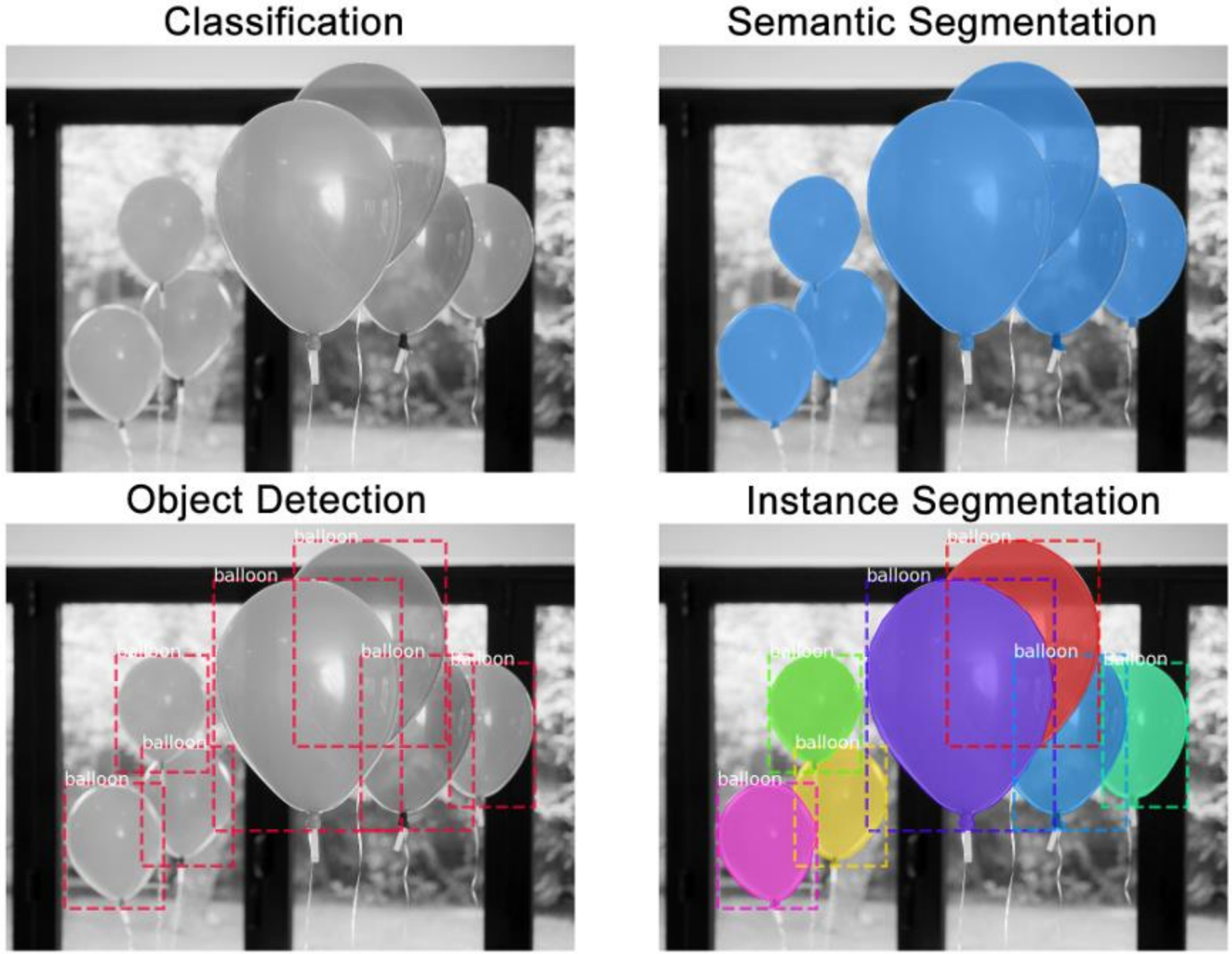
Mask R-CNN

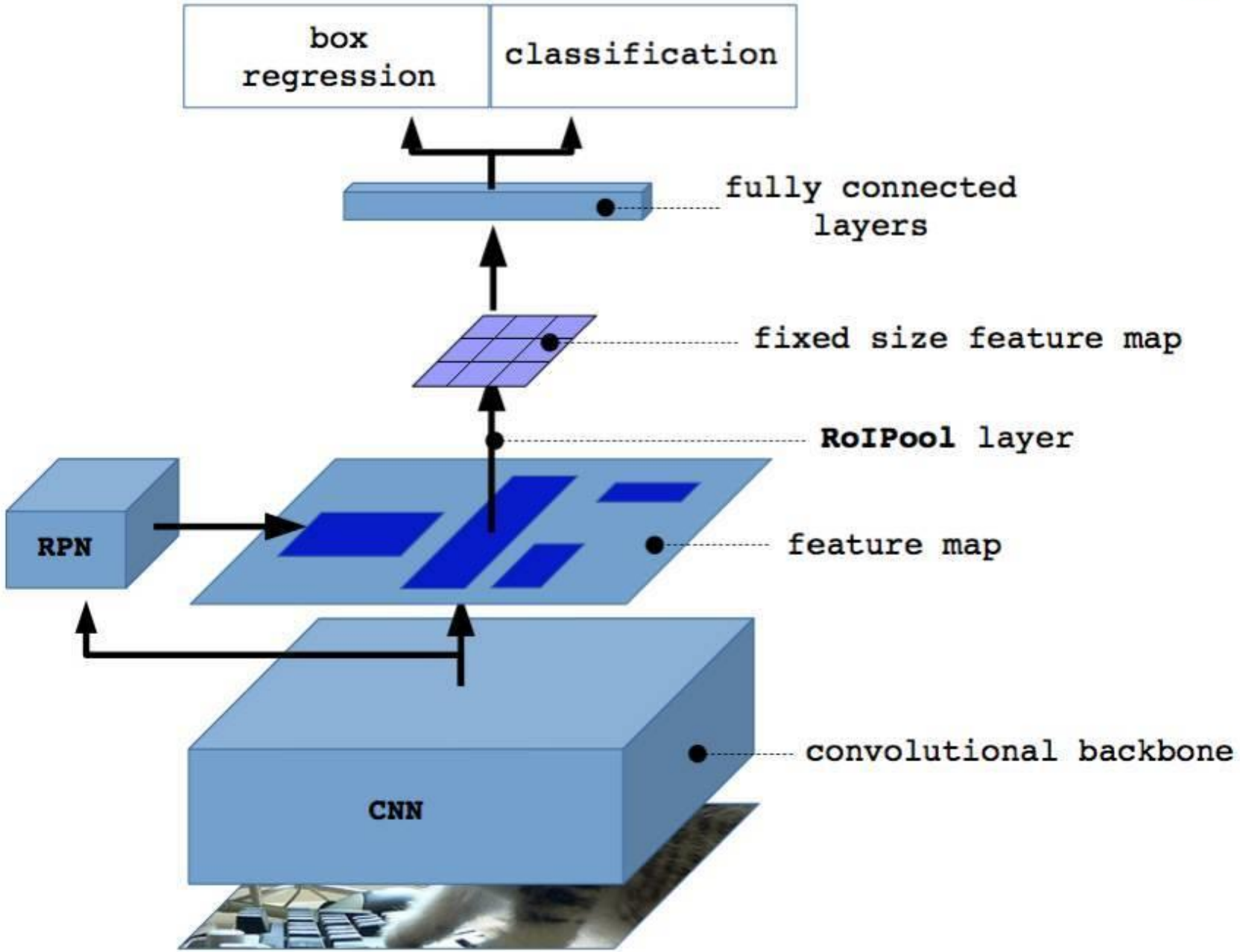
Kaiming He， Georgia Gkioxari， Piotr Dollár， Ross Girshick.

Mask R-CNN. ICCV, 2017

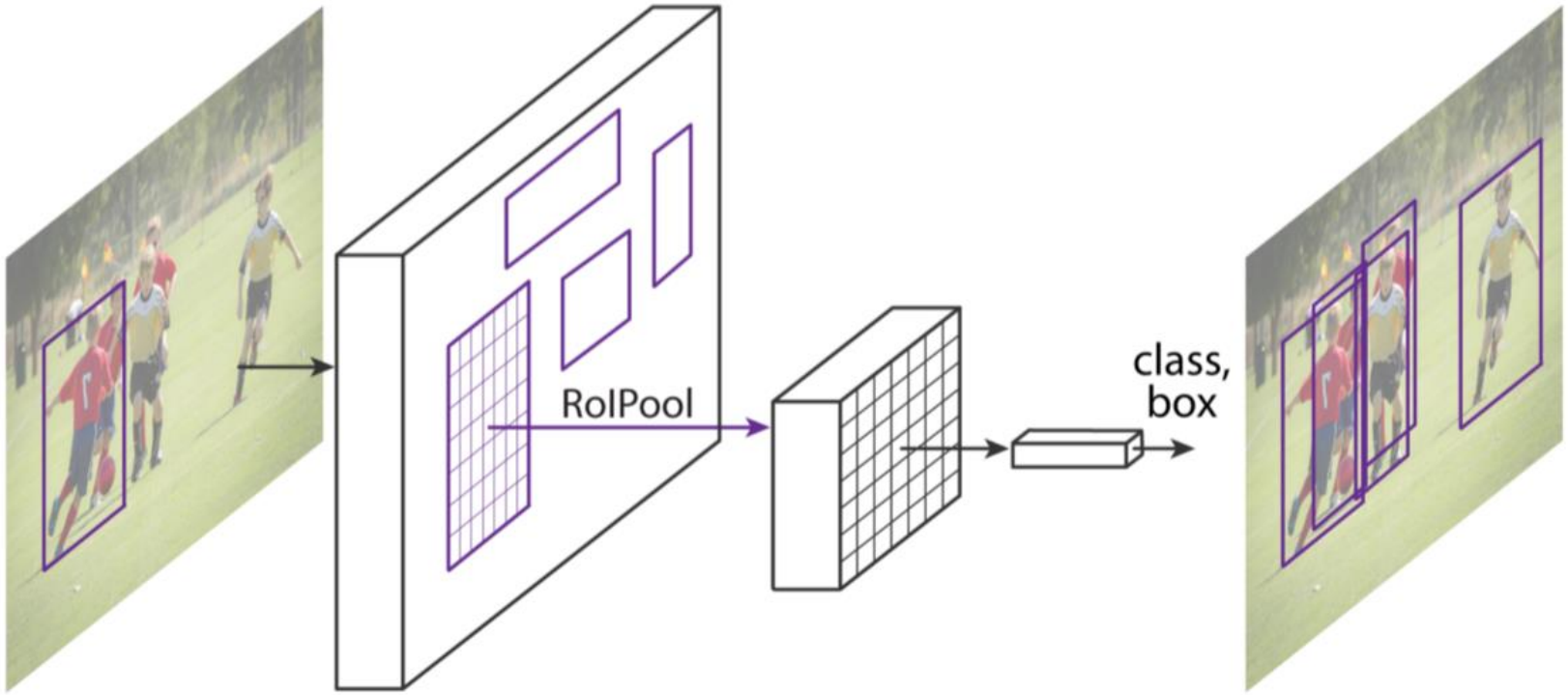
https://arxiv.org/abs/1703.06870


Classification
Semantic Segmentation
Object Detection
balloon
balloon
balloon
balloon
balloon
Instance Segmentation
balloon
balloon
balloon
balloon
balloon

