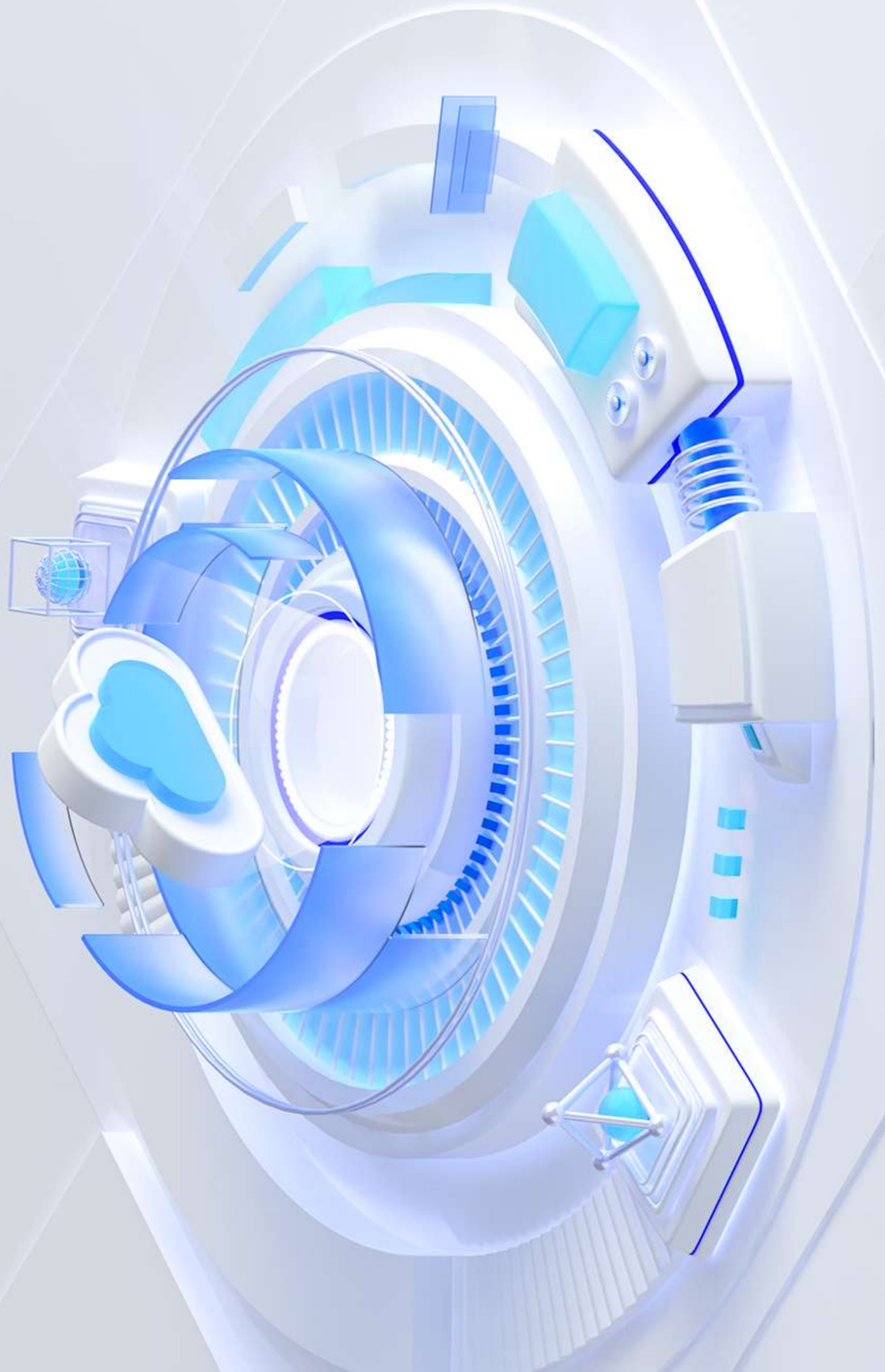
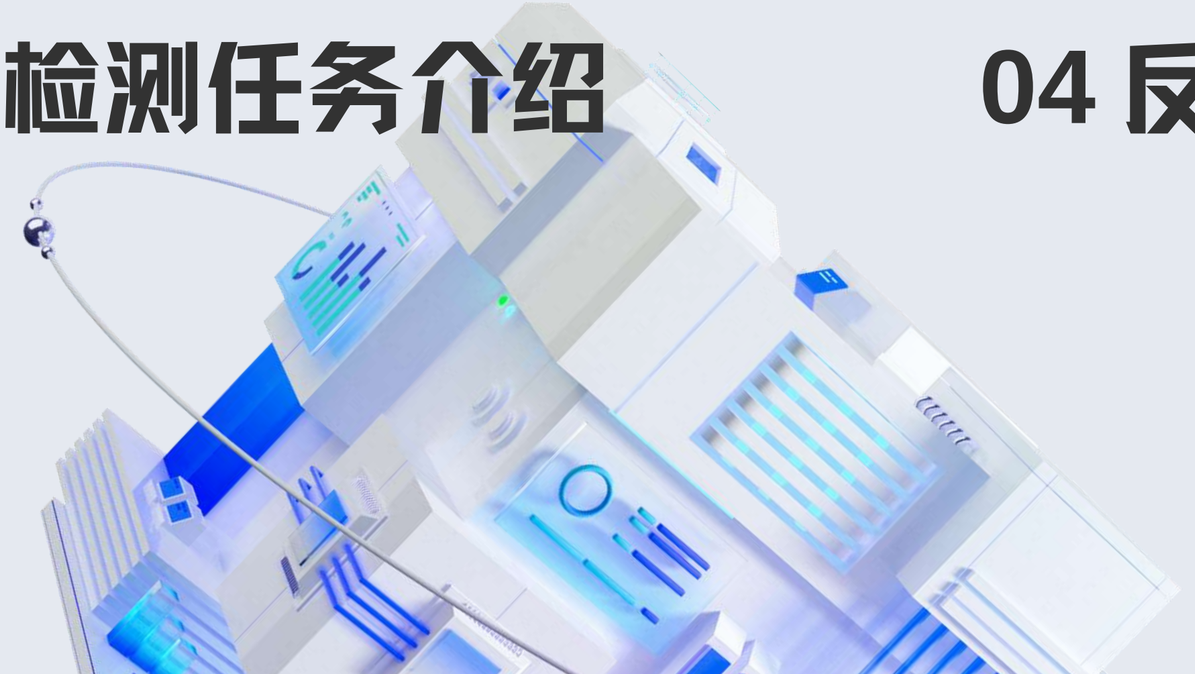
基于FCOS的2D图像目标检测器



