YOLOv4: Optimal Speed and Accuracy of Object Detection

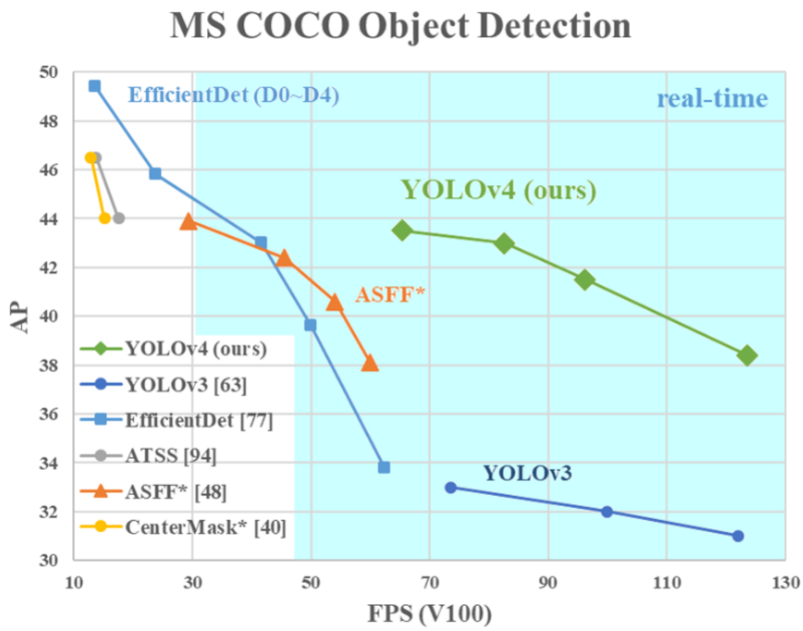
Paper Link

Abstract

We use new features: **WRC(Weighted-Residual-Connections)**, **CSP(Cross-Stage-Partial-connections)**, **CmBN(Cross mini-Batch Normalization)**, **SAT(Self-adversarial-training)**, **Mish activation**, **Mosaic data augmentation**, **DropBlock regularization**, and **CIoU loss**, and combine some of them to achieve state-of-the-art results: 43.5% AP (65.7% AP_{50}) for the MS COCO dataset at a real-time speed of ~65 FPS on Tesla V100.

我们使用了新功能: **加权残差连接，跨阶段部分连接，跨小批量归一化，自对抗训练，Mish激活函数，Mosaic 数据增强，DropBlock 正则化和CIoU 损失**，并将其中一些功能结合在一起，以获得最先进的结果：在 Tesla V100 上以 65 FPS 的实时速度对 MS COCO 数据集进行了 43.5% AP 处理（65.7% AP_{50}）。

1. Introduction



YOLOv4 runs twice faster than **EfficientDet** with comparable performance.

Improves **YOLOv3's AP** and **FPS** by 10% and 12%, respectively.

在性能相当的情况下，**YOLOv4** 的运行速度比 **EfficientDet** 快两倍。将 **YOLOv3** 的 **AP** 和 **FPS** 分别提高 10% 和 12%。

2. Related Work

2.1 Object detection models

A modern detector is usually composed of two parts, a **backbone** which is pre-trained on **ImageNet** and a **head** which is used to predict classes and bounding boxes of objects.

现代检测器通常由两部分组成，一部分是在 **ImageNet** 上预先训练好的**骨干网**，另一部分是用于预测物体类别和边界框的**头部**。

DetNAS

As to the head part, it is usually categorized into two kinds, i.e.,**one-stage object detector** and **two-stage object detector**.

头部网络通常分为两类，即**单级对象检测器**和**两级对象检测器**。

Object detectors developed in recent years often insert some layers between backbone and head, and these layers are usually used to **collect feature maps from different stages**.

We can call it the **neck** of an object detector.

近年来开发的物体检测器通常会在主干和头部之间插入一些层，这些层通常**用于收集不同阶段的特征图**。我们可以称其为物体检测器的**颈部**。

PAN

BiFPN

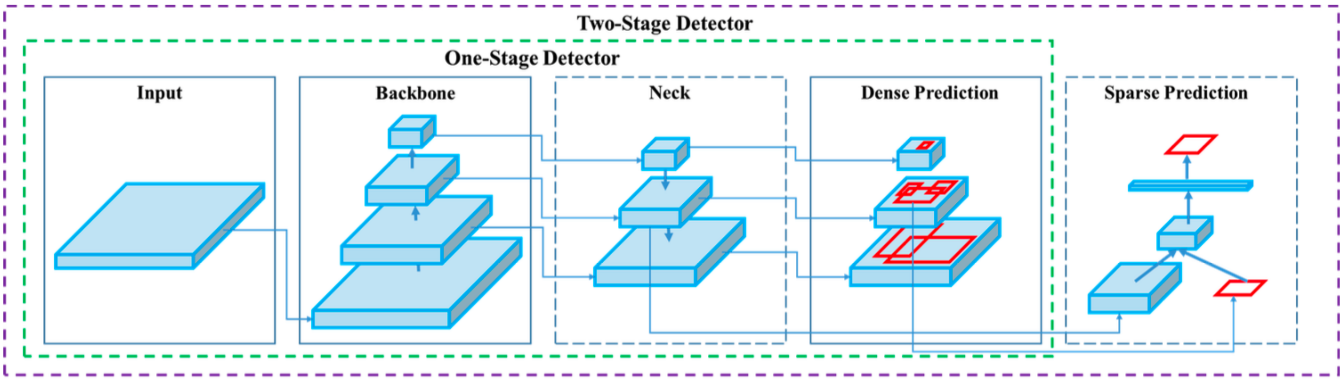
NAS-FPN

New whole model:

SpineNet

HitDetector

Supplementary for HitDetector



2.2 Bag of freebies

We call these methods that only **change the training strategy** or only **increase the training cost** as “bag of freebies.”

我们把这些只**改变训练策略**或只**增加训练成本**的方法称为 "免费包"。

What is often adopted by object detection methods and meets the definition of bag of freebies is **data augmentation**.

物体检测方法通常采用的、符合 "免费赠品包 "定义的方法是**数据增强**。

The purpose of data augmentation is to **increase the variability of the input images**, so that the designed object detection model has higher robustness to the images obtained from different environments.

数据增强的目的是**增加输入图像的可变性**，从而使设计的物体检测模型对不同环境下获得的图像具有更高的鲁棒性。

Photometric distortions and **geometric distortions** are two commonly used data augmentation method and they definitely benefit the object detection task.

In dealing with **photometric distortion**, we adjust the **brightness**, **contrast**, **hue**, **saturation**, and **noise** of an image.

For **geometric distortion**, we add random **scaling**, **cropping**, **flipping**, and **rotating**.

光度畸变和**几何畸变**是两种常用的数据增强方法，它们无疑对物体检测任务大有裨益。

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**噪点**。

对于**几何畸变**，我们会添加随机**缩放**、**裁剪**、**翻转**和**旋转**。

The data augmentation methods mentioned above are all **pixel-wise adjustments**, and **all original pixel information in the adjusted area is retained**.

上述数据增强方法都是**以像素为单位**进行调整，**调整区域内的所有原始像素信息都会保留**。

Random erase and **CutOut** can randomly select the rectangle region in an image and fill in a random or complementary value of zero.

随机擦除和**剪切输出**可以随机选择图像中的矩形区域，并填入一个随机或互补的零值。

hide-and-seek and **grid mask**, they randomly or evenly select multiple rectangle regions in an image and replace them to all zeros.

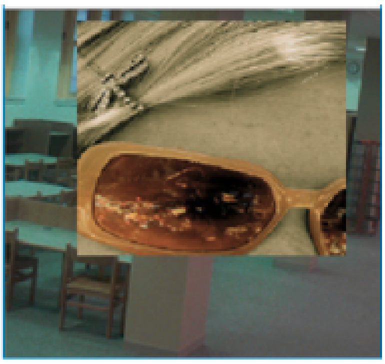
捉迷藏和**网格遮罩**，它们可以随机或均匀地选择图像中的多个矩形区域，并将其替换为所有的零。

MixUp uses two images to multiply and superimpose with different coefficient ratios, and then adjusts the label with these superimposed ratios.

MixUp使用两幅图像以不同的系数进行乘法叠加，然后根据这些叠加的系数调整标签。

CutMixi s to cover the cropped image to rectangle region of other images, and adjusts the label according to the size of the mix area.

CutMix将裁剪后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。



Some other bag of freebies methods are dedicated to solving the problem that the **semantic distribution in the dataset may have bias**. --- **imbalance between different classes**.

一些免费包方法致力于解决**数据集中的语义分布可能存在偏差**的问题——**不同类之间存在数据不平衡**。

Focal Loss

Another very important issue is that it is difficult to express the relationship of the **degree of association between different categories** with the one-hot hard representation.

另一个非常重要的问题是，很难表达**不同类别之间的关联程度**的关系。

The label smoothing proposed in **this paper** is to convert hard label into soft label for training, which can make model more robust.

In order to obtain a better soft label, **this paper** introduced the concept of **knowledge distillation** to design the label refinement network.

前者文献中提出的标签平滑法是将硬标签转换为软标签进行训练，这样可以使模型具有更多的鲁棒性。

为了获得更好的软标签，后者文献介绍了知识提炼的概念去设计标签提炼网络。

The last bag of freebies is the objective function of **Bounding Box (BBox) regression**.

The traditional object detector usually uses **Mean Square Error (MSE)** to directly perform regression on the **center point coordinates and height and width of the BBox**.

Asfor anchor-based method, it is to estimate the corresponding **offset**.

最后一袋免费包是**边界盒 (BBox) 回归**。

传统的目标检测器通常使用**平均平方误差**直接对**中心点坐标和边界框的高度和宽度**进行回归。

至于基于锚点的方法，则是估计相应的**偏移量**。

However, to directly estimate the coordinate values of each point of the BBox is to treat these points as independent variables, but in fact does not consider the integrity of the object itself.