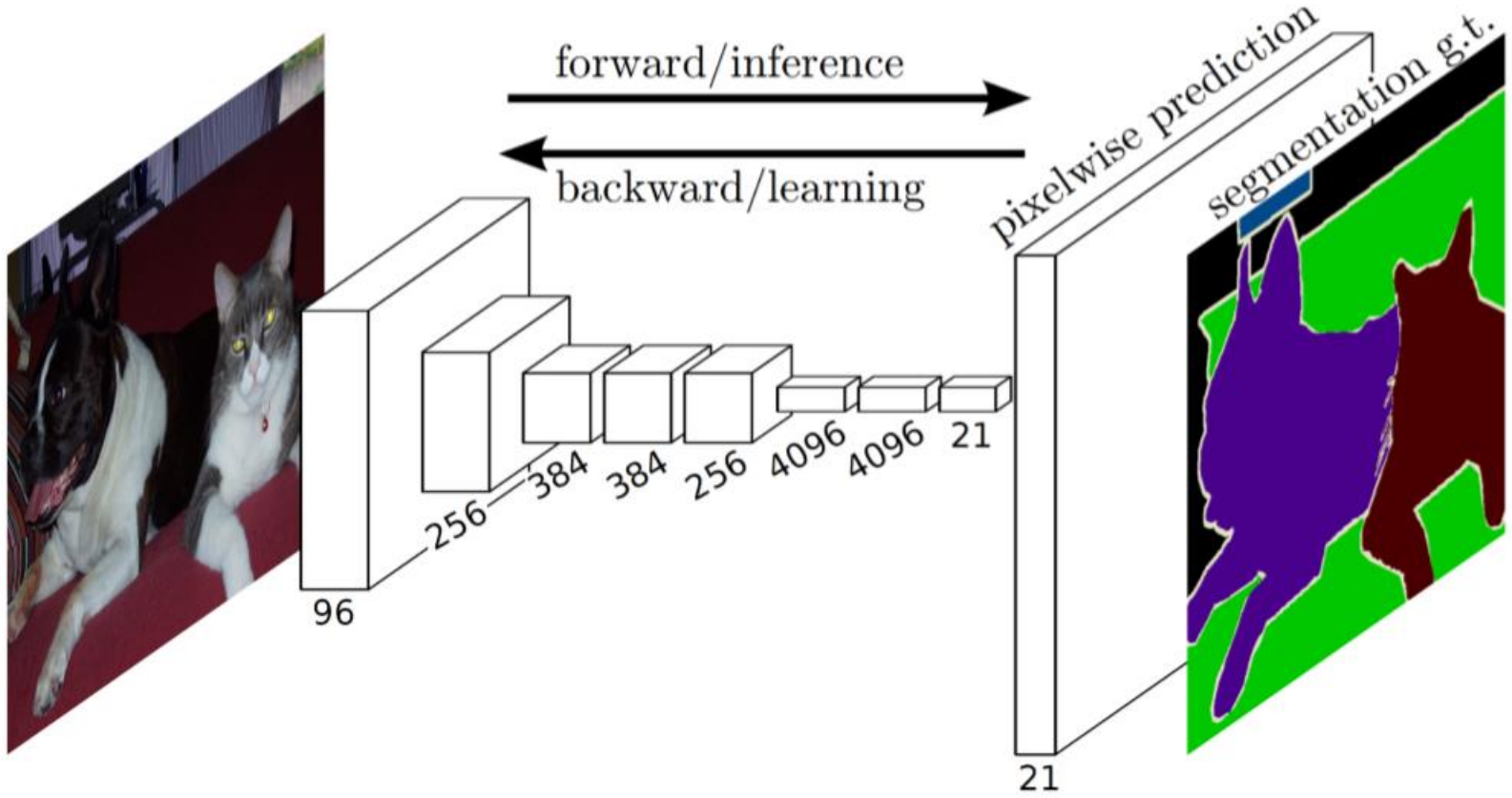
目标检测
Faster R-CNN



目标检测： Faster R-CNN



语义分割：Fully Convolutional Net (FCN)

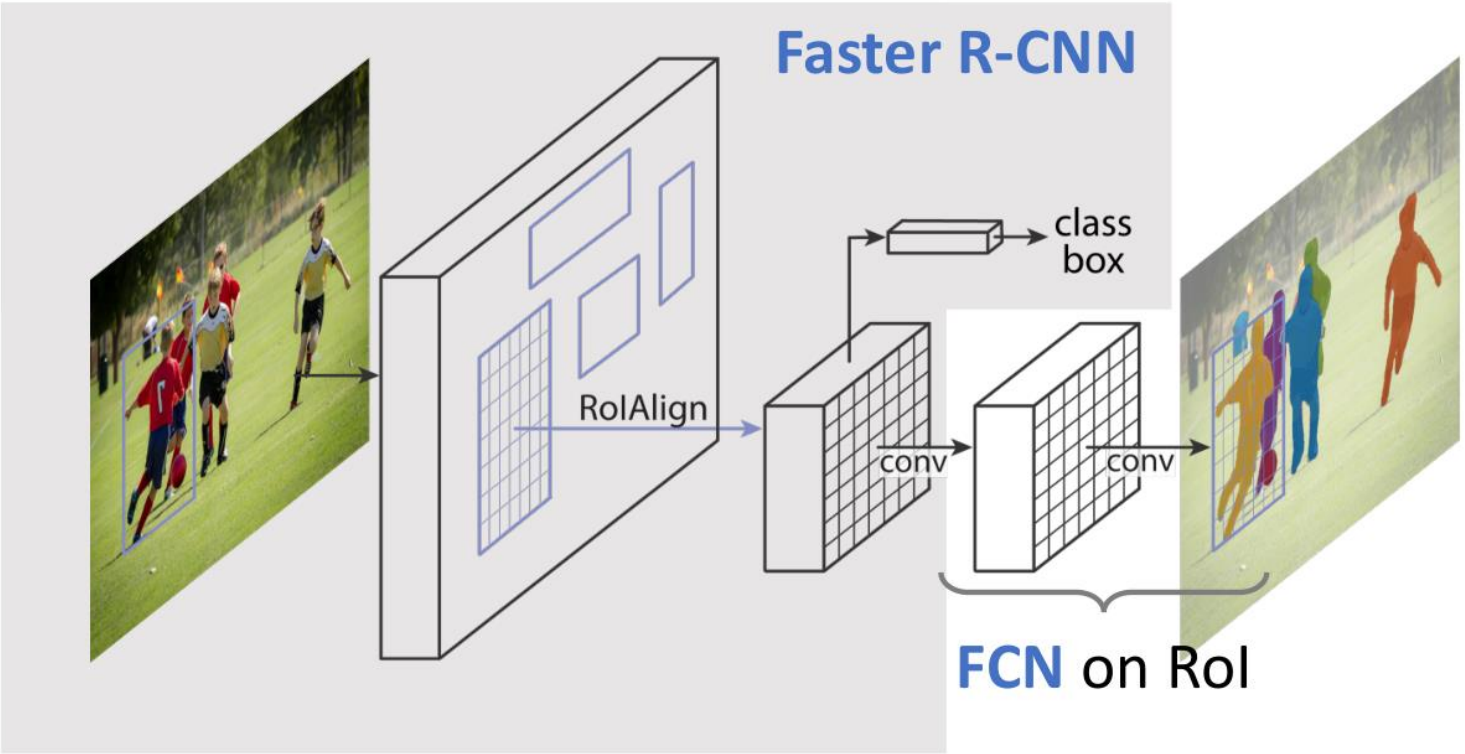


FCNs usually perform per-pixel multi-class categorization, which couples segmentation and classification

Jonathan Long, Evan Shelhamer, & Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015

Mask R-CNN framework for instance segmentation

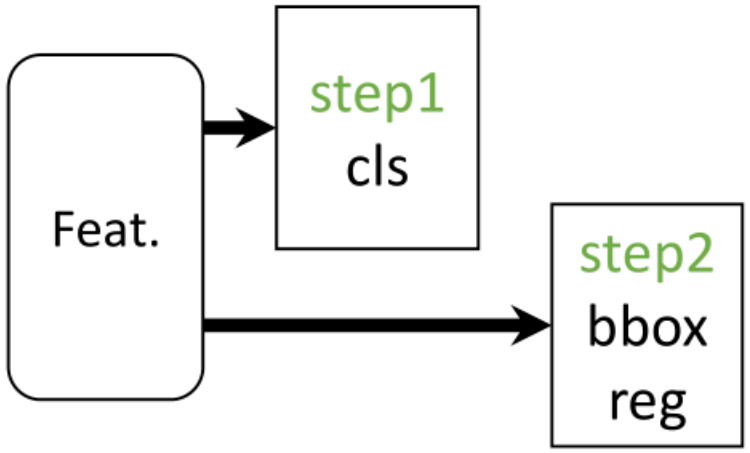
Mask R-CNN = Faster R-CNN with FCN on RoIs



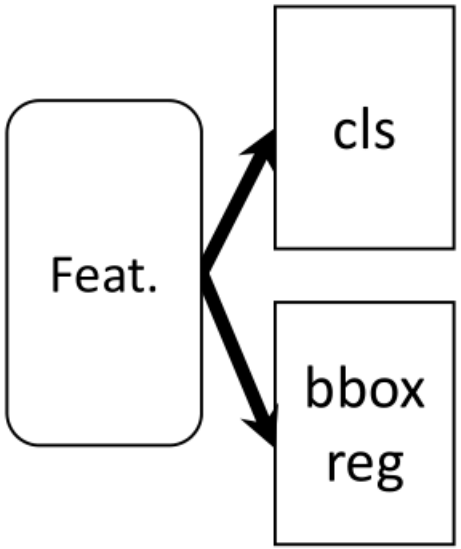
- Mask R-CNN的框架是对faster r-cnn的扩展
- 与bbox识别并行的增加一个mask分支来预测每一个RoI的分割掩码。
- mask分支是应用到每一个RoI上的一个小的FCN，以pix2pix的方式预测分割mask。