汇报人：刘卓瀚

目录



01

二维目标检测任务介绍

二维目标检测任务

输入：一张2D图片

输出：一个带类别标志的目标检测框

主要困难

1 一张图片可能有多个目标（且类别可能不同），需要把所有目标都检测出来

2 图片尺寸、分辨率不同

3 物体尺寸不同

02

FCOS

FULLY CONVOLUTIONAL ONE-STAGE OBJECT DETECTION

FCOS网络架构

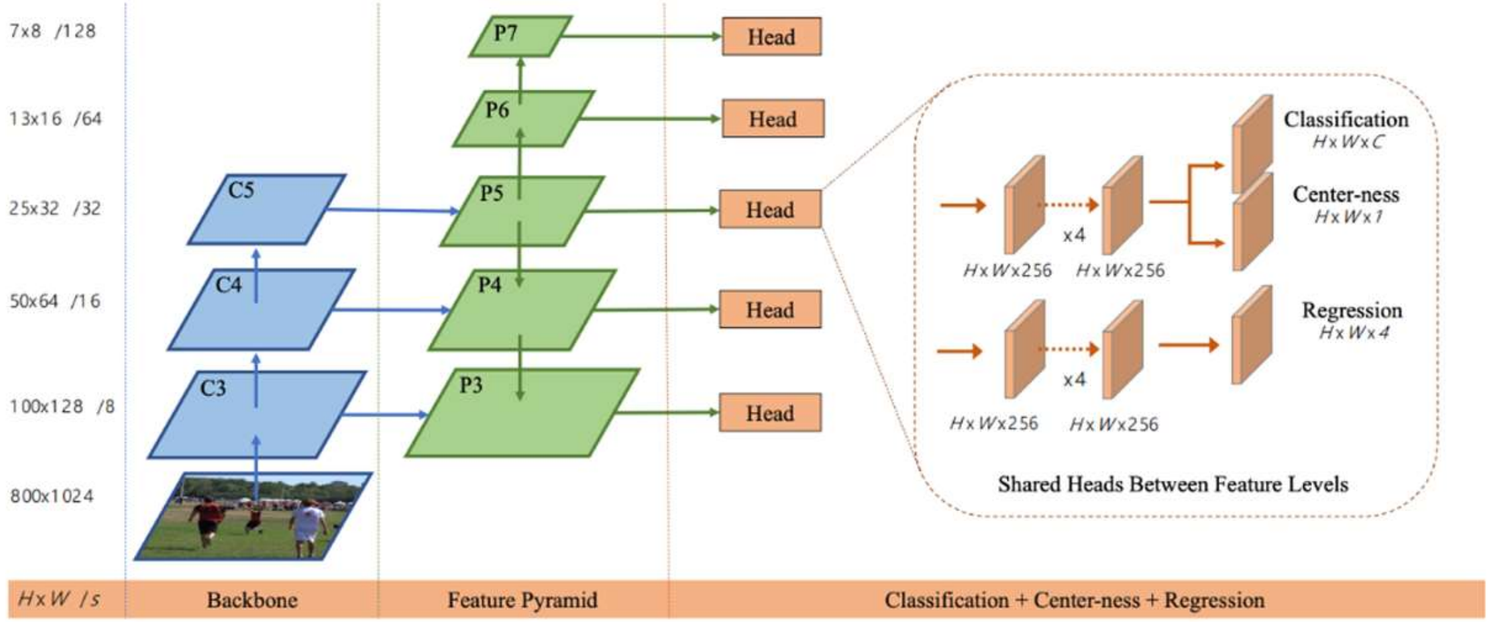


Figure 2 – The network architecture of FCOS, where C3, C4, and C5 denote the feature maps of the backbone network and P3 to P7 are the feature levels used for the final prediction. $H \times W$ is the height and width of feature maps. '/s' ($s = 8, 16, ..., 128$) is the down-sampling ratio of the feature maps at the level to the input image. As an example, all the numbers are computed with an 800×1024 input.

https://arxiv.org/pdf/1904.01355.pdf

FCOS网络架构

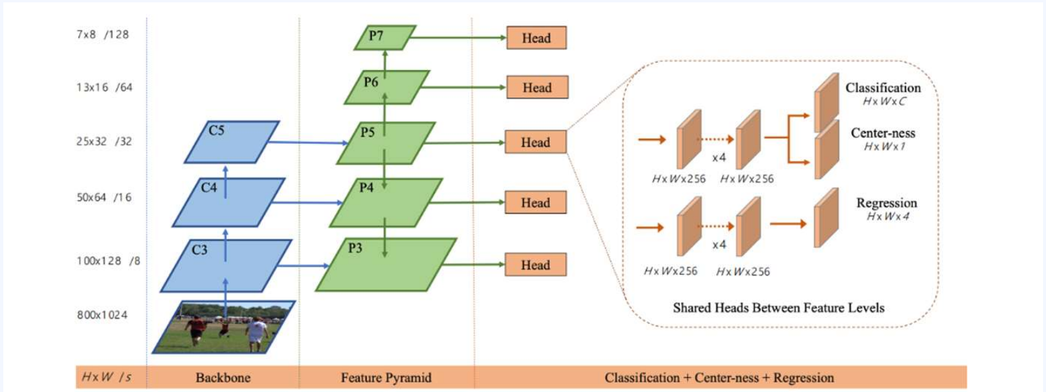


Figure 2 – The network architecture of FCOS, where C3, C4, and C5 denote the feature maps of the backbone network and P3 to P7 are the feature levels used for the final prediction. $H \times W$ is the height and width of feature maps. '$/s$' ($s = 8, 16, ..., 128$) is the down-sampling ratio of the feature maps at the level to the input image. As an example, all the numbers are computed with an 800×1024 input.