然而，直接估算边界框各点的坐标值是将这些点视为自变量，但实际上并没有考虑边界框的完整性。

IoU loss puts the coverage of predicted BBox area and ground truth BBox area into consideration.

IoU损失预测边界框区域和真实边界框区域的覆盖率。

Because **IoU** is a **scale invariant representation**, it can solve the problem that when traditional methods calculate the **l1** or **l2 loss** of {x, y, w, h}, the loss will **increase with the scale**.

由于 **IoU**是一种尺度不变表示法，它可以解决传统方法在计算 l1 或 l2 损失时所遇到的问题，**损失会随着尺度的增大而增大**。

CIoU: simultaneously considers the overlapping area, the distance between center points, and the aspect ratio

CIoU同时考虑了重叠区域、中心点之间的距离和长宽比。

2.3 Bag of specials

For those plugin modules and post-processing methods that only increase the inference cost by a small amount but can significantly improve the accuracy of object detection, we call them “**bag**

of specials”.

对于那些只增加少量推理成本，却能显著提高物体检测精度的插件模块和后处理方法，我们称之为 "**功能包**"。

Plugin modules are for **enhancing certain attributes** in a model, such as **enlarging receptive field**, **introducing attention mechanism**, or **strengthening feature integration capability**, etc.

Post-processing is a method for **screening model prediction results**.

插件模块用于**增强模型的某些属性**，如**扩大感受野**、**引入注意机制**或**加强特征整合能力**等。后处理则是**筛选模型预测结果**的一种方法。

Common modules that can be used to **enhance receptive field** are **SPP**, **ASPP**, and **RFB**.

可用于增强感受野的常见模块有 **SPP**、**ASPP** 和 **RFB**。

SPP: split feature map into several $d \times d$ equal blocks, where d can be $\{1, 2, 3, ...\}$ thus forming spatial pyramid, and then extracting features by max-pooling operation.

分割成若干个$d \times d$的等分块，其中 d 可以是 *{1，2，3，...}*，从而形成空间金字塔，然后通过最大池化运算提取特征。

improved SPP: concatenation of max-pooling outputs with kernel size $k \times k$, where $k = \{1, 5, 9, 13\}$, and stride equals to 1.

Under this design, a relatively large $k \times k$ max-pooling effectively increase the receptive field of backbone feature.

串联最大池化输出，内核大小为 $k{\times}k$，其中 *k = {1，5，9，13}*，步长等于 1。

在这种设计下，相对较大的$k \times k$最大池化可以有效增加主干特征的感受野。

ASPP: The difference in operation between **ASPP** module and **improved SPP** module is mainly from the original $k{\times}k$ kernel size, max-pooling of stride equals to 1 to **several 3 × 3 kernel size**,

dilated ratio equals to k, and **stride equals to 1** in dilated convolution operation.

区别主要是从原来的 *k×k* 内核大小、步长等于 1 的最大池化到多个 *3×3* 内核大小、稀释比等于 k、稀释卷积操作中步长等于 1。

RFB: use several dilated convolutions of $k \times k$ kernel, dilated ratio equals to k, and stride equals to 1 to obtain a more comprehensive spatial coverage.

RFB模块使用多个 $k \times k$ 内核的稀释卷积，稀释比等于 k，步长等于 1，从而获得更全面的空间覆盖。

The **attention module** that is often used in object detection is mainly divided into **channel-wise attention** and **point-wise attention**, and the representatives of these two attention models are **Squeeze-and-Excitation (SE)** and **Spatial Attention Module (SAM)**, respectively.

物体检测中经常使用的**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是**Squeeze-and-Excitation**和**Spatial Attention Module (SAM)**。

Feature integration module: integrate low-level physical feature to high-level semantic feature.

将低层物理特征集成到高层语义特征中。

SFAM (2019): use **SE** module to execute channel-wise level re-weighting on multi-scale concatenated feature maps.

ASFF (2019) : uses softmax as point-wise level re- weighting and then adds feature maps of different scales.

BiFPN (2020) : the multi-input weighted residual connections is proposed to execute scale-wise level re-weighting, and then add feature maps of different scales.

SFAM: 使用 SE 模块对多尺度串联特征图执行信道级再加权。

ASFF: 使用 softmax 作为点向级再加权，然后添加不同尺度的特征图。

BiFPN: 使用多输入加权残差连接来执行按比例的水平再加权，然后添加不同比例的特征图。

The post-processing method commonly used in deep-learning-based object detection is **NMS**, which can be used to **filter those BBoxes that badly** predict the same object, and only r**etain the candidate BBoxes with higher response**.

基于深度学习的物体检测中常用的后处理方法是 **NMS**，它可以用来过滤那些对同一物体预测不佳的边界框，只保留响应度较高的候选边界框。

It is worth mentioning that, since none of above post-processing methods directly refer to the captured image features, **post-processing is no longer required in the subsequent development of an anchor-free method**.

