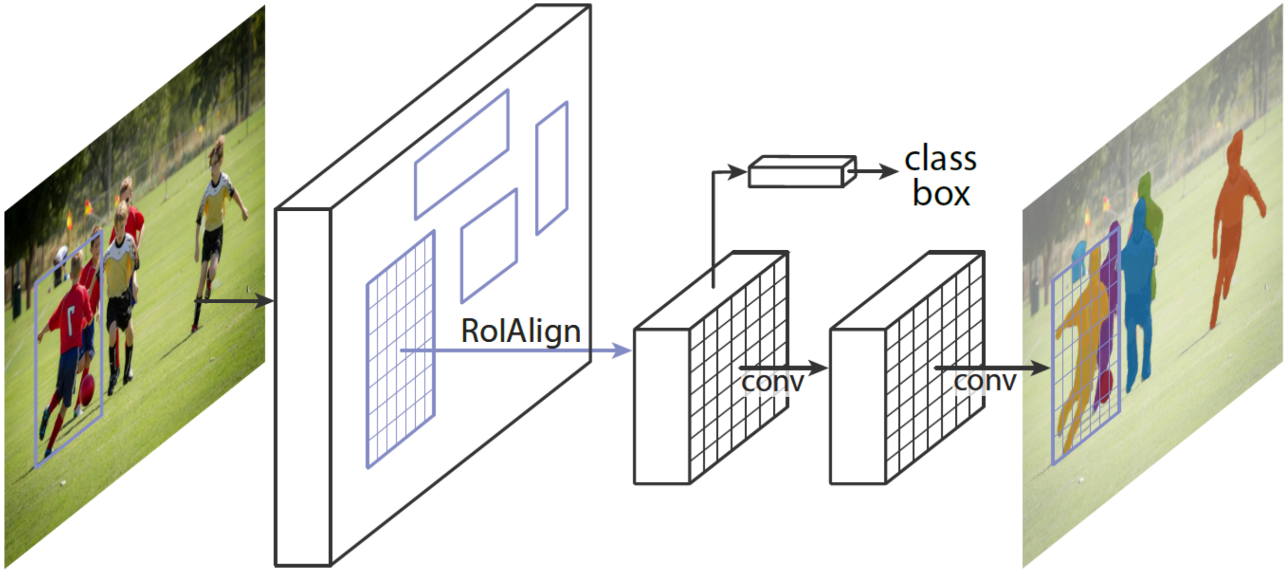


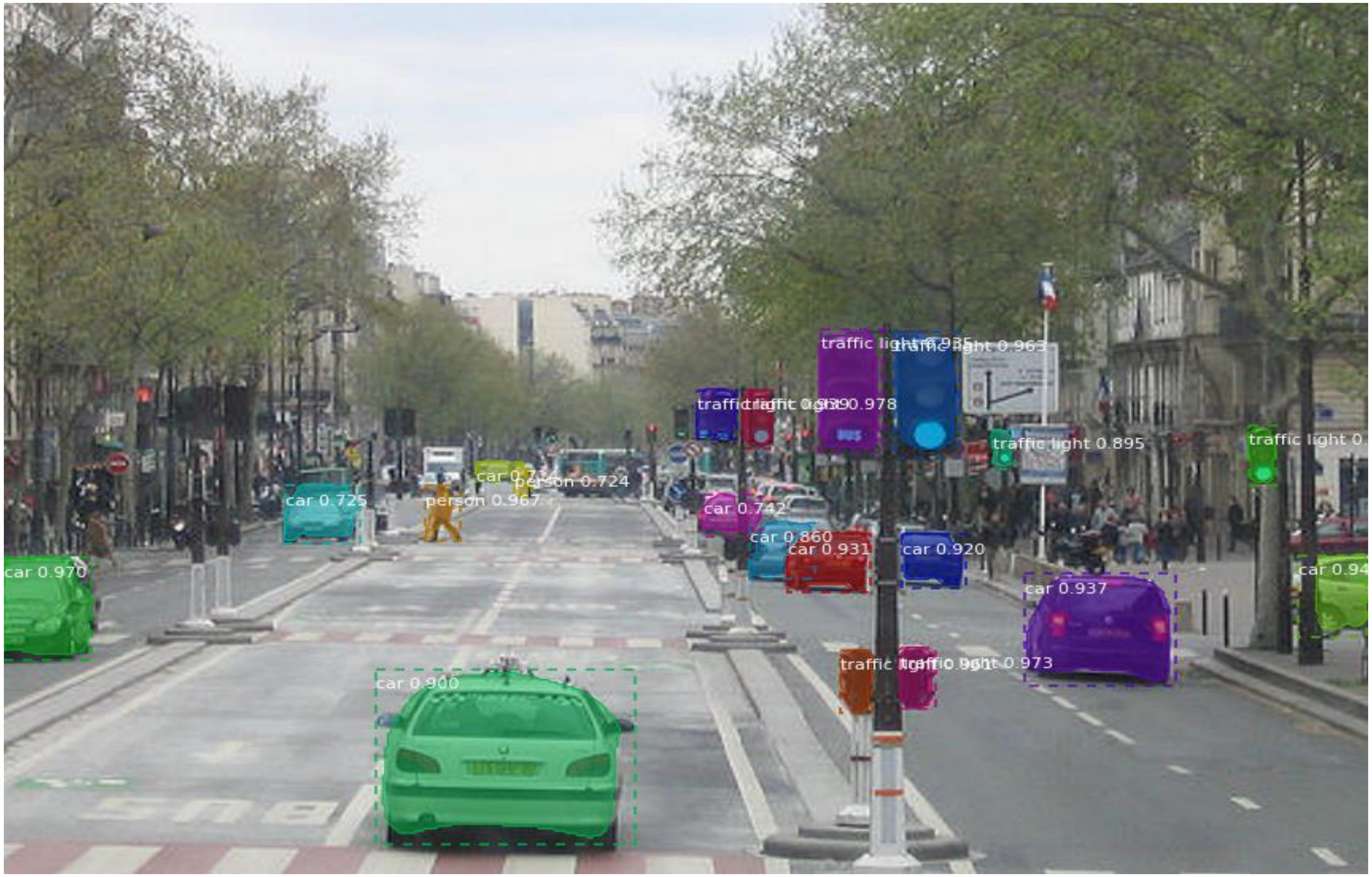
Instance Segmentation：Mask R-CNN

徐培

2019.07.13

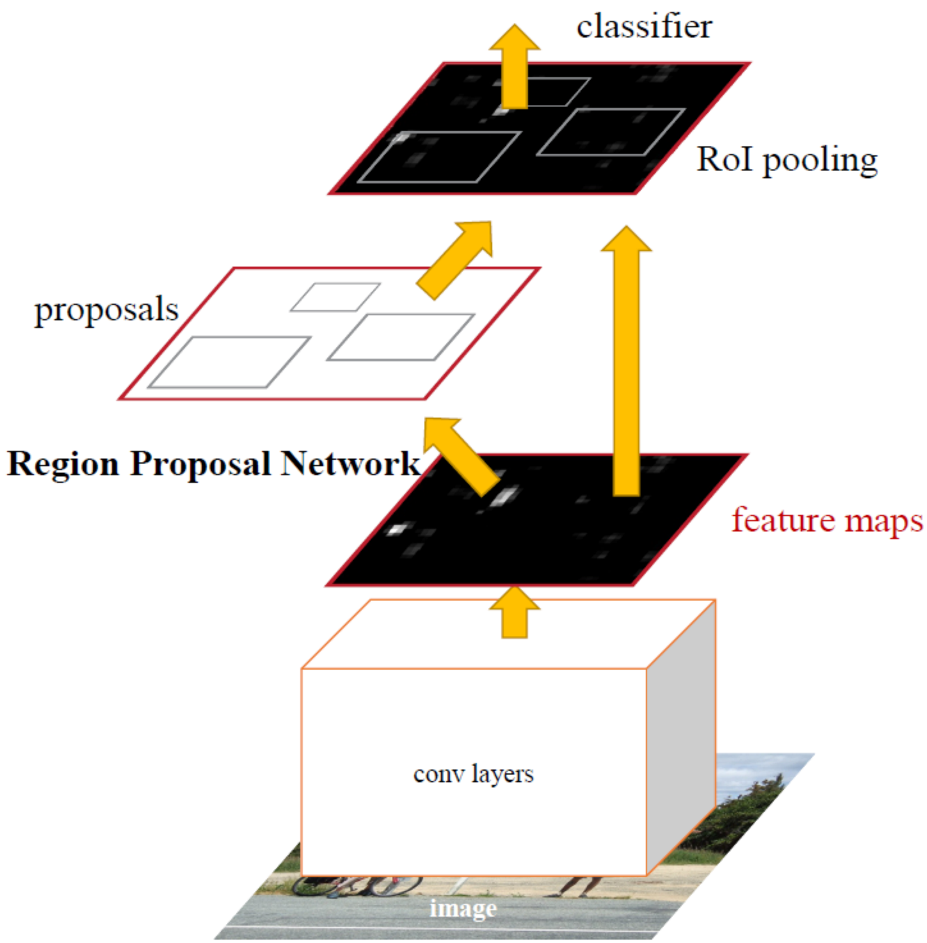
From Faster R-CNN to Mask R-CNN

Mask R-CNN = Faster R-CNN + Mask Branch



Faster R-CNN

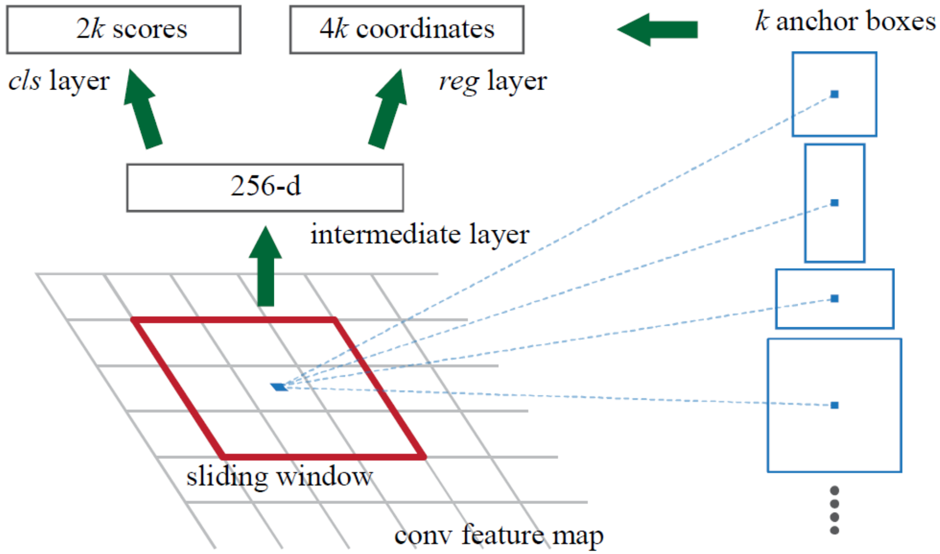
Two-stage object detection network



- Faster R-CNN = Fast R-CNN + RPN
- RPN取代离线的Selective Search 模块
- RPN 和检测网络共享卷积计算
- 基于Attention注意机制引导Fast R-CNN关注区域
- Region proposals 量少质优（~300，高precision，高recall）