BACKBONE NETWORK：使用一个预训练的CNN特征提取网络，获取图像的FEATURE MAP

特征提取网络可以是resnet、vggnet、regnet等等

FCOS网络架构：FPN

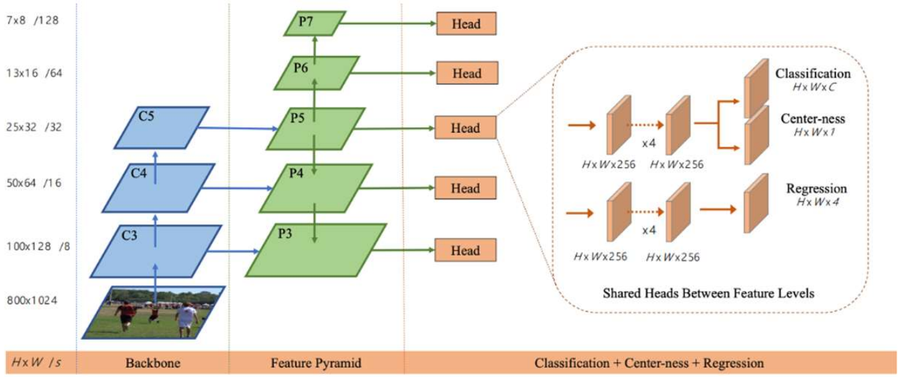
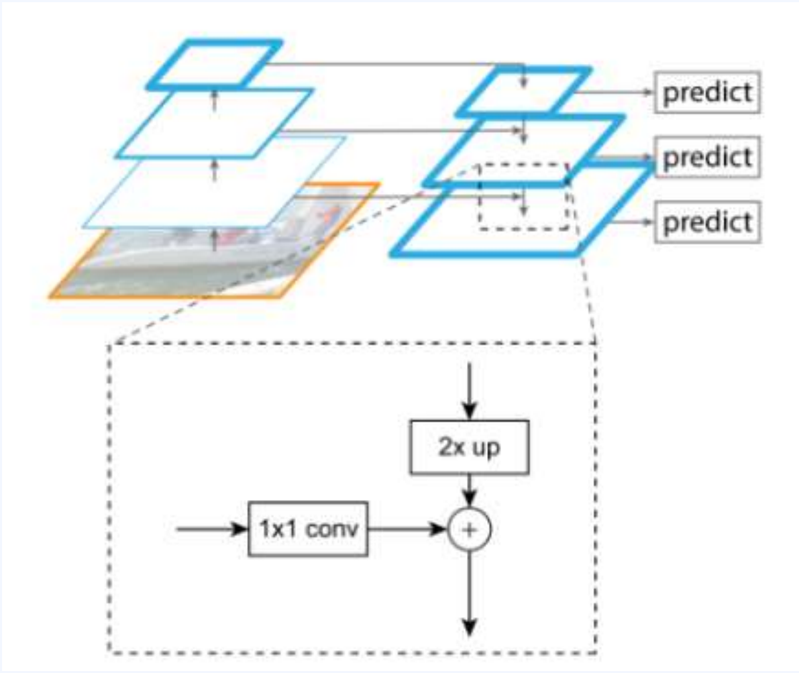


Figure 2 – The network architecture of FCOS, where C3, C4, and C5 denote the feature maps of the backbone network and P3 to P7 are the feature levels used for the final prediction. $H \times W$ is the height and width of feature maps. '/s' ($s = 8, 16, ..., 128$) is the down-sampling ratio of the feature maps at the level to the input image. As an example, all the numbers are computed with an 800×1024 input.



FPN: FEATURE PYRAMID NETWORK

- BOTTOM-UP：自底向上进行CNN得到特征图，然后对特征图进行1 * 1 CONVOLUTION（目的是对齐通道数）
- top-down：对高层特征图进行上采样，通常的方法是最近邻插值法，使得高层特征图的大小和低层特征图的大小一样
- lateral connections：将高层特征图和底层特征图进行融合，通常是简单的逐元素相加，得到一个更加丰富的特征图
- 3 * 3 convolution：对融合后的特征图进行3 * 3的卷积操作，得到最终的特征图

为何使用FPN

- 为了解决多尺度的问题，即有些物体很大，包含很多像素点，但也有些物体很小，只包含很少的像素点
- 对于多层的FEATURE MAP，底层FEATURE MAP往往包含物体的位置信息，高层FEATURE MAP往往包含物体的语义信息
- 在一些目标检测器中，只使用了最后一层的FEATURE MAP，这样可能会导致小物体的信息丢失（比如我们得到的特征图是60 * 60，而原图是600 * 600，那么特征图上一个像素点对应原图上的10 * 10的像素点，那么原图上可能有许多小于10 * 10像素点的物体，它们的信息就容易丢失，或说至少位置信息容易丢失）
- 通过将高层特征和底层特征进行融合，让识别和定位都更加准确