What is Mask R-CNN: Parallel Heads

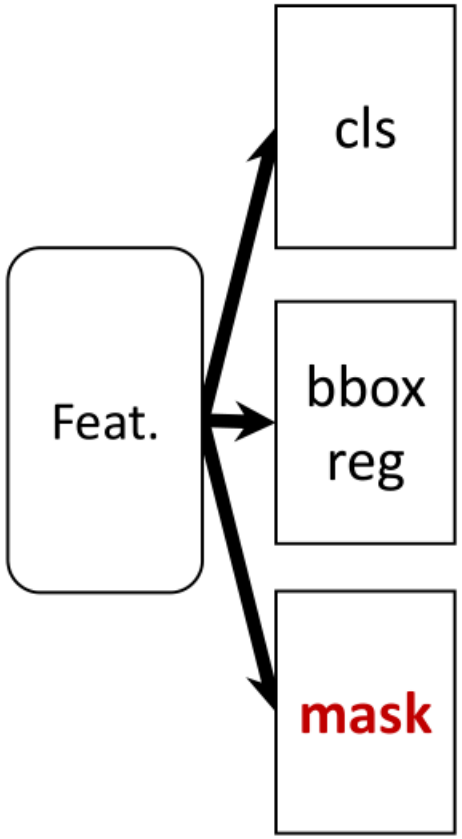
Easy, fast to implement and use



(slow) R-CNN

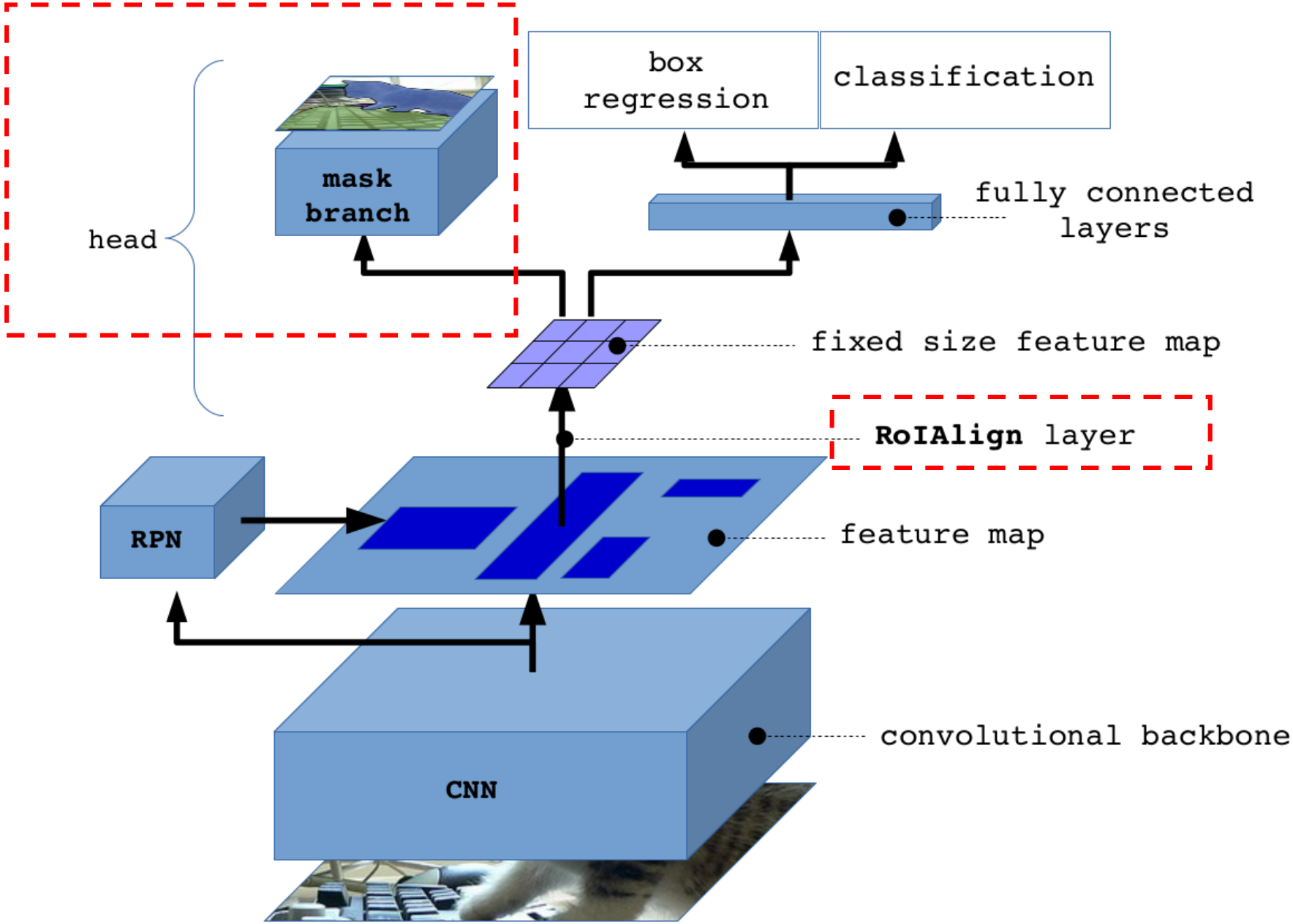


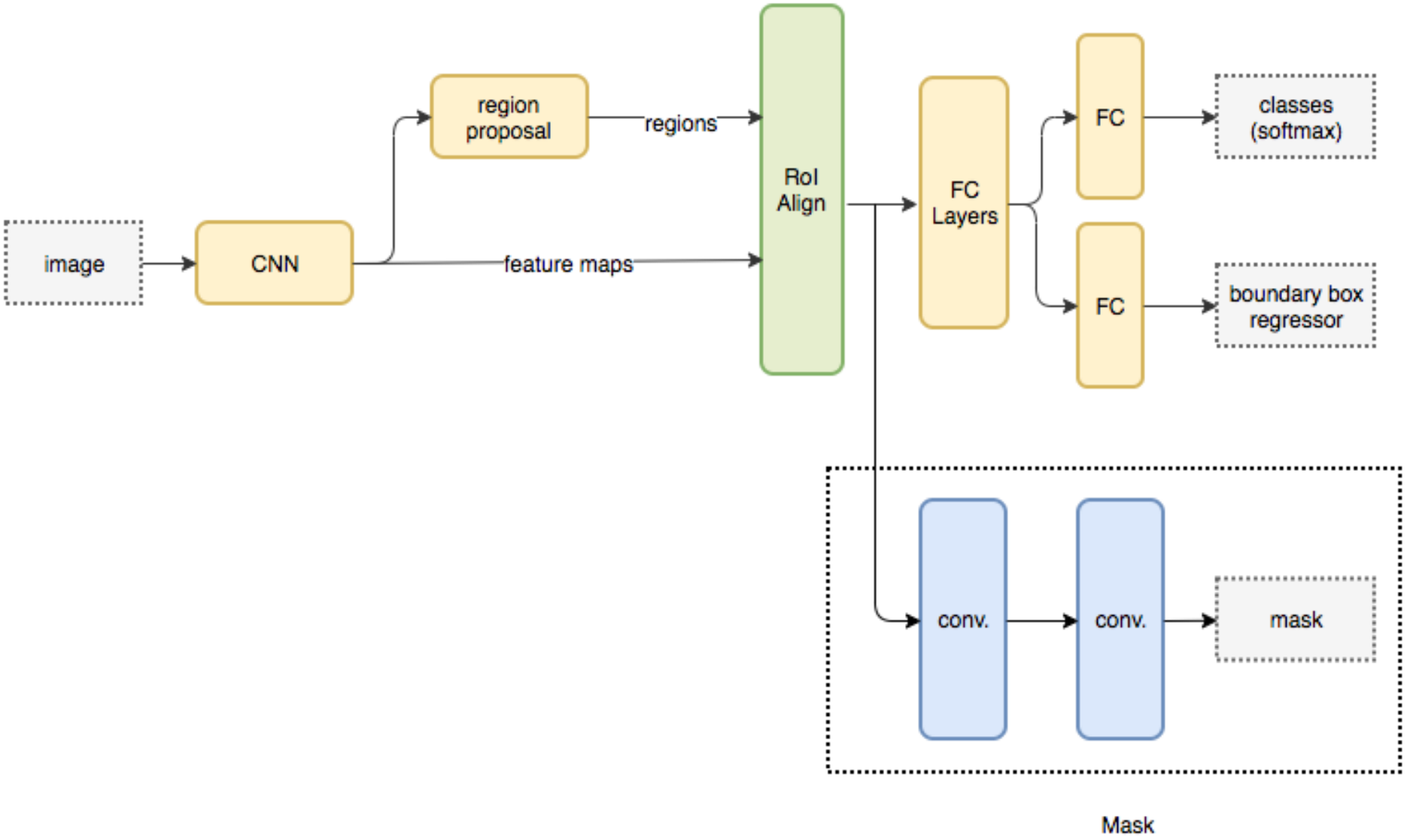
Fast/er R-CNN



Mask R-CNN

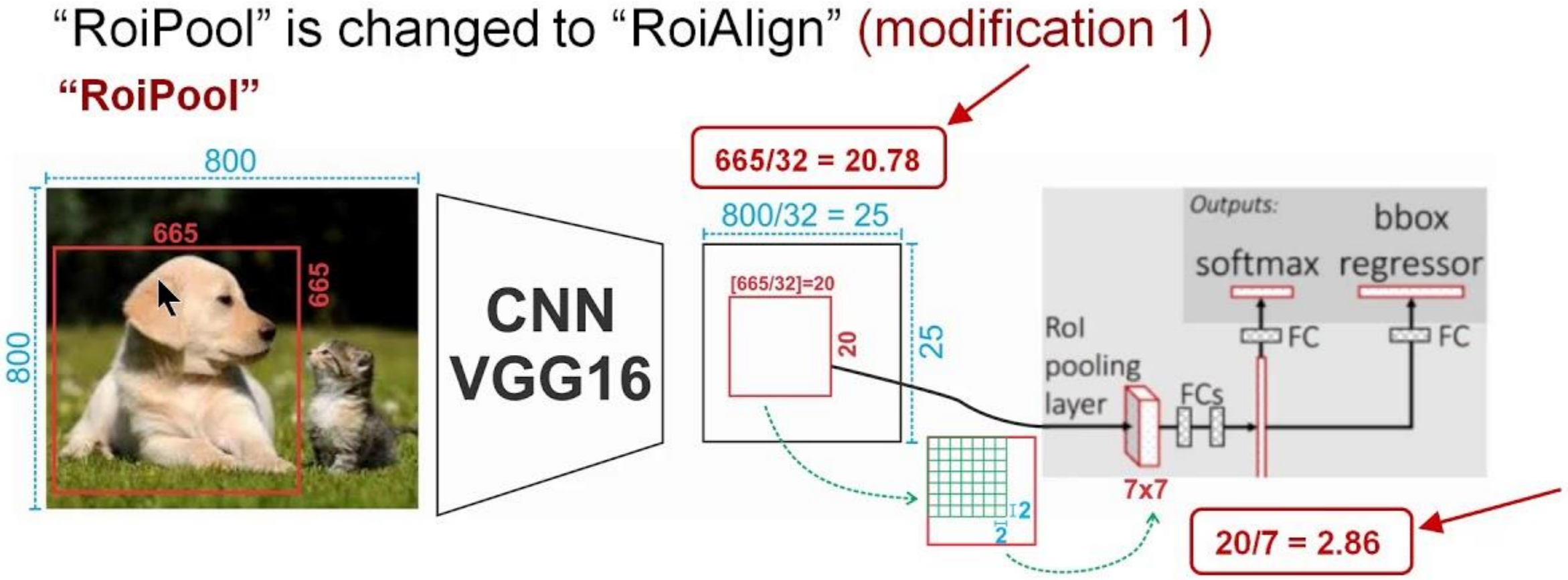
Mask R-CNN


box regression
classification
mask branch
head
fully connected layers
fixed size feature map
RoIAlign layer
RPN
feature map
CNN
convolutional backbone