- 比SPPnet and Fast R-CNN更快（5fps for VGG16 backbone）.

Faster R-CNN

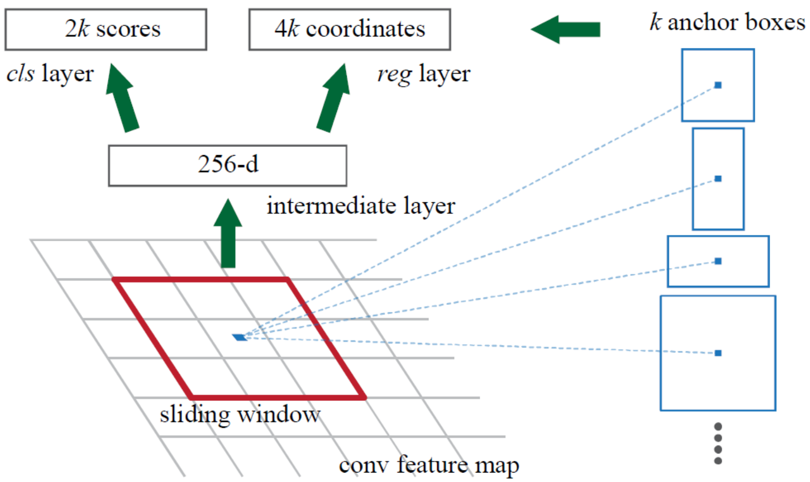
Predict object bounds and objectness scores by RPN



- 每个位置3*3个anchors，3个尺度和3个宽高比
- 对于W x H 的卷积特征图，共产生WxHxk个anchors

Faster R-CNN

Predict object bounds and objectness scores by RPN



- 3 x 3，256-d （256-d for ZF and 512-d for VGG）卷积层 + ReLU ← 输入图片Conv5特征.
- 1x1, 4k-d卷积层 → 输出k组proposal的offsets（t_x, t_y, t_w, t_h）
- 1x1, 2k-d卷积层 → 输出k组（object score, non-object score）

Faster R-CNN

RPN loss function

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*).$$

- i：每个mini-batch中anchor的索引；
- p_i 第i个anchor 预测为物体的概率，p_i*为 ground-truth label(1 for positive)；
- t_i 是预测的bounding box 参数, t_i* 是正样本anchor的ground-truth box 参数;
- L_{cls} is log loss over two classes(object vs. not object);
- L_{reg} is the robust loss function (smooth L1) defined in fast R-CNN;
- The two terms are normalized by N_{cls} (mini-batch size)and N_{reg} (the number of anchor locations)and weighted by a balancing parameter λ.