FCOS网络架构

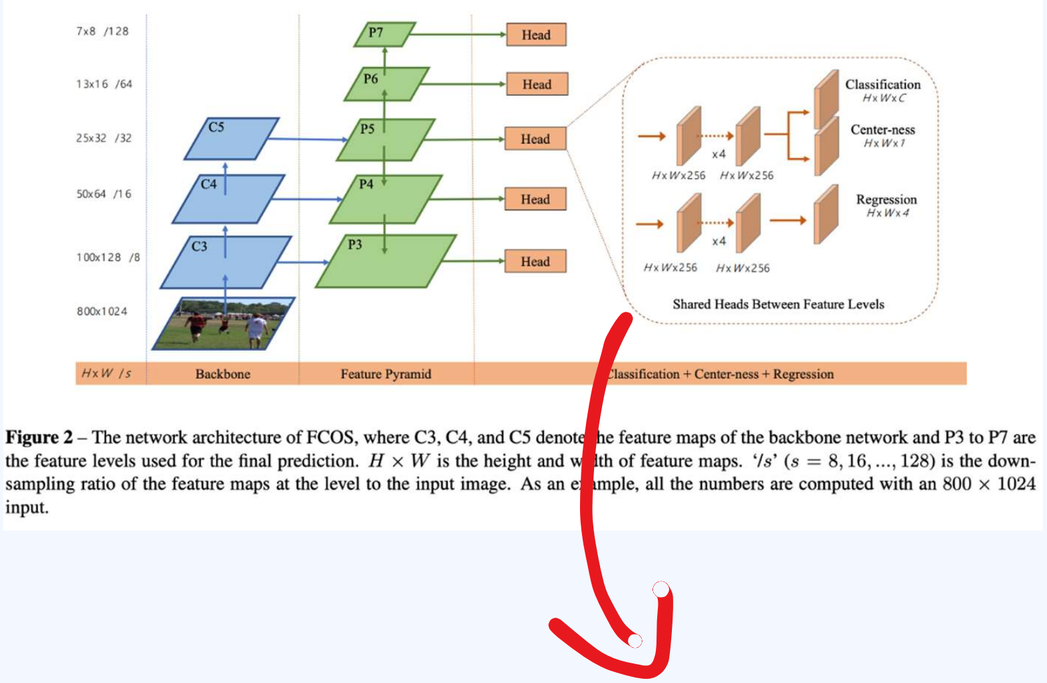
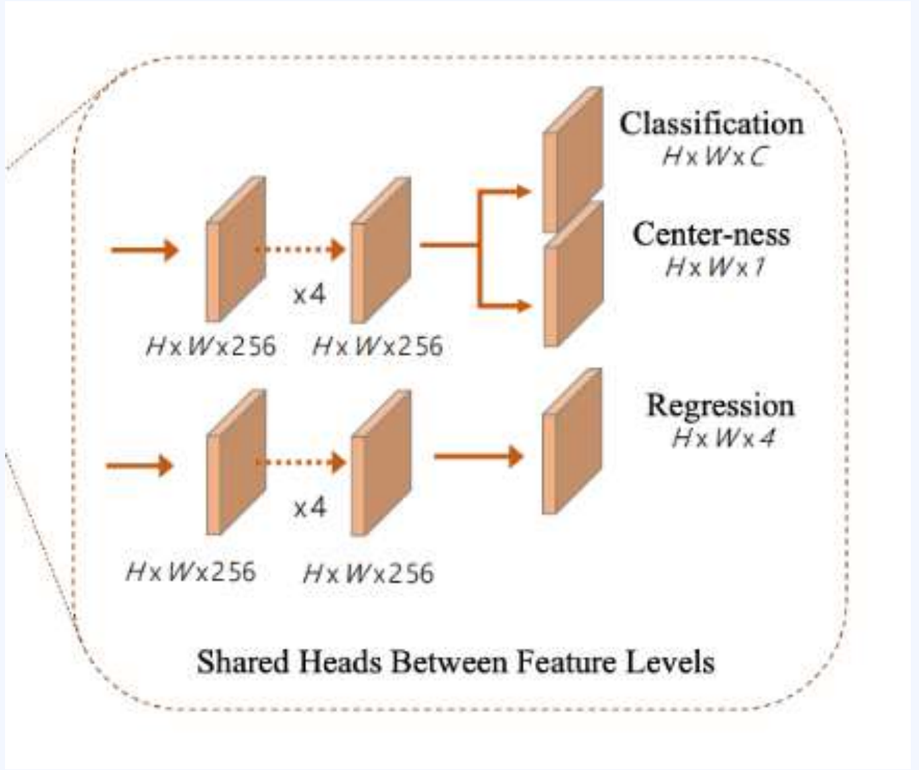


Figure 2 – The network architecture of FCOS, where C3, C4, and C5 denote the feature maps of the backbone network and P3 to P7 are the feature levels used for the final prediction. $H \times W$ is the height and width of feature maps. '/s' ($s = 8, 16, ..., 128$) is the down-sampling ratio of the feature maps at the level to the input image. As an example, all the numbers are computed with an 800×1024 input.



- 对于FPN输出的FEATURE MAP的**每一个点**，都预测类别、CENTERNESS、BOX REGRESSION，通过CNN来预测，所以可以方便的对每一个点都进行预测
- 我的实现中没有采用4个256通道的卷积，而是采用了2个128通道的卷积，KERNEL SIZE = 3，PADDING = STRIDE = 1
- 之后有研究表明CENTERNESS和REGRESSION的预测放在一起比CLASSIFICATION和CENTERNESS的预测放在一起更好，我的实现采用了这种方法

loss的计算

对于每一层FEATURE MAP的每一个点，都计算LOSS

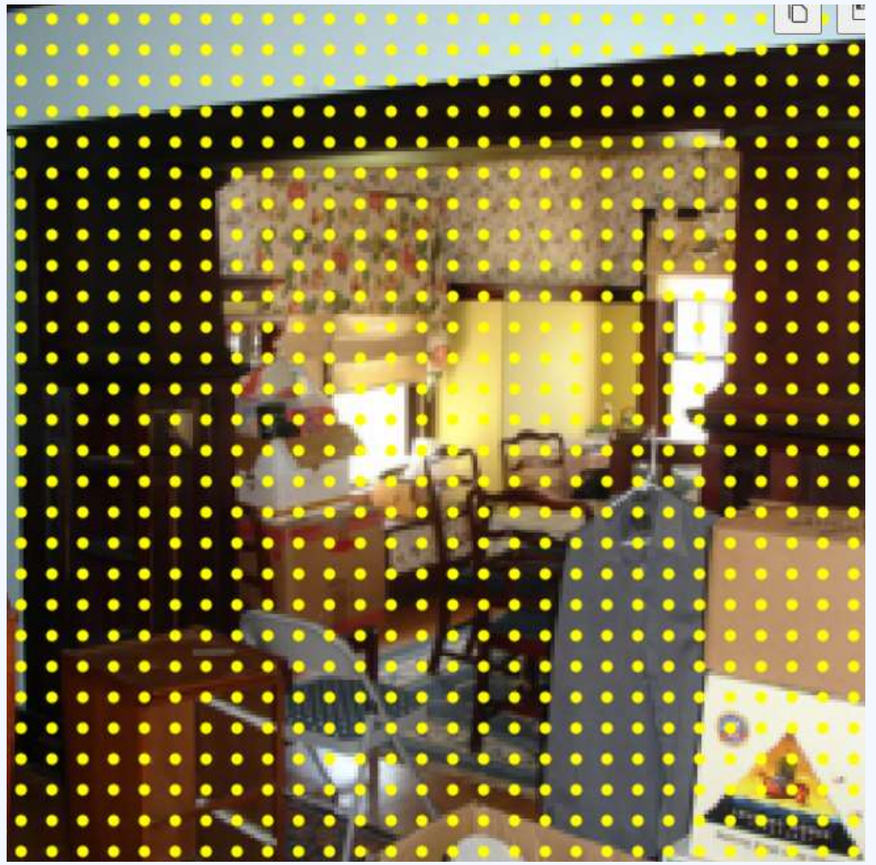
我的实现中只用了P3、P4、P5层的FEATURE MAP，大小分别是(28,28),(14,14),(7,7)