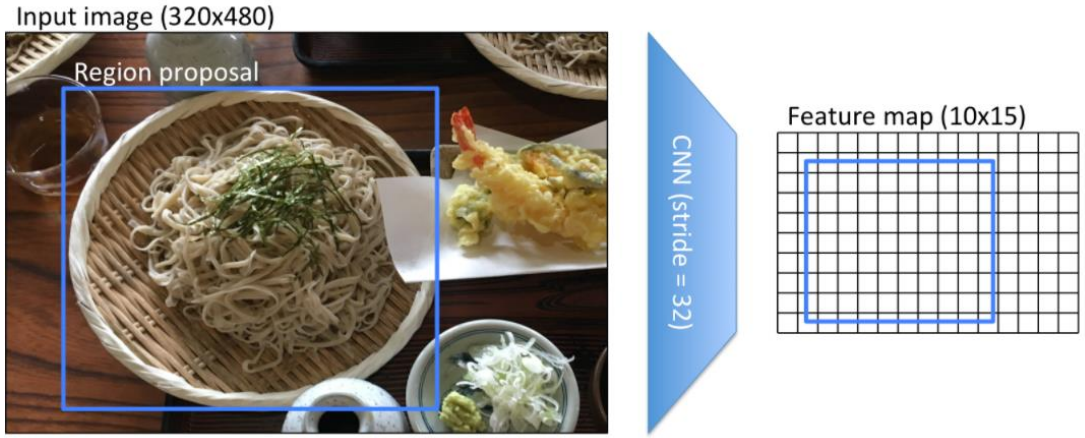
Piggy back 2 convolutional layers to build the mask

RoIPool vs. RoIAlign



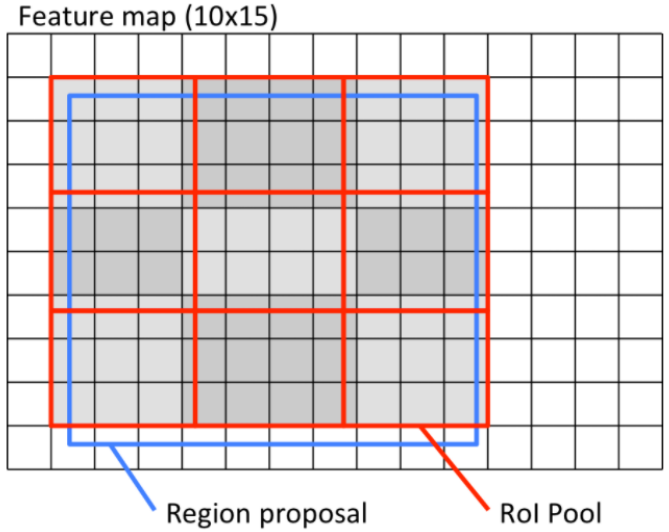
这些量化造成了RoI和提取特征之间的不对准而产生位置偏差

RoIPool vs. RoIAlign

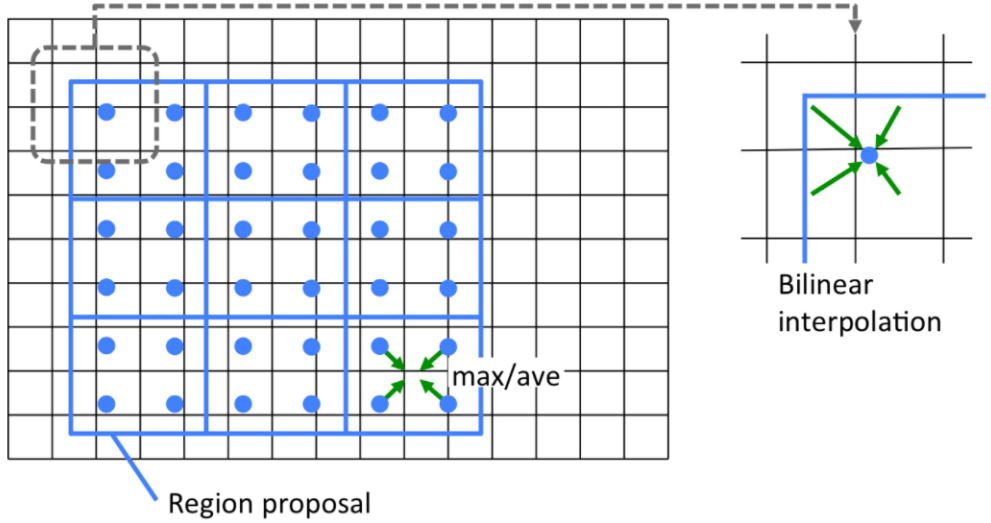


- 避免对RoI边界或bin进行任何量化
- 使用双线性插值计算每个RoI bin中四个采样位置的输入特征的精确值，并使用最大值池化或平均池化。

RoI Pool

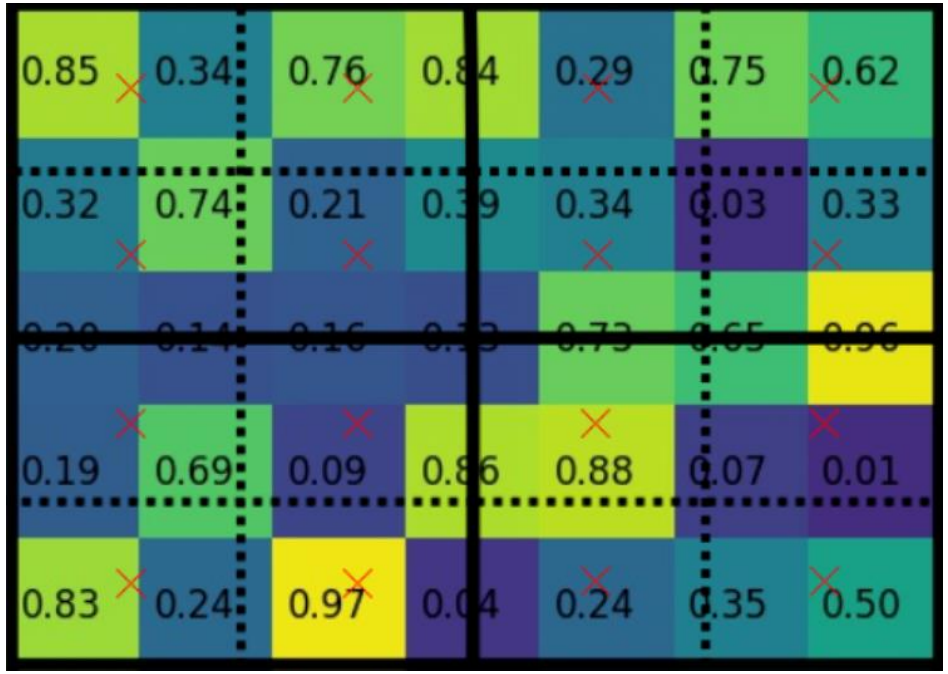


RoI Align



RoI Align

第一个proposal映射阶段不进行量化，在第二个池化阶段也不进行量化，这样最后得到的若干区域的大小为：(665/32/7)x(665/32/7)=2.97x2.97

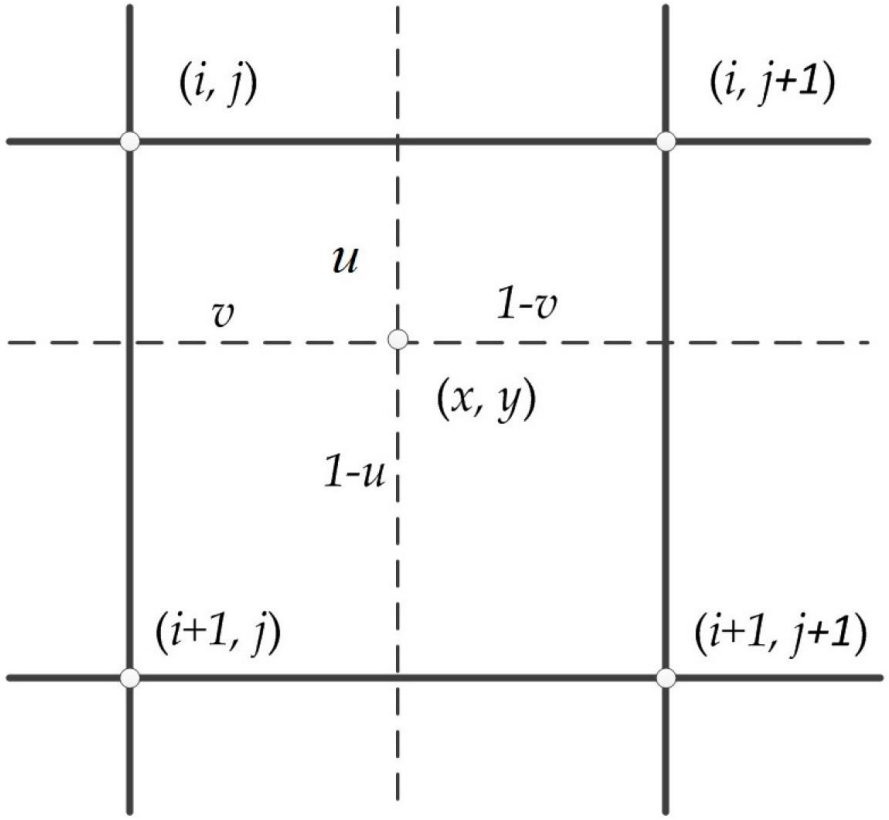


对于每个ROI，映射之后坐标保持浮点数，在此基础上再平均切分成k×k个bin，这个时候也保持浮点数。再把每个bin平均分成 4 个小区域(bin中更小的bin)，然后计算每个更小的bin的中心点的像素点对应的概率值。