Faster R-CNN

Bounding box regression

$$
\begin{aligned}
t_{\mathrm{x}} &= (x - x_{\mathrm{a}})/w_{\mathrm{a}}, \quad t_{\mathrm{y}} = (y - y_{\mathrm{a}})/h_{\mathrm{a}}, \\
t_{\mathrm{w}} &= \log(w/w_{\mathrm{a}}), \quad t_{\mathrm{h}} = \log(h/h_{\mathrm{a}}), \\
t^{*}_{\mathrm{x}} &= (x^{*} - x_{\mathrm{a}})/w_{\mathrm{a}}, \quad t^{*}_{\mathrm{y}} = (y^{*} - y_{\mathrm{a}})/h_{\mathrm{a}}, \\
t^{*}_{\mathrm{w}} &= \log(w^{*}/w_{\mathrm{a}}), \quad t^{*}_{\mathrm{h}} = \log(h^{*}/h_{\mathrm{a}}),
\end{aligned}
$$

- where x, y, w, and h denote the box's center coordinates and its width and height.
- Variables x, x_a, and x* are for the predicted box, anchor box, and ground-truth box respectively (likewise for y,w, h);
- thought of as bounding-box regression from an anchor box to a nearby ground-truth box.

Faster R-CNN

Bounding box regression

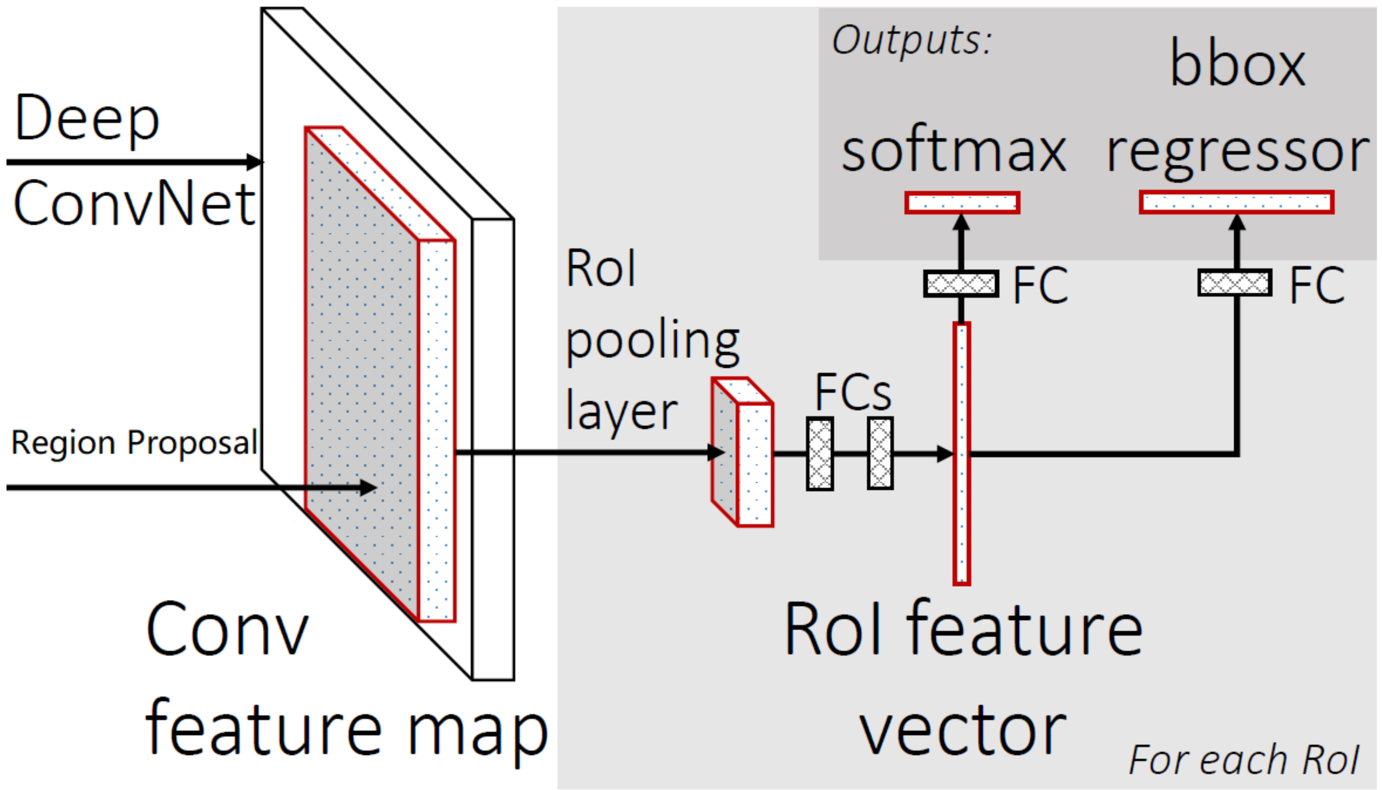
Table 1: the learned average proposal size for each anchor using the ZF net (numbers for $s = 600$).