LOSS分别有三项：分类的LOSS，CENTERNESS的LOSS，BOX REGRESSION的LOSS

- 最后的LOSS是所有LOSS的总和对图中GROUND TRUTH正类个数（即不为背景）求平均

ground truth的获得

- 对于feature map的每一个像素点，都会匹配一个相对于原图的中心点（根据stride（即放缩比例）的大小，均匀地匹配上去）；比如原图是224x224，feature map是28x28，那么feature map的每一个像素点对应原图的8x8的区域，其中心点的坐标依次为(4,4),(4,12),(4,20),...,(4,220),(12,4),(12,12),...,(220,220)



对于每张图片，为不同FEATURE MAP对应的中心点匹配GT BOX（GROUND TRUTH BOX），匹配规则是：

- 如果此中心点位于某个GT BOX内，那么将其匹配到这个GT BOX；如果在多个GT BOX内，那么匹配到面积最小的那个GT BOX
- 如果此中心点不在任何GT BOX内，那么将其匹配为BACKGROUND，对应GT BOX为(-1,-1,-1,-1,-1)
- 要根据FEATURE MAP的放缩比例，候选的GT BOX并不是所有的GT BOX，而是匹配合适大小的GT BOX，因为还有上一级和下一级的FEATURE MAP，上一级会匹配更大的GT BOX，下一级会匹配更小的GT BOX

loss-classification

采用sigmoid_focal_loss对分类结果计算loss

即先应用sigmoid函数，再计算focal loss

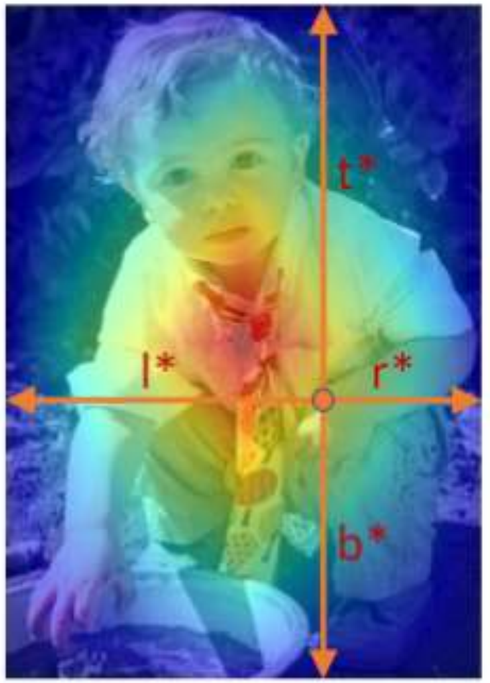
$$L_{fl} = \begin{cases} -\alpha(1-\hat{y})^{\gamma}\log\hat{y}, \text{当} y=1 \\ -(1-\alpha)\hat{y}^{\gamma}\log(1-\hat{y}), \text{当} y=0 \end{cases}$$

loss-box regression

预测的不是bounding box的绝对位置，而是相对于中心点的偏移大小LTRB（left、top、right、bottom），并且要除以特征图相对于原图的放缩比例

ground truth也需要根据中心点和gt box计算LTRB

loss是两个LTRB的L1范数



loss-centerness

ground truth：用如上公式计算centerness

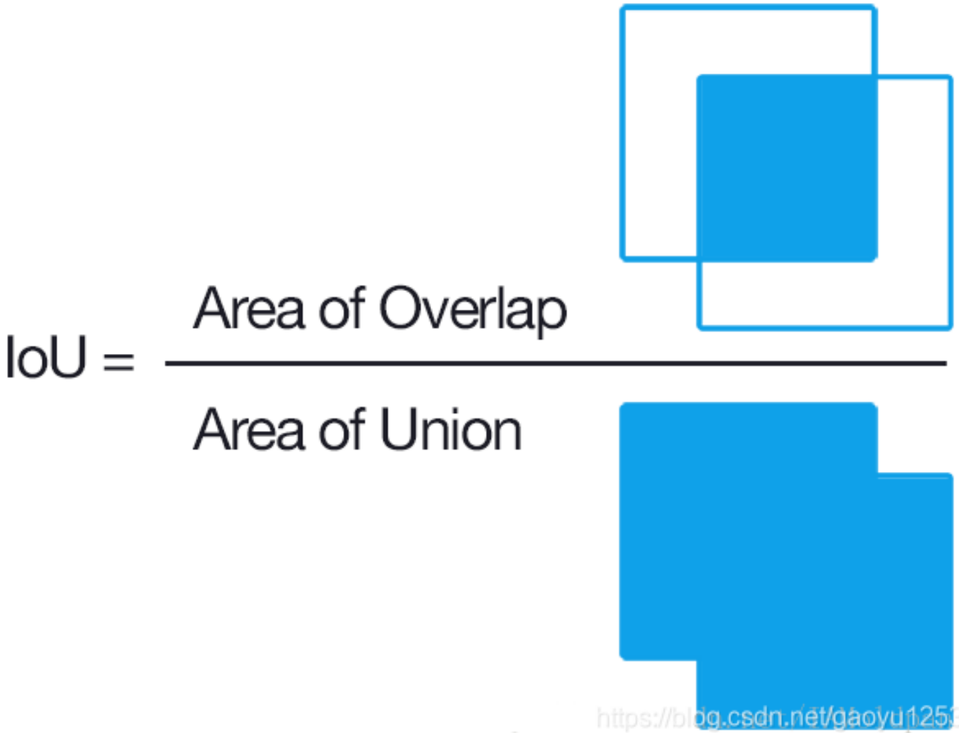
$$centerness = \sqrt{\frac{\min(left, right) \cdot \min(top, bottom)}{\max(left, right) \cdot \max(top, bottom)}}$$