值得一提的是，由于上述后处理方法都不直接参考捕捉到的图像特征，因此**在无锚点方法的后续开发中不再需要后处理**。

3. Methodology

3.1 Selection of architecture

CSPResNext50 is considerably better compared to **CSPDarknet53** in terms of **object classification** on the **ILSVRC2012 (ImageNet) dataset**.

与 **CSPDarknet53** 相比，**CSPResNext50** 在 ILSVRC2012（ImageNet）数据集上的物体分类效果要好得多。

The **CSPDarknet53** is better compared to **CSPResNext50** in terms of **detecting objects** on the **MS COCO dataset**.

在MS COCO数据集上，**CSPDarknet53** 在检测物体方面要优于 **CSPResNext50**。

Detector requires:

- Higher input network size (**resolution**) – for detecting multiple small-sized objects;
- **More layers** – for a higher receptive field to cover the increased size of input network;
- **More parameters** – for greater capacity of a model to detect multiple objects of different sizes in a single image;

检测器需要：

- 更大的输入网络尺寸（分辨率）- 用于检测多个小尺寸物体；
- 层数更多--可获得更大的感受野，以覆盖输入网络的更大尺寸；
- 更多参数--提高模型在一张图像中检测多个不同大小物体的能力；

The influence of the **receptive field** with different sizes:

- Up to the object size - allows viewing the entire object;
- Up to networksize-allows viewing the context around the object;
- Exceeding the network size - increases the number of connections between the image point and the final activation;

不同尺寸的感受野的影响：

- 与物体大小相当--可观察整个物体；
- 达到网络大小--可以查看物体周围的环境；
- 超过网络大小--增加图像点与最终激活之间的连接数量；

YOLOv4 final architecture

CSSPDarknet53	
SPP block	
PANet	instead of the FPN used in YOLOv3
YOLOv3 head	

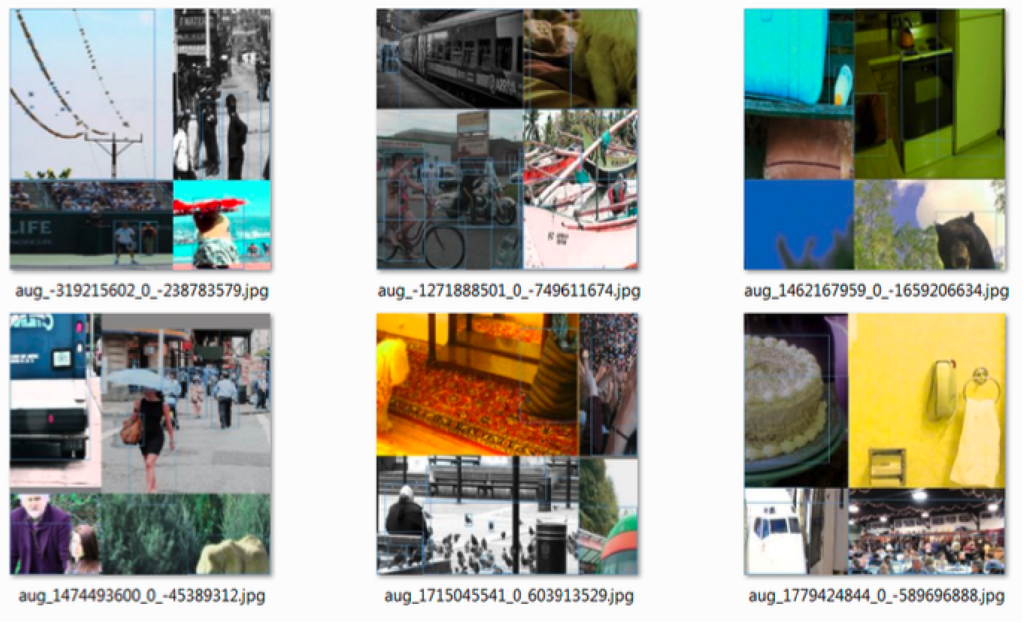
3.2 Selection of BoF and BoS

DropBlock

3.3 Additional improvements

- new method of data augmentation **Mosaic**, and **Self-Adversarial Training (SAT)**
- select optimal hyper-parameters while applying **genetic algorithms**
- modify some exsiting methods to make our design suitble for efficient training and detection - **modified SAM**, **modified PAN**, and **Cross mini-Batch Normalization(CmBN)**
- 引入了一种新的数据增强方法 **Mosaic 和自对抗训练 (SAT)**
- **应用遗传算法**的同时选择最优超参数
- 修改了一些现有方法，使我们的设计适合高效训练和检测--**修改后的 SAM**、**修改后的 PAN 和交叉小批量正则化（CmBN）**

Mosaic represents a new data augmentation method that **mixes 4 training images**. Thus 4 different contexts are mixed, while **CutMix** mixes only 2 input images.


aug_-319215602_0_-238783579.jpg
aug_-1271888501_0_-749611674.jpg
aug_1462167959_0_-1659206634.jpg
aug_1474493600_0_-45389312.jpg
aug_1715045541_0_603913529.jpg
aug_1779424844_0_-589696888.jpg

Self-Adversarial Training (SAT) represents a new data augmentation technique that operates in **2 forward backward stages**.