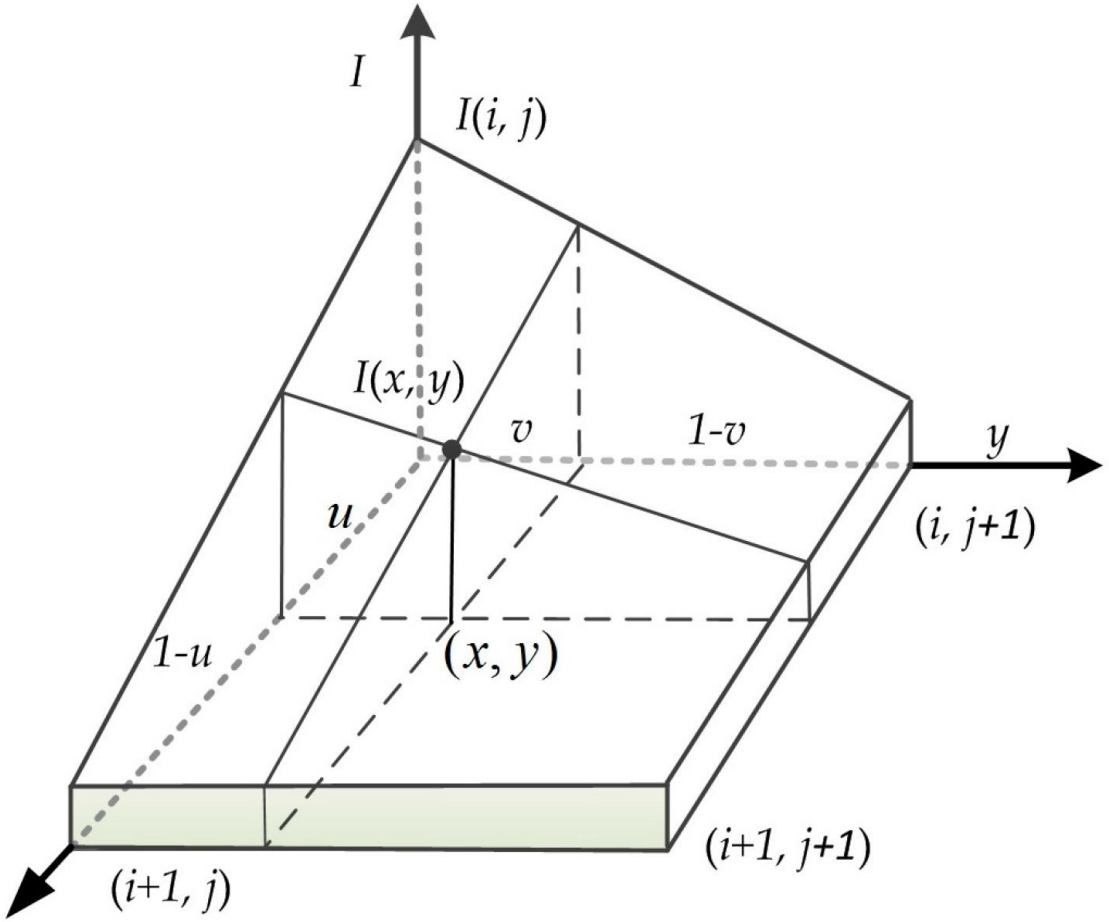
对于每个小区域，取中心点的值作为该区域的像素值，或者每个小区域内取四个点之后取最大值。

每个点的像素值由双线性插值法得到。

双线性插值



(a)

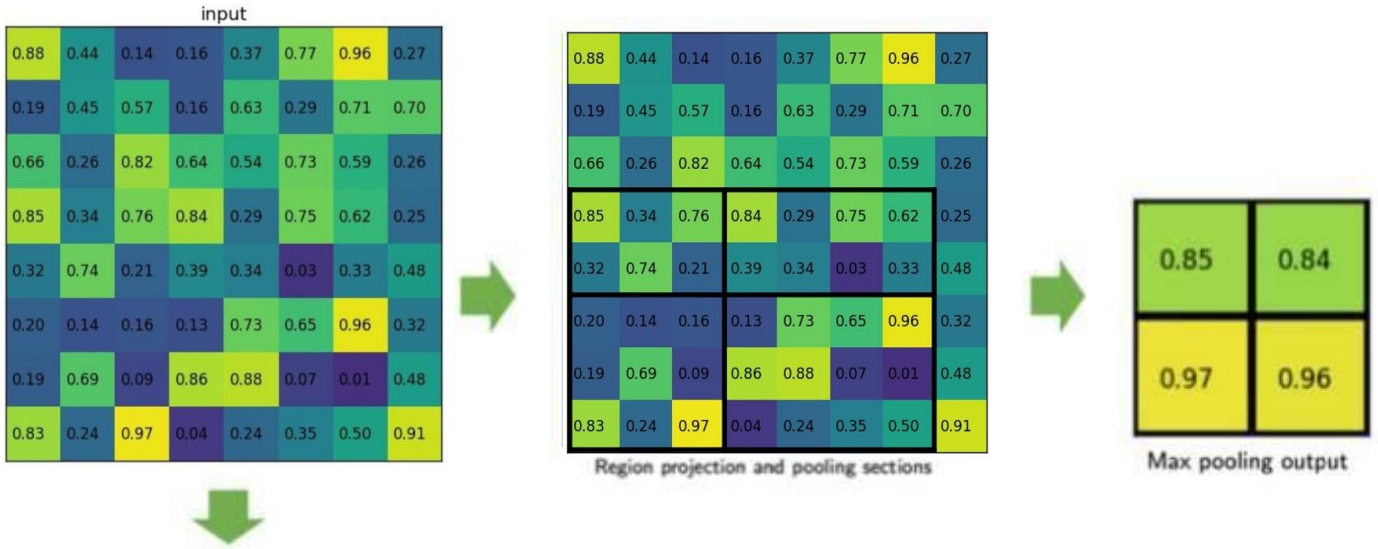


(b)

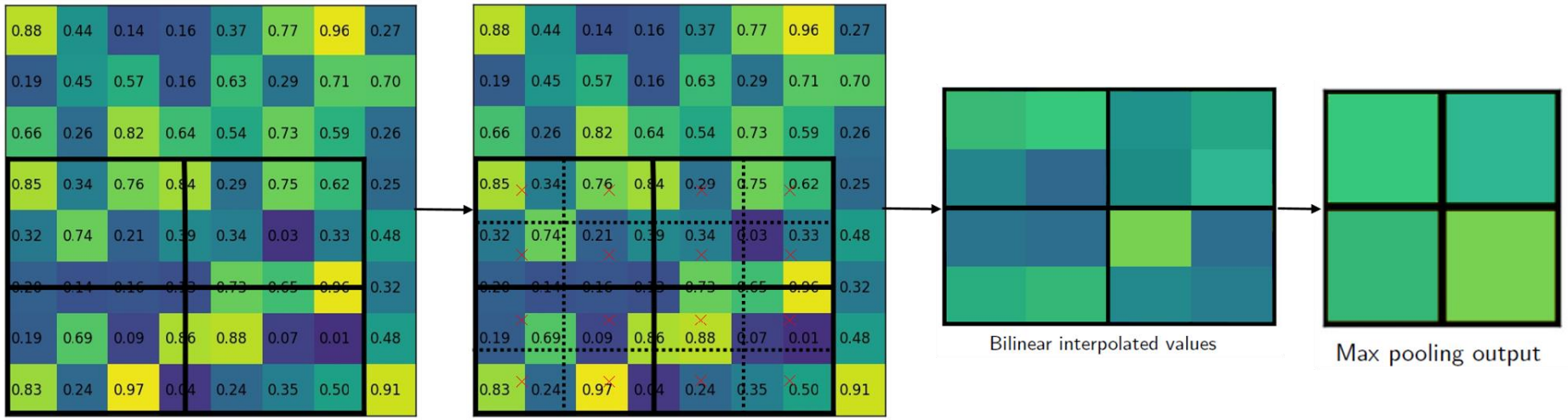
那么， RoI Pooling和RoI Align的区别在于哪里呢？如何能够精确的反向找到对应像素点边缘？

对Pooling的划分不能按照Pooling的边缘， 而是要按照像素点缩放后的边缘。

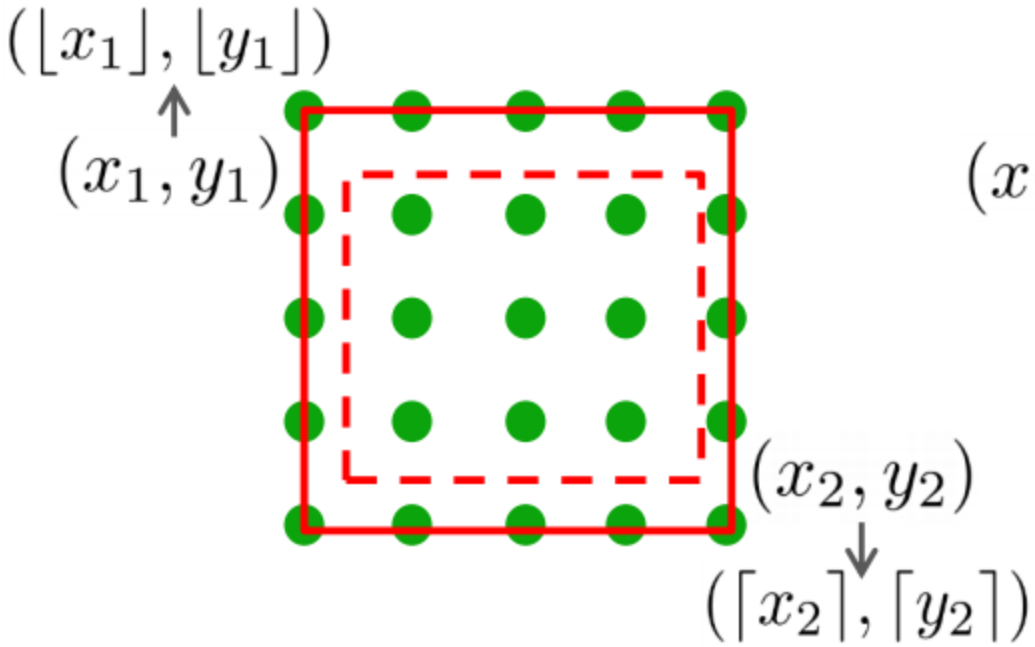
RoI Pooling



RoI Align



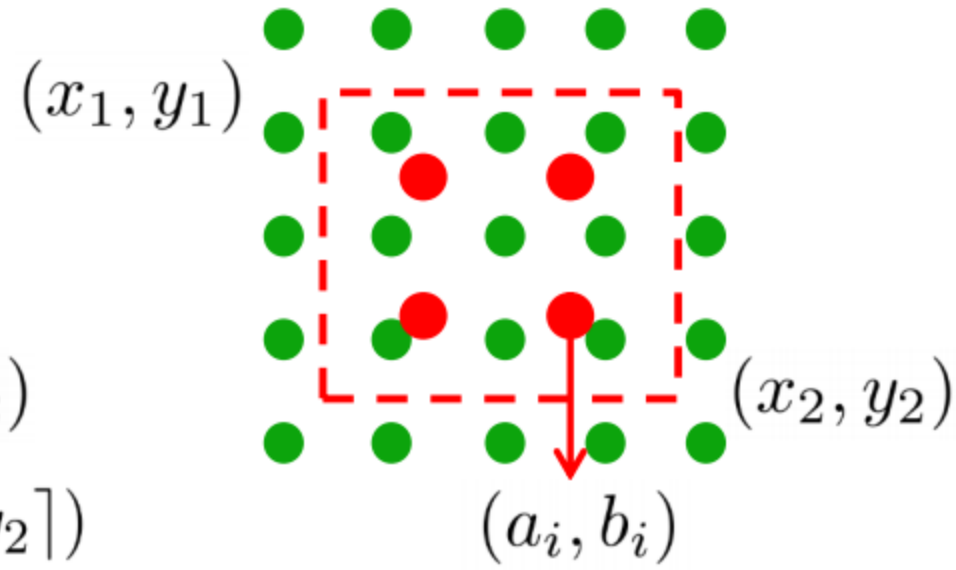
1. RoI Pooling



平均池化：

$$\frac{\sum_{i=\lfloor x_1 \rfloor}^{\lceil x_2 \rceil} \sum_{j=\lfloor y_1 \rfloor}^{\lceil y_2 \rceil} w_{i,j}}{(\lceil x_2 \rceil - \lfloor x_1 \rfloor + 1) \times (\lceil y_2 \rceil - \lfloor y_1 \rfloor + 1)}$$