与预测的centerness计算cross entropy loss

为何要搞个CENTERNESS：

最后推理的时候，置信度是由类别概率和CENTERNESS相乘得到的；由于是对每个像素点都进行预测，所以势必会有许多重复的检测，而一些检测框可能质量比较低（即中心点非常偏离检测框的中心位置），这样CENTERNESS就会比较小，通过CENTERNESS可以过滤掉这些质量比较低的检测框

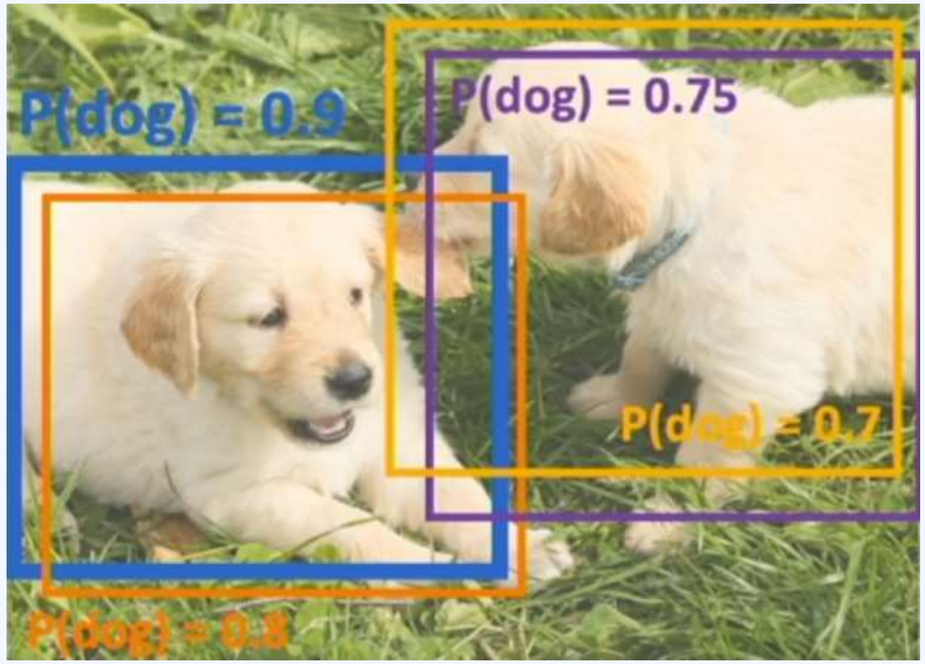
iou



IOU是两个区域重叠的部分除以
两个区域的集合部分得出的结果

NMS算法

Non-Max Suppression：
为了应对多重检测框的问题



算法步骤：

设定一个NMS_THRESHOLD，对于输入的BOUNDING BOXES进行以下步骤：

1.对于某个类别，选取BOXES中这个类别的预测分数最大的哪一个，记为BOX_BEST，并保留它

2.计算BOX_BEST与其余的BOX的IOU

3.如果其IOU>NMS_THRESHOLD了，那么就舍弃这个BOX

4.对于剩余的BOXES，GOTO STEP 1

03

实验

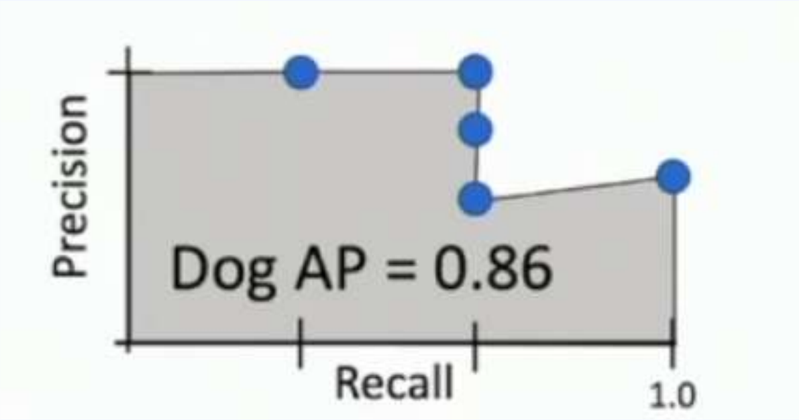
实验

代码框架参考课程UMich EECS 498的ASSIGNMENT 4

实现了FCOS目标检测器，使用PASCAL VOC 2007数据集（20个数据类别），在训练集上训练了9000轮，在测试集上进行了测试并且得出mAP指标，在电视剧《狂飙》中截了一些图来测试模型的实际效果

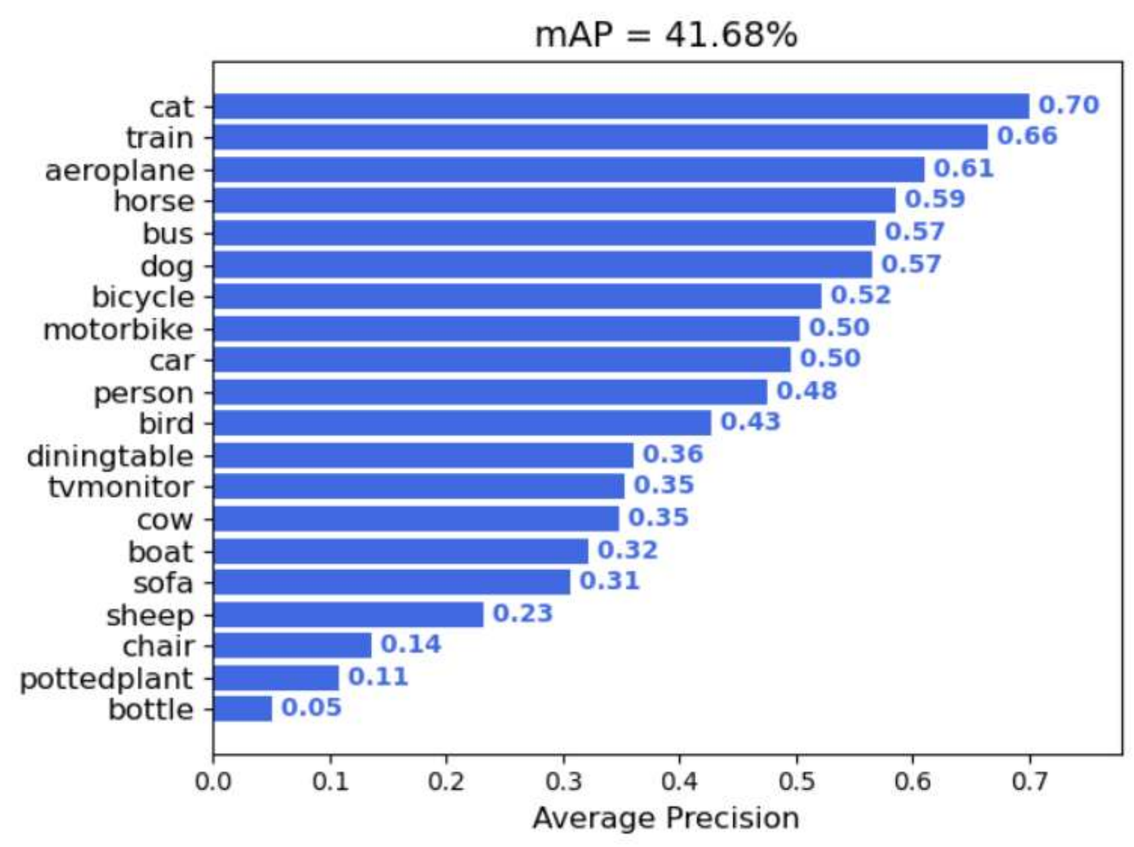
项目代码详见我的代码仓库https://github.com/anzeameol/UMich-EECS498