anchor	128^2, 2:1	128^2, 1:1	128^2, 1:2	256^2, 2:1	256^2, 1:1	256^2, 1:2	512^2, 2:1	512^2, 1:1	512^2, 1:2
proposal	188×111	113×114	70×92	416×229	261×284	174×332	768×437	499×501	355×715

- 算法允许比潜在感受野更大的预测。这样的预测并非不可能——如果一个物体的中心是可见的，仍然可以粗略地推断出这个物体的范围（管中窥豹）.

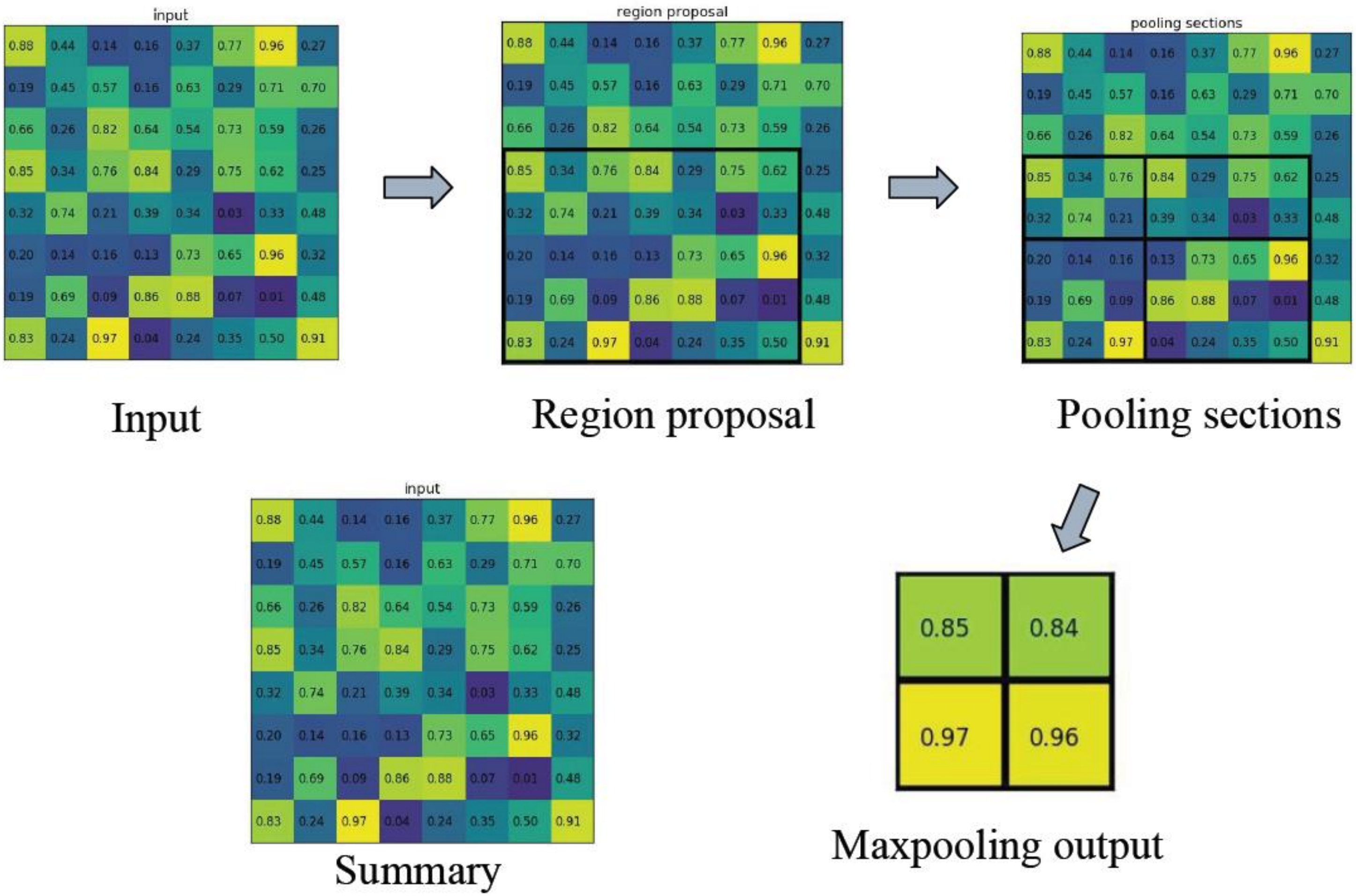