2. RoI Align



平均池化：

$$\sum_{i=1}^{N} f(a_i, b_i)/N$$

网络架构

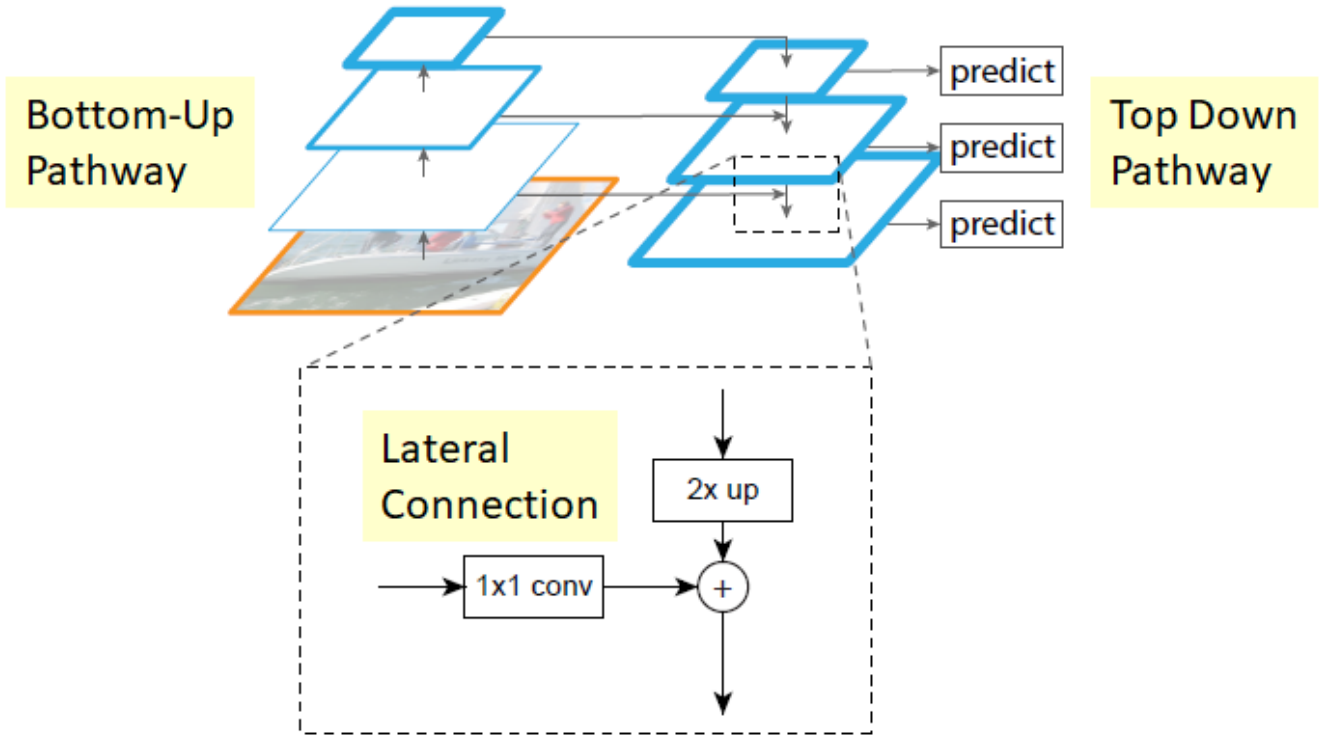
- 为了演示方法的一般性，作者使用多种架构来构建Mask R-CNN。
  区分：（i）用于整个图像上的特征提取的卷积主干架构；
  以及（ii）网络头（head）用于对每个RoI单独应用的边界框识别（分类和回归）和掩码预测。

- 作者首先评估了深度为50或101层的ResNet 和ResNeXt网络。 Faster R-CNN与ResNets 的实现从第4阶段的最终卷积层中提取了特征，称之为C4。 例如，ResNet-50的这个主干由ResNet-50-C4表示。

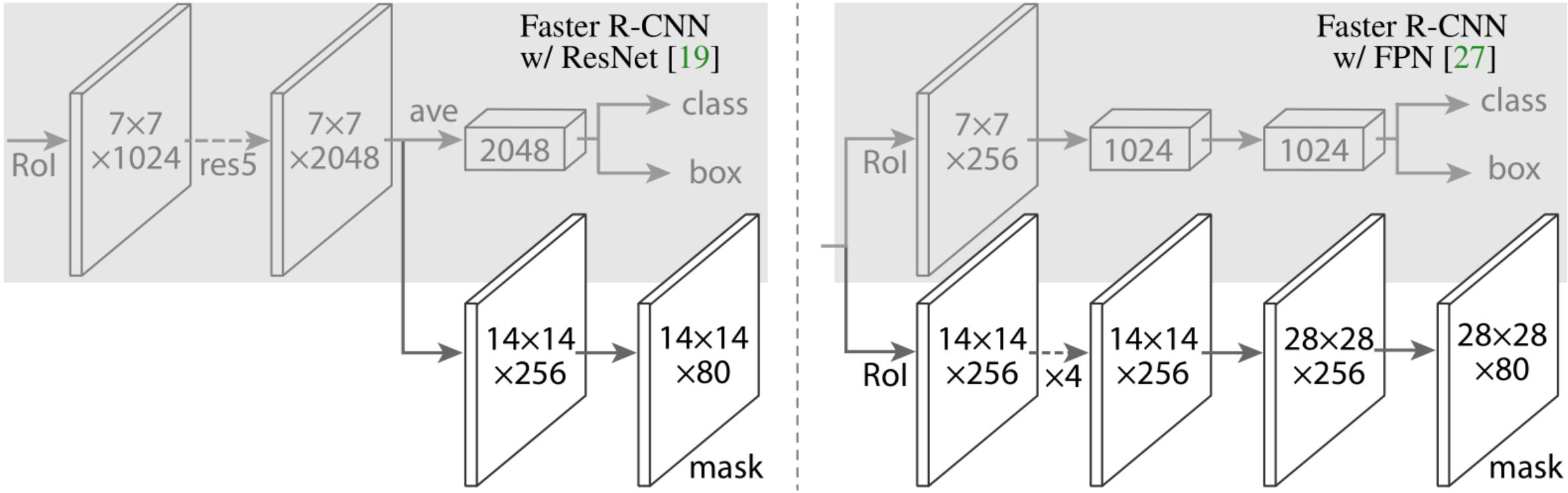
网络架构

- 作者还使用了另一个更有效的称为特征金字塔网络（FPN）主干。
- 具有FPN主干的Faster R-CNN根据其规模从特征金字塔的不同级别提取RoI特征。
- 使用ResNet-FPN主干网通过Mask R-CNN进行特征提取，可以在精度和速度方面获得极佳的提升。

特征金字塔网络（FPN）

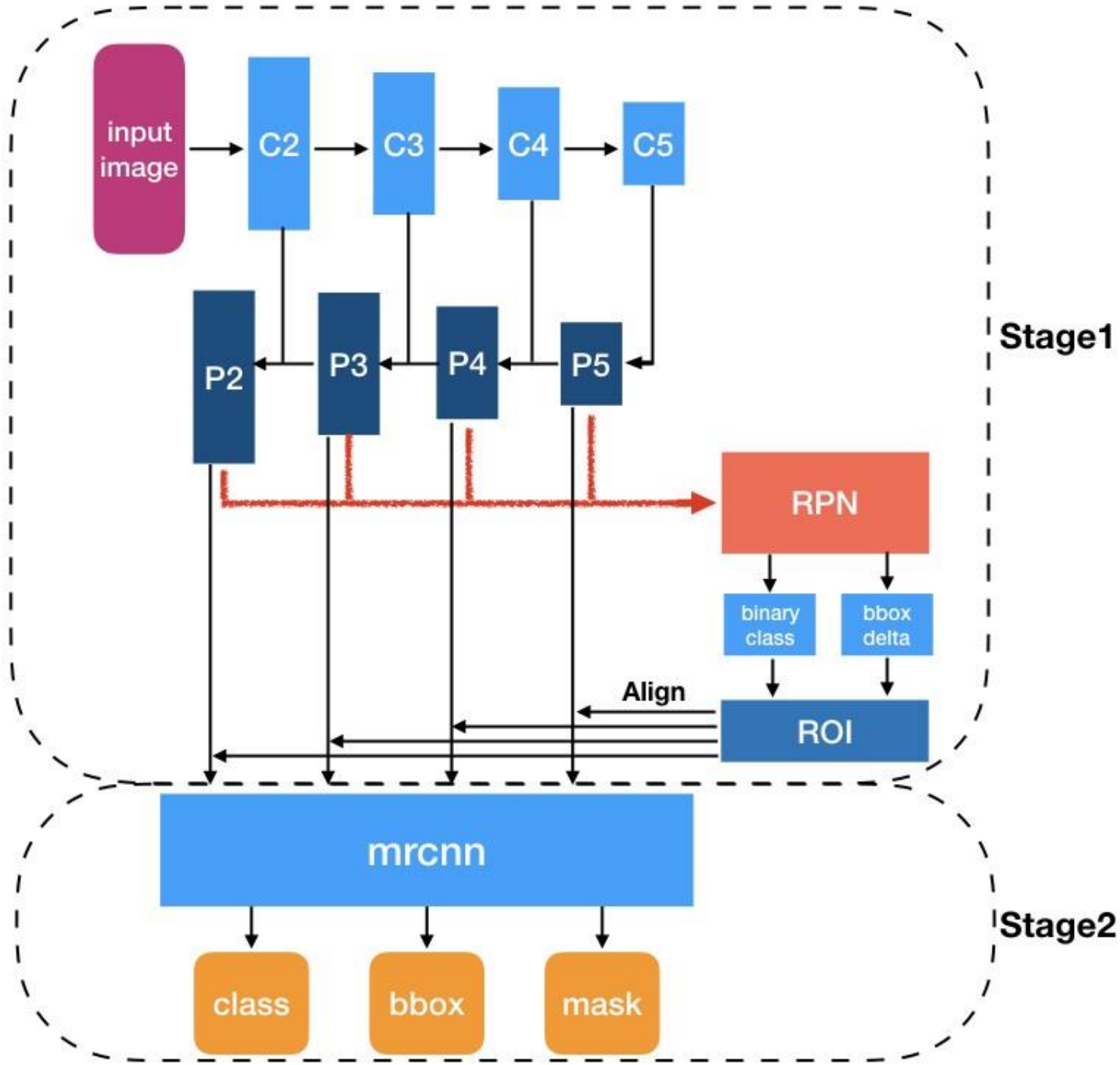


FCN Head Architecture



扩展两个Faster R-CNN heads.

