-shot object detection. *arXiv preprint arXiv:1911.09516*, 2019. 2, 4, 13

[49] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8759–8768, 2018. 1, 2, 7

[50] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2016. 2, 11

[51] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 4

[52] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[53] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. ShuffleNetV2: Practical guidelines for efficient cnn

architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018. 2

[54] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 30, page 3, 2013. 4

[55] Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 2019. 4

[56] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 807–814, 2010. 4

[57] Jing Nie, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Enriched feature guided refinement network for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9537–9546, 2019. 12

[58] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra R-CNN: Towards balanced learning for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–830, 2019. 2, 12

[59] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 4

[60] Abdullah Rashwan, Agastya Kalra, and Pascal Poupart. Matrix Nets: A new deep architecture for object detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCV Workshop)*, pages 0–0,2019. 2

[61] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 2

[62] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7263–7271, 2017. 2

[63] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2,4, 7, 11

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 2

[65] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. 3

[66] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. 2

[67] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 761–769, 2016. 3

[68] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[69] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee. Hide-and-Seek: A data augmentation technique for weakly-supervised localization and beyond. *arXiv preprint arXiv:1811.02545*, 2018. 3

[70] Saurabh Singh and Shankar Krishnan. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. *arXiv preprint arXiv:1911.09737*, 2019. 6

[71] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. DropOut: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 3

[72] K-K Sung and Tomaso Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(1):39–51, 1998. 3

[73] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 3

[74] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. MNASNet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820–2828, 2019. 2

[75] Mingxing Tan and Quoc V Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, 2019. 2

[76] Mingxing Tan and Quoc V Le. MixNet: Mixed depthwise convolutional kernels. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 5

[77] Mingxing Tan, Ruoming Pang, and Quoc V Le. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 13

[78] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9627–9636, 2019. 2

[79] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, 2015. 6

[80] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using DropConnect. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1058–1066, 2013. 3

[81] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A new backbone that can enhance learning capability of cnn. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPR Workshop)*, 2020. 2, 7

[82] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2965–2974, 2019. 12

[83] Shaoru Wang, Yongchao Gong, Junliang Xing, Lichao Huang, Chang Huang, and Weiming Hu. RDSNet: A new deep architecture for reciprocal object detection and instance segmentation. *arXiv preprint arXiv:1912.05070*, 2019. 13

[84] Tiancai Wang, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Learning rich features at high-speed for single-shot object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1971–1980, 2019. 11

[85] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 1, 2, 4

[86] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017. 2

[87] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. RepPoints: Point set representation for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9657–9666, 2019. 2, 12

[88] Lewei Yao, Hang Xu, Wei Zhang, Xiaodan Liang, and Zhenguo Li. SM-NAS: Structural-to-modular neural architecture search for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 13

[89] Zhuliang Yao, Yue Cao, Shuxin Zheng, Gao Huang, and Stephen Lin. Cross-iteration batch normalization. *arXiv preprint arXiv:2002.05712*, 2020. 1, 6

[90] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. UnitBox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 3

[91] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. 3

[92] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. MixUp: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[93] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7151–7160, 2018. 6

[94] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 13

[95] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4203–4212, 2018. 11

[96] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. FreeAnchor: Learning to match anchors for visual object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 12

[97] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, 2018. 2

[98] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 9259–9266, 2019. 2,4, 11

[99] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU Loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3, 4

[100] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 3

[101] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides. Soft anchor-point object detection. *arXiv preprint arXiv:1911.12448*, 2019. 12

[102] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 840–849, 2019. 11