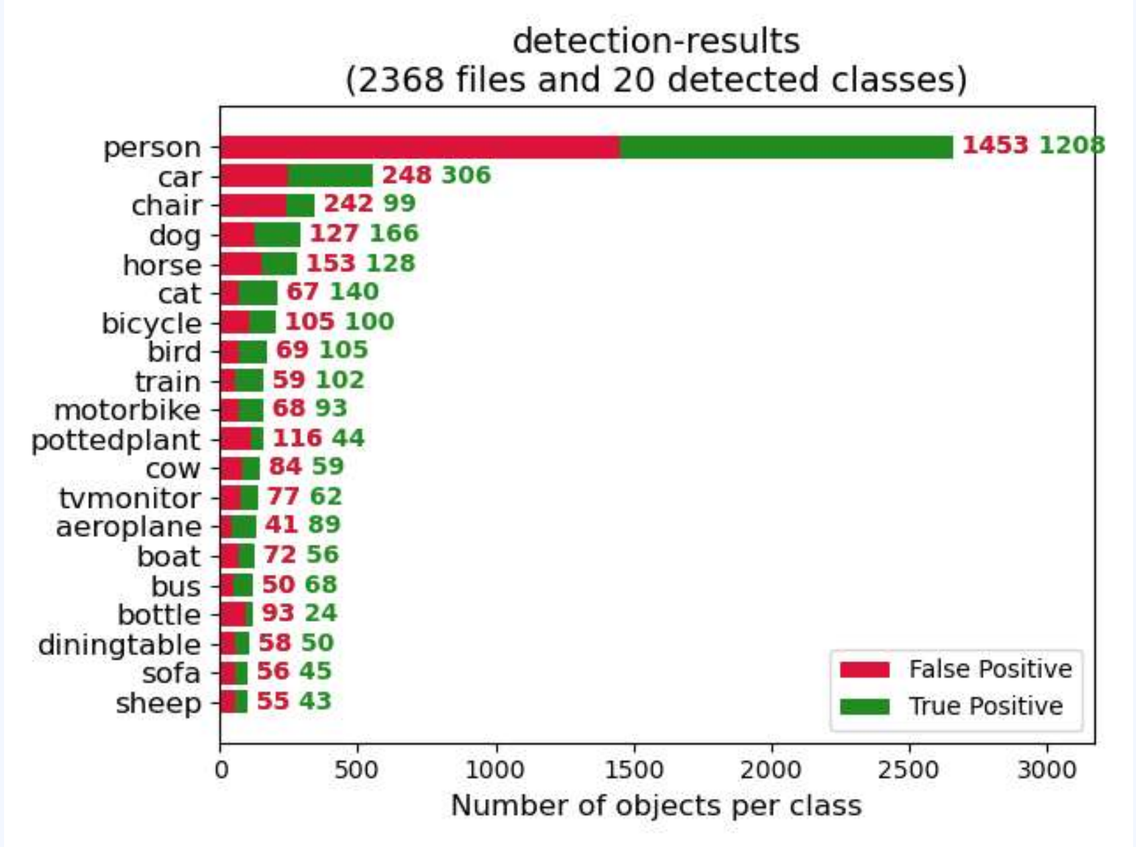
评价指标：mAP@N



- 对于某个类别，如果GROUND TRUTH BOUNDING BOX和预测的BOUNDING BOX之间的IOU大于N，就标记为正例，否则标记为负例
- 遍历所有GROUND TRUTH BOUNDING BOX和预测的BOUNDING BOX，标记出所有的正例和负例
- 画出PRECISION-RECALL曲线，其面积就是这个类别的AP
- 所有类别的AP取平均就是MAP@N

实验结果

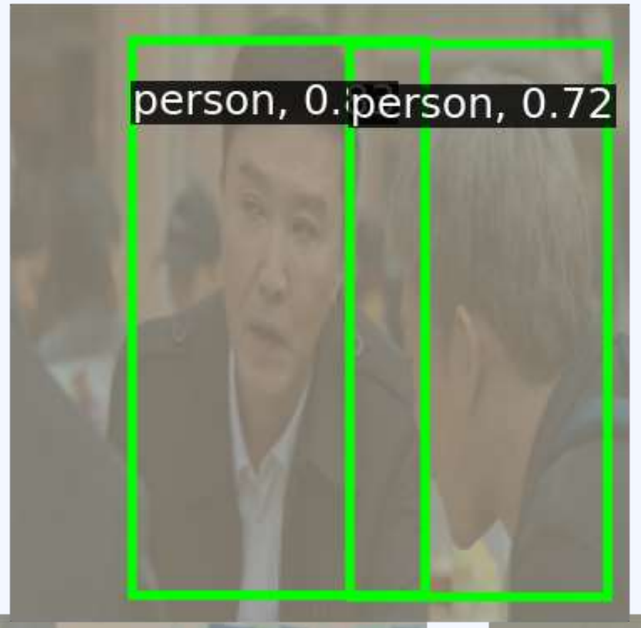

mAP = 41.68%
cat 0.70
train 0.66
aeroplane 0.61
horse 0.59
bus 0.57
dog 0.57
bicycle 0.52
motorbike 0.50
car 0.50
person 0.48
bird 0.43
diningtable 0.36
tvmonitor 0.35
cow 0.35
boat 0.32
sofa 0.31
sheep 0.23
chair 0.14
pottedplant 0.11
bottle 0.05
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
Average Precision