Left/Right panels show the heads for the ResNet C4 and FPN backbones, respectively.


input image
C2
C3
C4
C5
P2
P3
P4
P5
RPN
binary class
bbox delta
Align
ROI
Stage1
mrcnn
class
bbox
mask
Stage2

损失函数

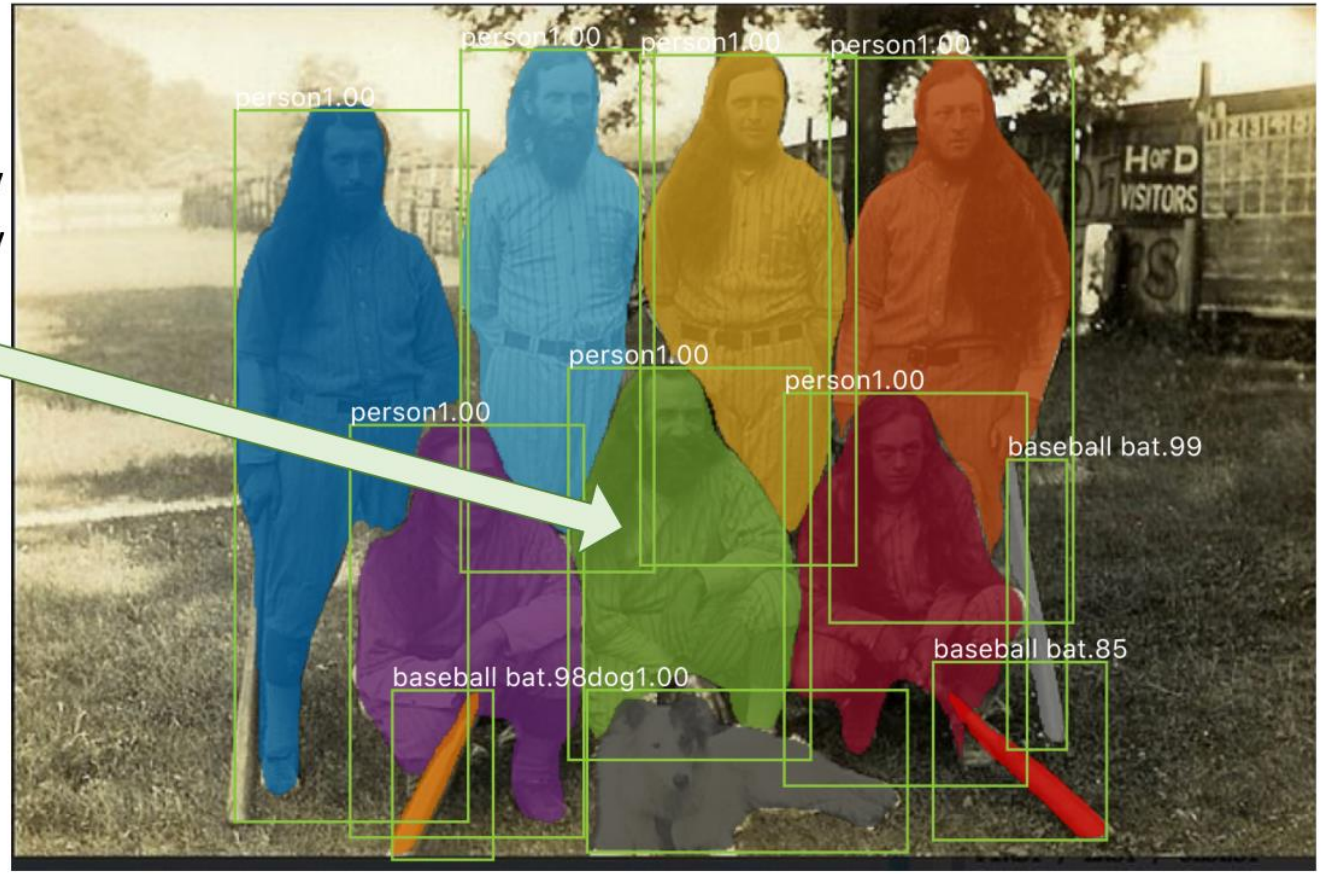
- 在训练期间，将每个采样的RoI上的多任务损失定义为$L = L_{cls} + L_{box} + L_{mask}$。掩码分支对于每个RoI具有$Km^2$维输出，其编码分辨率为$m \times m$的$K$个二进制掩码，每个$K$类对应一个。

- 为此，使用每像素sigmoid，并将L_{mask}定义为平均二元交叉熵损失。 对于与GT类k相关联的RoI， L_{mask}仅在第k个掩码上定义（其他掩码输出不会导致损失）。