Faster R-CNN

Detection Network use Fast R-CNN



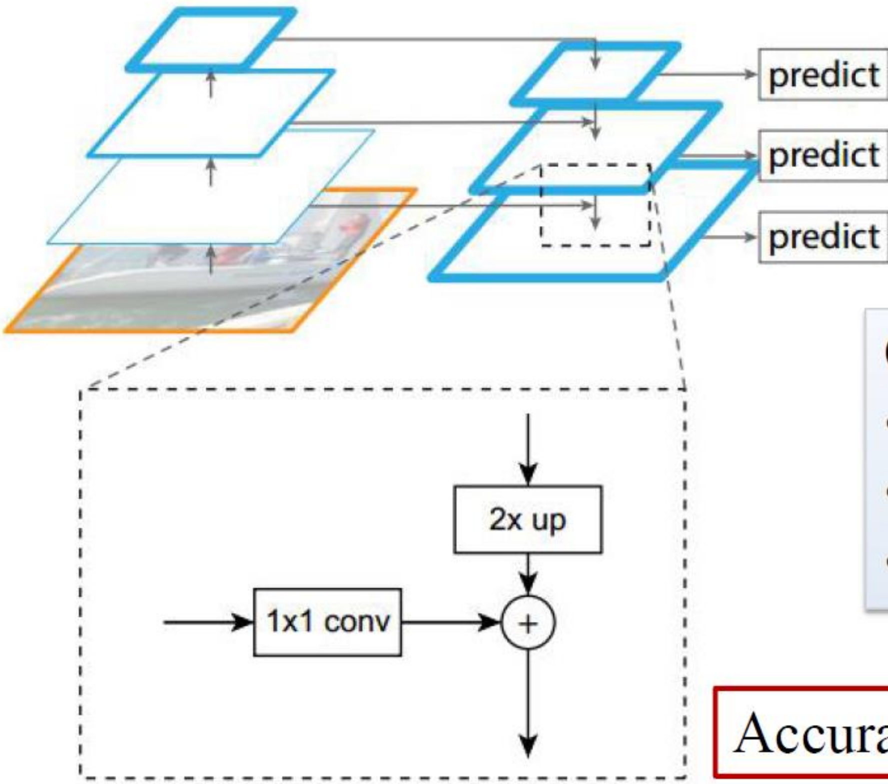
Faster R-CNN

ROI pooling



FPN

Feature Pyramid Network



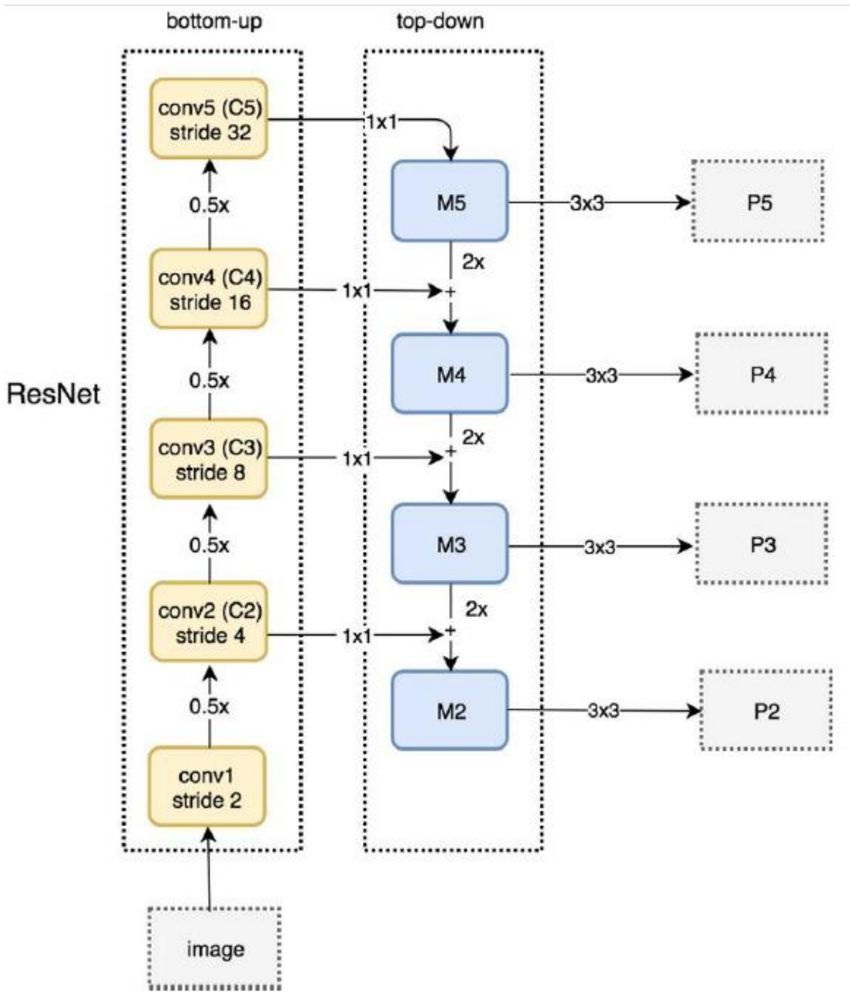
Construction：
- Bottom-up pathway
- Top-down pathway
- Lateral connections