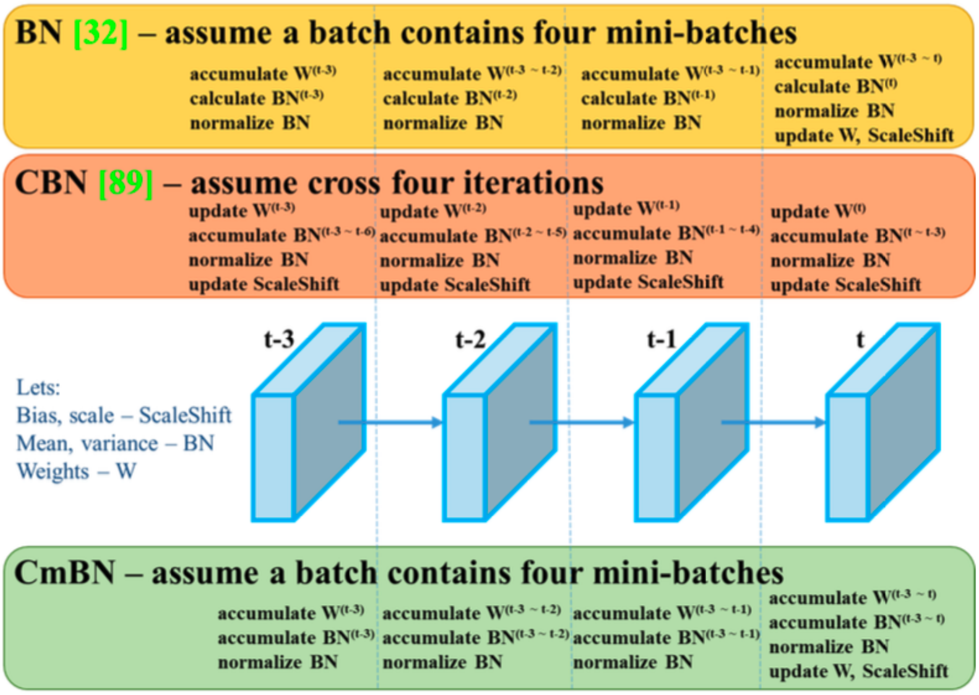
1. In the **1st stage** the neural network **alters the original image instead of the network weights**. In this way the neural network executes an adversarial attack on itself, altering the original image to create the deception that there is no desired object on the image.

2. In the **2nd stage**, the neural network is trained to detect an object on this modified image in the normal way.

自适应训练是一种新的数据增强技术，它分为**两个前向后的阶段**。

1. 在第一阶段，**神经网络改变的是原始图像而不是网络权重**。这样，神经网络就会对自己实施对抗性攻击，改变原始图像，制造图像上没有目标的假象。

2. 在第二阶段，对神经网络进行训练，使其能够以正常的方式检测出被修改图像上的物体。

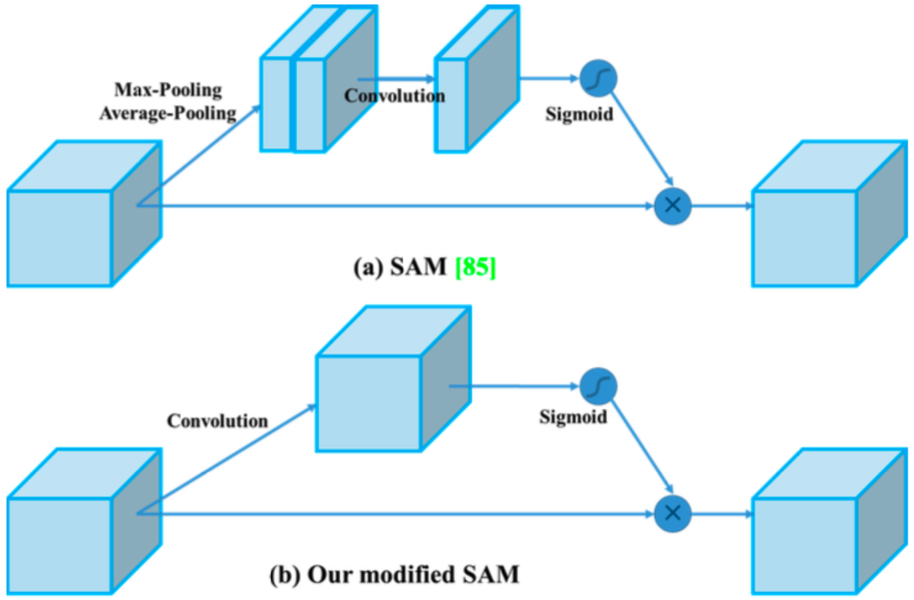


Cross mini-Batch Normalization (CmBN) collects statistics only between mini-batches within a single batch, represents a **CBN** modified version.