- mask预测和class预测去耦合: 对每个类别独立的预测一个二值mask，而不依赖分类分支的预测结果。

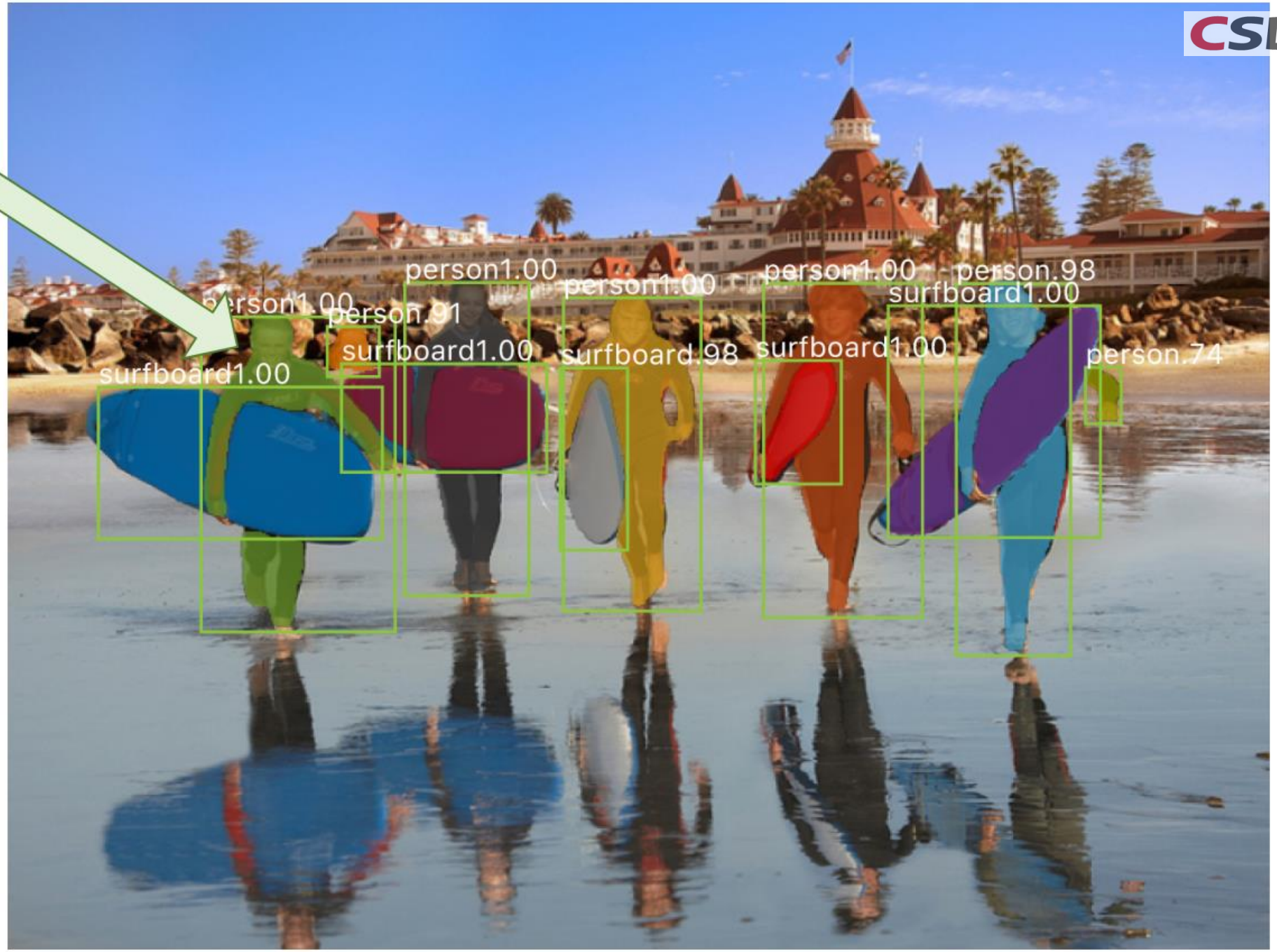
Examples

surrounded by same-category objects

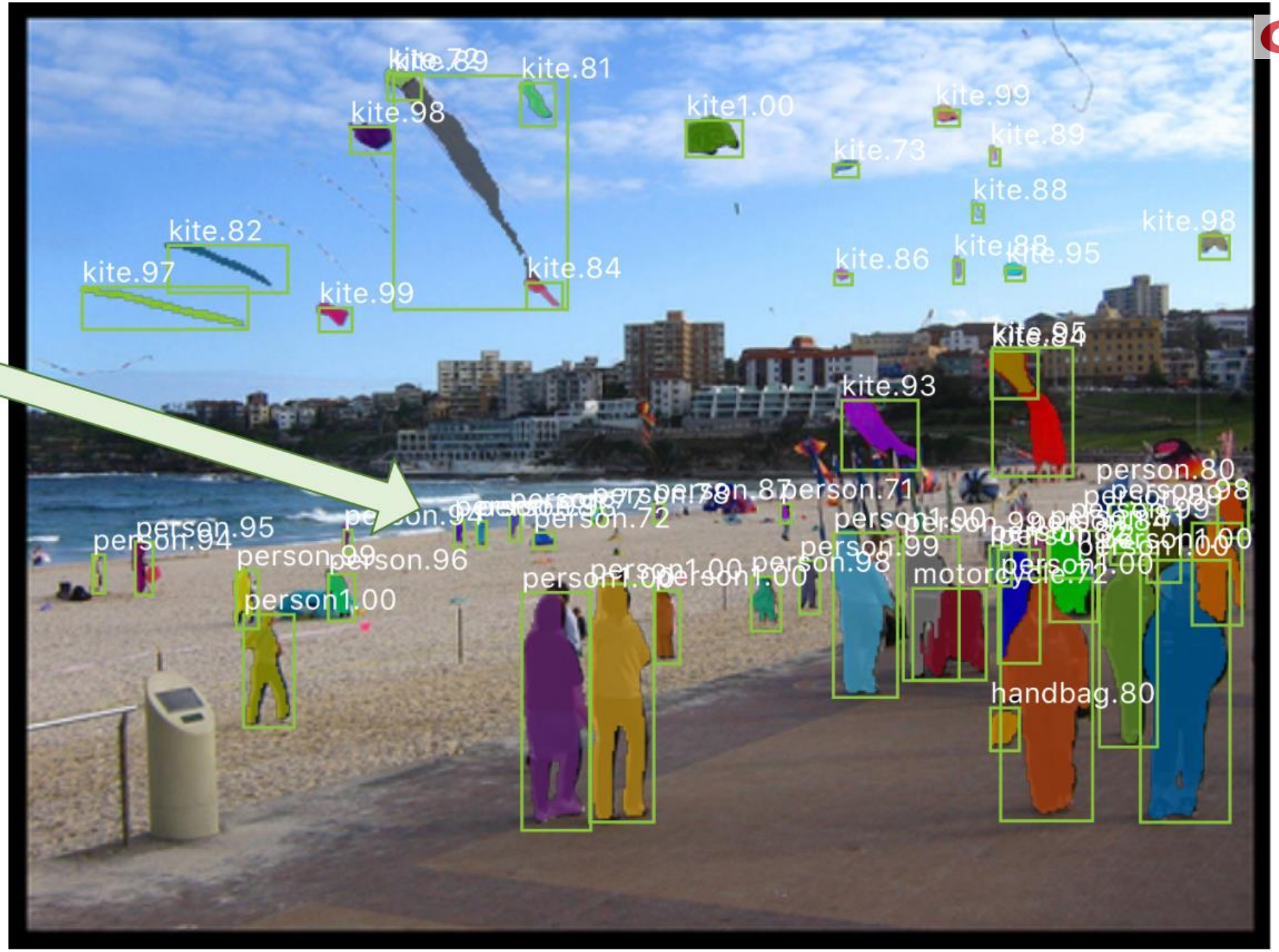


Mask R-CNN on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP

disconnected objects

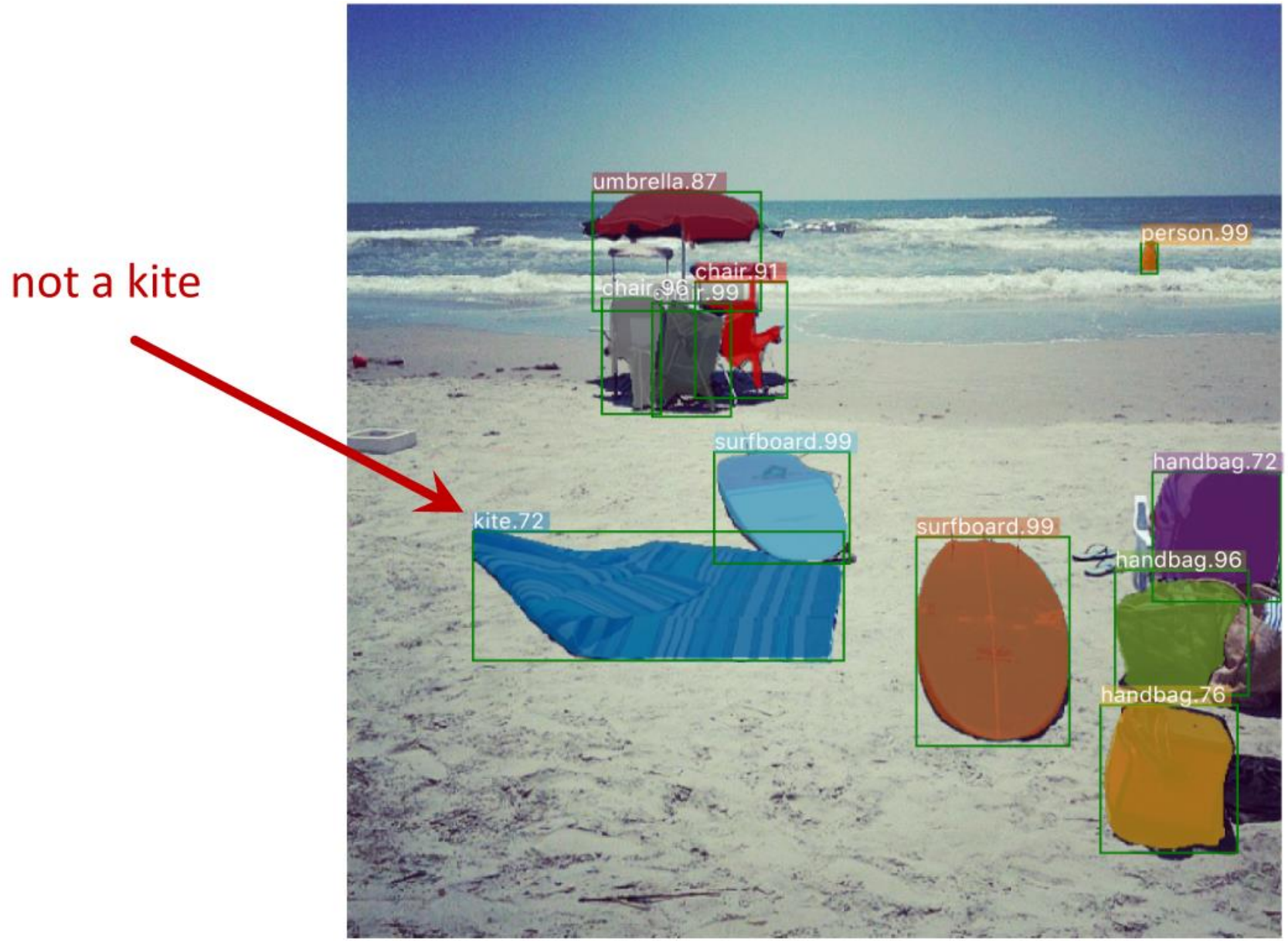


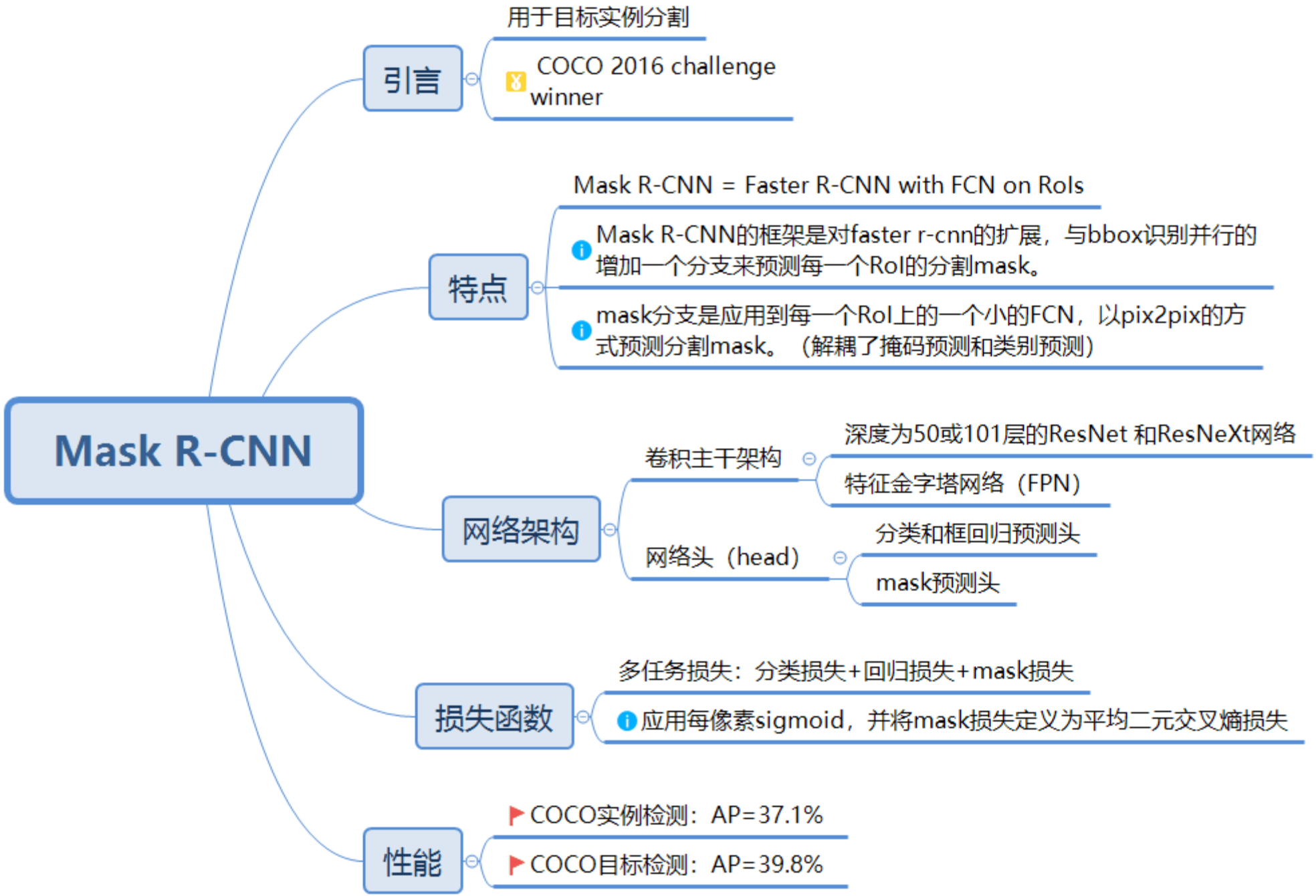
small objects



Failure: detection/segmentation

Failure: recognition




Mask R-CNN
引言
用于目标实例分割
COCO 2016 challenge winner
特点
Mask R-CNN = Faster R-CNN with FCN on RoIs
Mask R-CNN的框架是对faster r-cnn的扩展，与bbox识别并行的增加一个分支来预测每一个RoI的分割mask。
mask分支是应用到每一个RoI上的一个小的FCN，以pix2pix的方式预测分割mask。（解耦了掩码预测和类别预测）
网络架构
卷积主干架构
深度为50或101层的ResNet 和ResNeXt网络
特征金字塔网络（FPN）
网络头（head）
分类和框回归预测头
mask预测头
损失函数
多任务损失：分类损失+回归损失+mask损失
应用每像素sigmoid，并将mask损失定义为平均二元交叉熵损失
性能
COCO实例检测：AP=37.1%
COCO目标检测：AP=39.8%