Accurate location + High semantics

➢ FPN是一种具有横向连接的自顶向下体系结构，用于构建各种尺度的高级语义特征图。

FPN

ResNet-FPN



Down to top

ResNet backbone
Output: C2, C3, C4, C5
Strides: 4, 8, 16, 32

Top to down

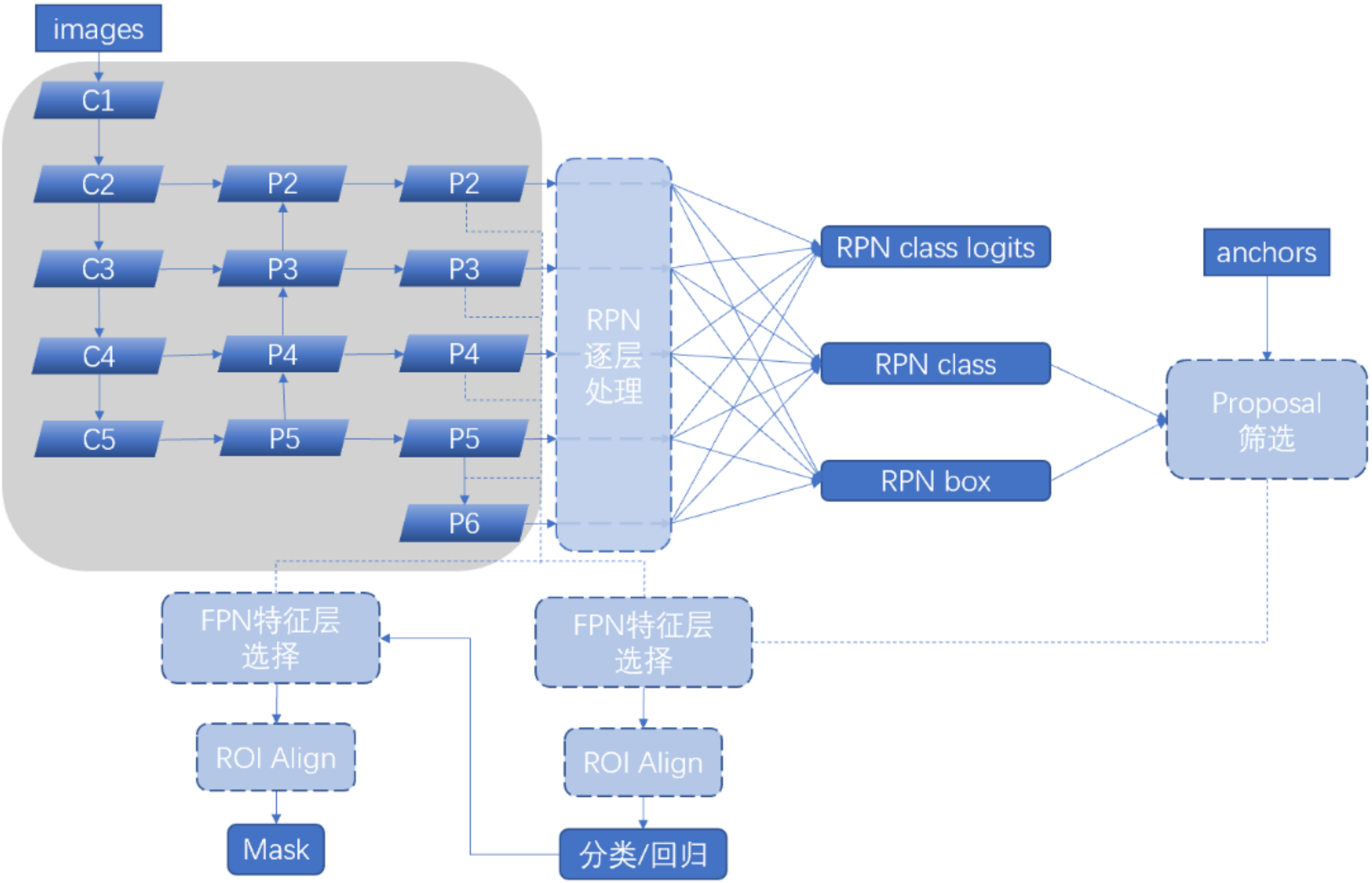
Nearest neighbor upsample
Store high-level semantics