CmBN代表**CBN**的改进版本，只收集单个批次内迷你批次之间的统计数据。

Modify SAM from spatial-wise attention to point-wise attention.

将**SAM**从空间关注改为点关注。

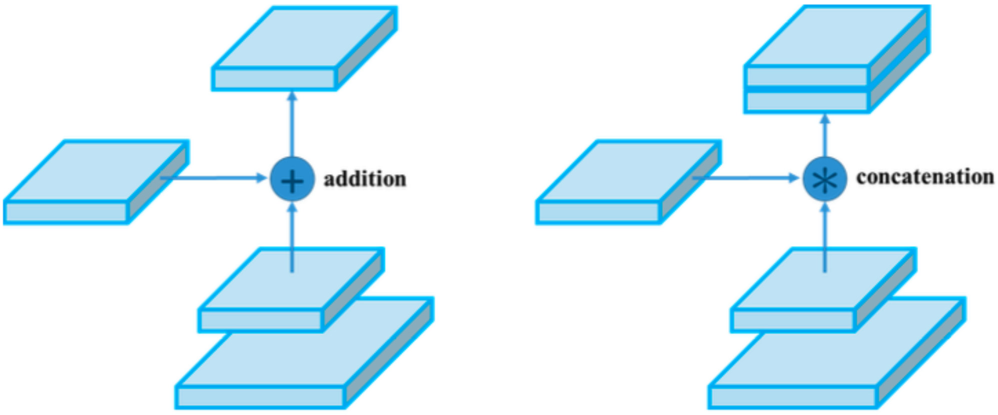


(a) SAM [85]

(b) Our modified SAM

Replace shortcut connection of **PAN** to concatenation.

将**PAN**的快捷连接改为串联。



(a) PAN [49]

(a) Our modified PAN

3.4 YOLOv4

Details of YOLOv4:

- **Backbone**: CSPDarknet53 (2020CVPR)
- **Neck**: SPP, PAN (2018CVPR)
- **Head**: YOLOv3
- **BoF for backbone**: CutMix and Mosaic data augmentation, DropBlock regularization, Class label smoothing

- **BoS for backbone**: Mish activation, Cross-stage partial connections (CSP), Multi- input weighted residual connections (MiWRC)
- **BoF for detector**: CIoU-loss, CmBN, DropBlock regularization, Mosaic data augmentation, Self-Adversarial Training, Eliminate grid sensitivity, Using multiple anchors for a single ground truth, Cosine annealing scheduler [52], Optimal hyperparameters, Random training shapes
- **BoS for detector**: Mish activation, SPP-block, SAM-block, PAN path-aggregation block, DIoU-NMS

4. Experiments

Test the influence of different training improvement techniques on accuracy of the classifier on **ImageNet (ILSVRC 2012 val) dataset**, and then on the accuracy of the detector on **MS COCO (test-dev 2017) dataset**.

在 **ImageNet ILSVRC 2012 val**数据集上测试了不同训练改进技术对分类器准确性的影响，然后在 **MS COCO test-dev 2017**数据集上测试了检测器准确性的影响。

4.1 Experimental setup

In ImageNet classification experiment:

- training steps -> 8000000
- batch size -> 128
- mini-batch size -> 32
- the polynomial decay learning rate scheduling strategy initial learning rate 0.1

  多项式衰减学习率调度策略，初始学习率 0.1
- warm-up steps -> 1000
- momentum -> 0.9
- weight decay -> 0.005

In MS COCO object detection experiments:

- training steps -> 500500
- batch size -> 64
- mini-batch size -> 8 or 4
- step decay learning rate scheduling strategy initial learning rate 0.01, multiply with a factor 0.1 at the 400,000 steps and the 450,000 steps, respectively.