detection-results
(2368 files and 20 detected classes)
person 1453 1208
car 248 306
chair 242 99
dog 127 166
horse 153 128
cat 67 140
bicycle 105 100
bird 69 105
train 59 102
motorbike 68 93
pottedplant 116 44
cow 84 59
tvmonitor 77 62
aeroplane 41 89
boat 72 56
bus 50 68
bottle 93 24
diningtable 58 50
sofa 56 45
sheep 55 43
False Positive
True Positive
0 500 1000 1500 2000 2500 3000
Number of objects per class

实际使用展示


person, 0.50


car, 0.85


person, 0.72


person, 0.71


car, 0.60


person, 0.81


person, 0.79

04

反思与总结

FCOS的优点

- 其他目标检测模型，例如RetinaNet、SSD、YOLOv3、Faster R-CNN都依赖于预先定义的anchor box，相比之下，FCOS不依赖预先定义的anchor box和region proposal。
- 通过去除预先定义的anchor box，FCOS完全的避免了关于anchor box的复杂运算，例如训练过程中计算重叠度，而且节省了训练过程中的内存占用。
- 而且，FCOS避免了和anchor box有关且对最终检测结果非常敏感的所有超参数。
- FCOS模型简单，运算速度快

FCOS的缺点

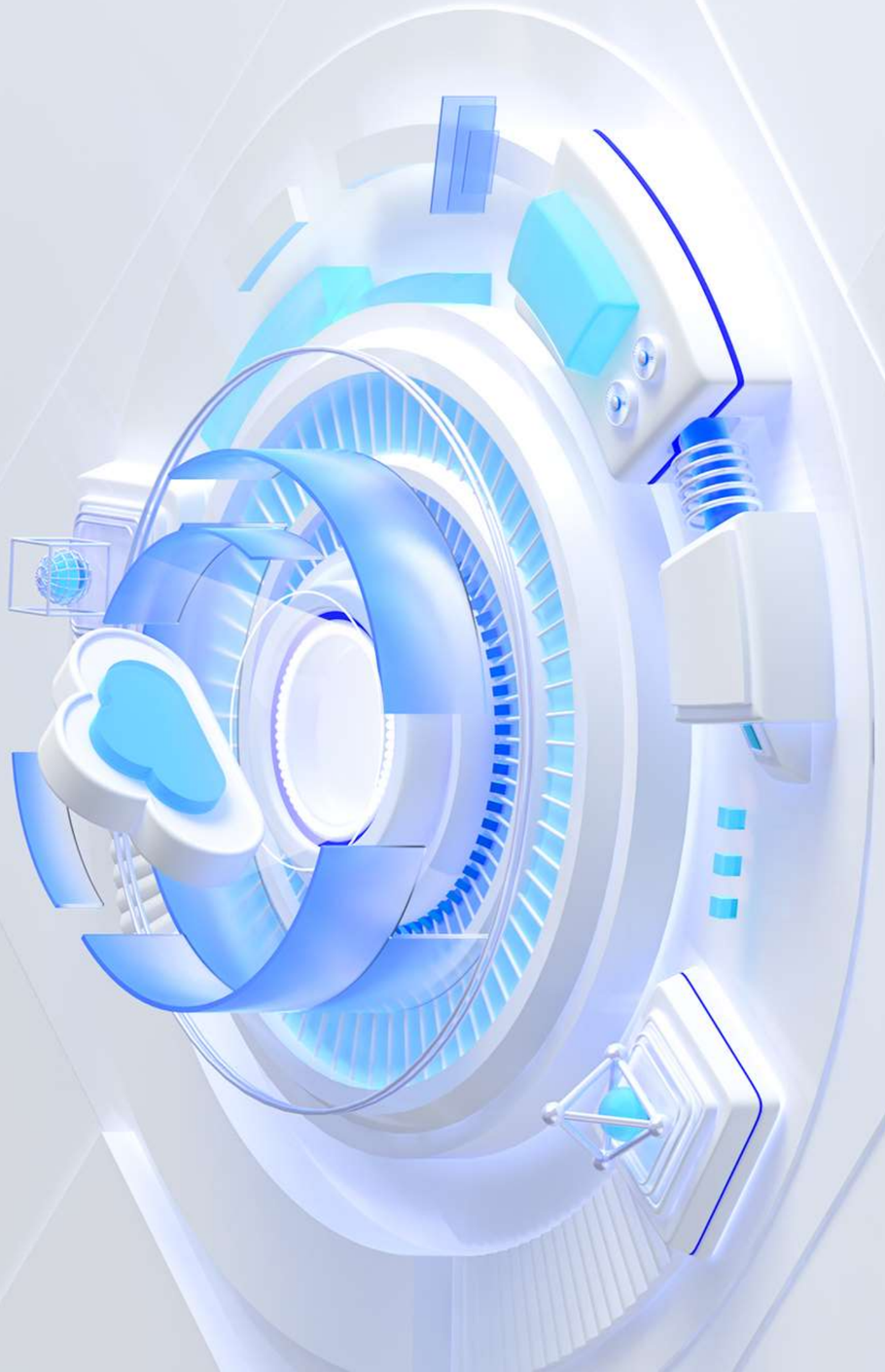
- 相对于二阶段目标检测器，其最终效果可能会有所下降
- 语义模糊性，即如果两个物体的中心点落在了同一个网格中，那么检测效果可能会较差

mAP较低的原因

- 计算资源不足，获取feature map的模型不够大，训练次数不够多
- 超参数的设置不合理

谢 谢 观 看

THANK YOU ALL FOR WATCHING



汇报人：刘卓瀚