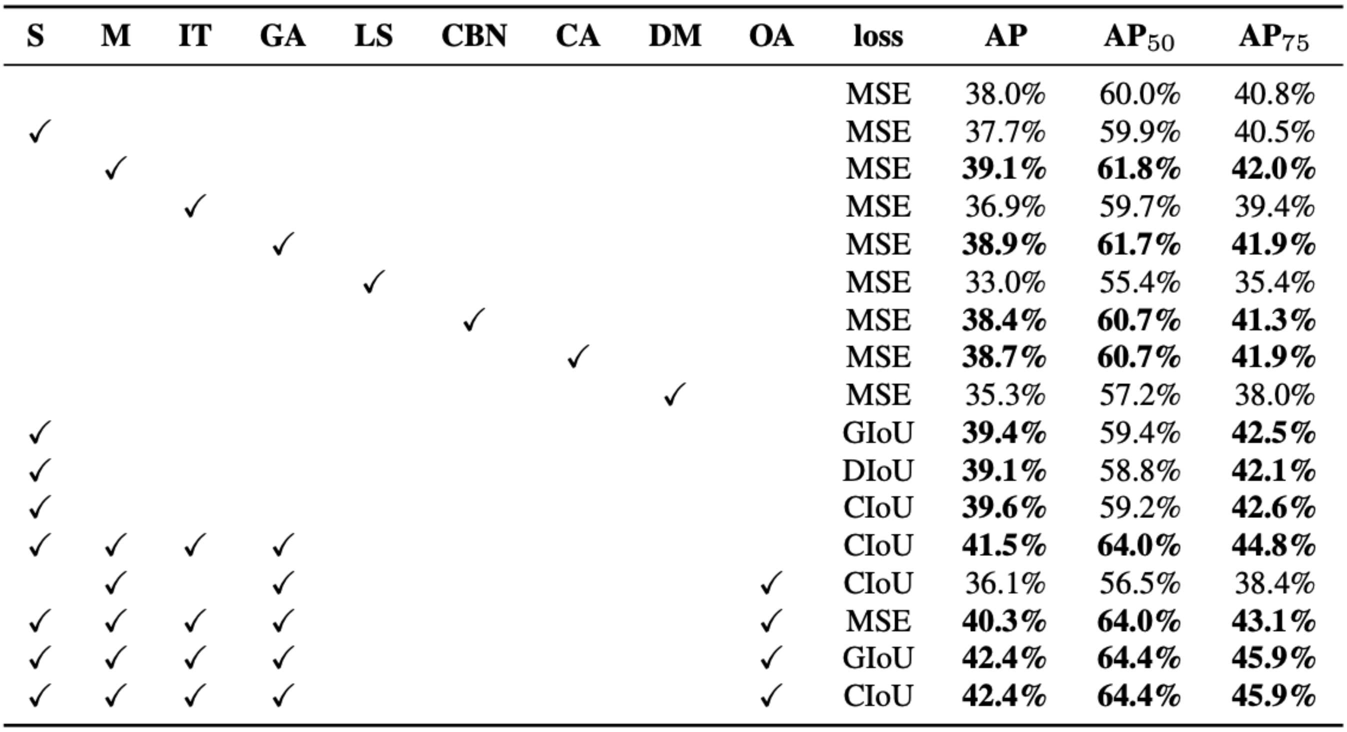
  步长衰减学习率调度策略的初始学习率为 0.01，在 400 000 步和 450 000 步时分别乘以系数 0.1。
- momentum -> 0.9
- weight decay -> 0.0005
- genetic algorithm -> YOLOv3-SPP, GIoU loss, 300 epochs, searched learning rate 0.00261, momentum 0.949, IoU threshold for assigning ground truth 0.213, and loss normalizer 0.07

4.2 Influence of different features on Classifier training

The classifier's accuracy is improved by introducing the features such as: **CutMix** and **Mosaic** data augmentation, **Class label smoothing**, and **Mish** activation.

通过引入 **CutMix** 和 **Mosaic** 数据增强、**类别标签平滑**和 **Mish** 激活等特征，分类器的准确率得到了提高。

4.3 Influence of different features on Detector training



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%**

- **S**: Eliminate grid sensitivity

  消除格子敏感性

- **M**: Mosaic data augmentation, 4 images

  马赛克数据增强技术，混合4张图片

- **IT**: IoU threshold

- **GA**: Genetic algorithms, first 10% periods

  遗传算法，在前10%的训练时期

- **LS**: Class label smoothing for sigmoid activation

  类别标签平滑技术，使用sigmoid激活函数

- **CBN**: CmBN, Cross mini-Batch Normalization for collecting statistics inside the entire batch

  跨小批次归一化，用于收集整个批次内的统计数据

- **CA**: Cosine annealing scheduler

  余弦退火调度程序

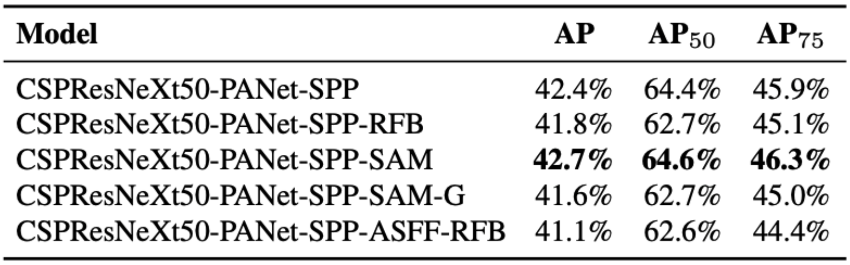
- **DM**: Dynamic mini-batch size - automatic increase of mini-batch size during small resolution training by using Random training shapes

动态迷你批量大小，通过使用随机训练尺寸，在小分辨率训练期间自动增加迷你批量大小

- **OA**: Optimized Anchors, using the optimized anchors for training with the 512x512 network resolution

  以 512x512 网络分辨率进行训练

- **loss**: GIoU, CIoU, DIoU, MSE for bounded box regression



Model	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%

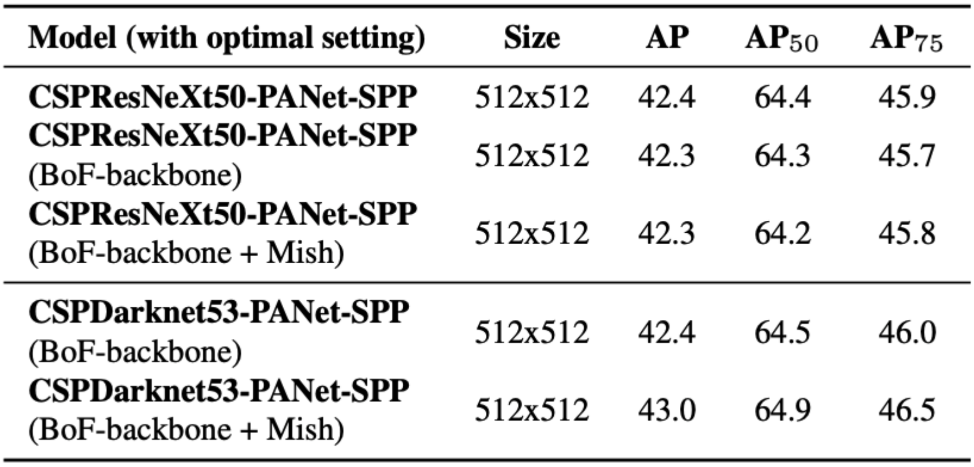
The detector gets best performance when using **SPP**, **PANet**, and **SAM**.

在使用**SPP**、**PANet**、**SAM**时，检测器达到最佳性能。

4.4 Influence of different backbones and pretrained weightings on Detector training

We notice that the model characterized with the best classification accuracy is not always the best in terms of the detector accuracy.

我们注意到，分类准确率最高的模型并不总是检测器准确率最高的模型。



Model (with optimal setting)	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

In classification task, **CSPResNeXt50** is better; In object detection task, **CSPResNeXt53** is better.

在分类任务中，**CSPResNeXt50**表现更好；在目标检测任务中，**CSPResNeXt53**表现更好。

Using **BoF** and **Mish** for the **CSPResNeXt50** classifier training increases its classification accuracy, but further application of these pre-trained weightings for detector training **reduces the detector accuracy**.

在 **CSPResNeXt50** 分类器训练中使用 **BoF** 和 **Mish** 可以提高分类器的分类精度，但在检测器训练中进一步使用这些预训练加权则会降低检测器的精度。

However, using **BoF** and **Mish** for the **CSPDarknet53** classifier training increases the accuracy of both the classifier and the detector which uses this classifier pre-trained weightings.

然而，在 **CSPDarknet53** 分类器训练中使用 **BoF** 和 **Mish** 会提高分类器和使用该分类器预训练加权的检测器的准确性。

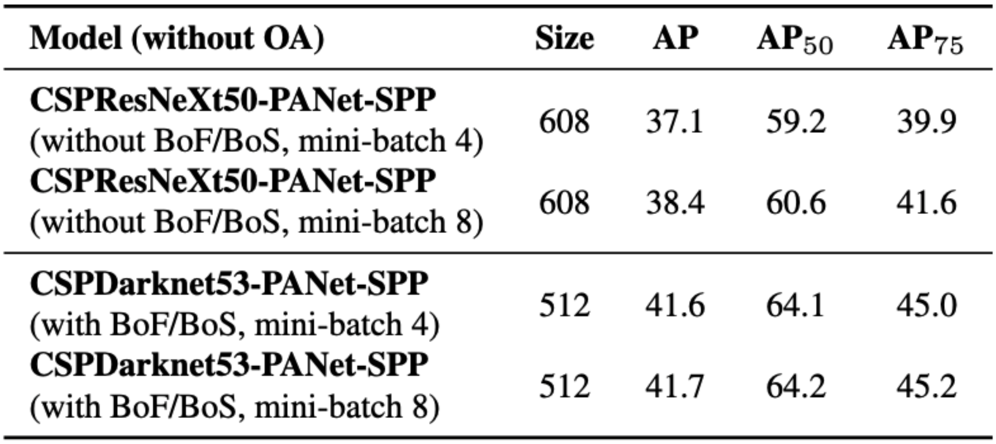
CSPDarknet53 is more suitable for the detector than for **CSPResNeXt50**.

CSPDarknet53 比 **CSPResNeXt50** 更适合检测器。

The **depth** of **CSPDarknet53**, **bottleneck ratio**, **width growth ratio** between stages are 65, 1, and 2, respectively.

CSPDarknet53 的深度、瓶颈比、级间宽度增长比分别为 65、1 和 2。

4.5 Influence of different mini-batch size on Detector training



Model (without 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

The **mini-batch size** has **almost no effect** on the detector’s performance.

5. Results

6. Conclusions

The original concept of one-stage anchor-based detectors has proven its viability. We have verified a large number of features, and selected for use such of them for improving the accuracy of both the classifier and the detector.

基于锚点的单级检测器的原始概念已经证明了其可行性。我们验证了大量特征，并选择了其中一些特征用于提高分类器和检测器的准确性。