Lateral connections

Reduce channels by 1x1 conv
Enhance location information
Remove aliasing by 3X3 conv

Mask-RCNN

Mask-RCNN Inference Model



Mask-RCNN

Question: How to assign RoIs of different scales to the pyramid levels?

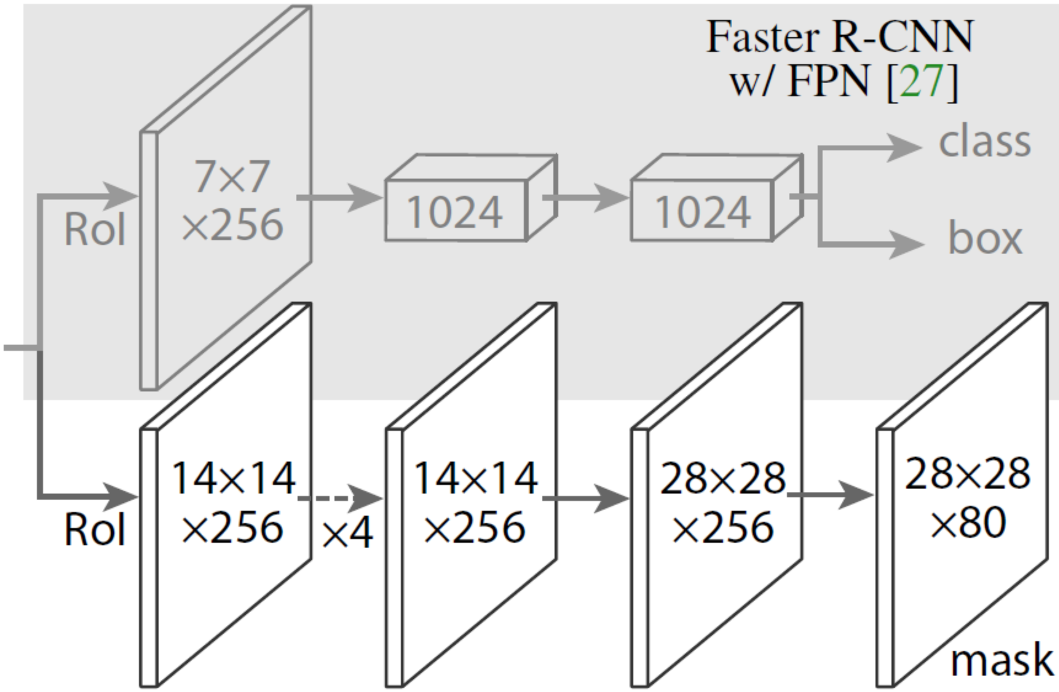
Assign an RoI of width **w** and height **h** to the level P_k of feature pyramid by:

$$k = \lfloor k_0 + \log_2(\sqrt{wh}/224) \rfloor.$$

- 224： is the canonical ImageNet pre-training size；
- k_0 ： is the target level on which an RoI with $w*h = 224^2$ should be mapped into.

Mask-RCNN

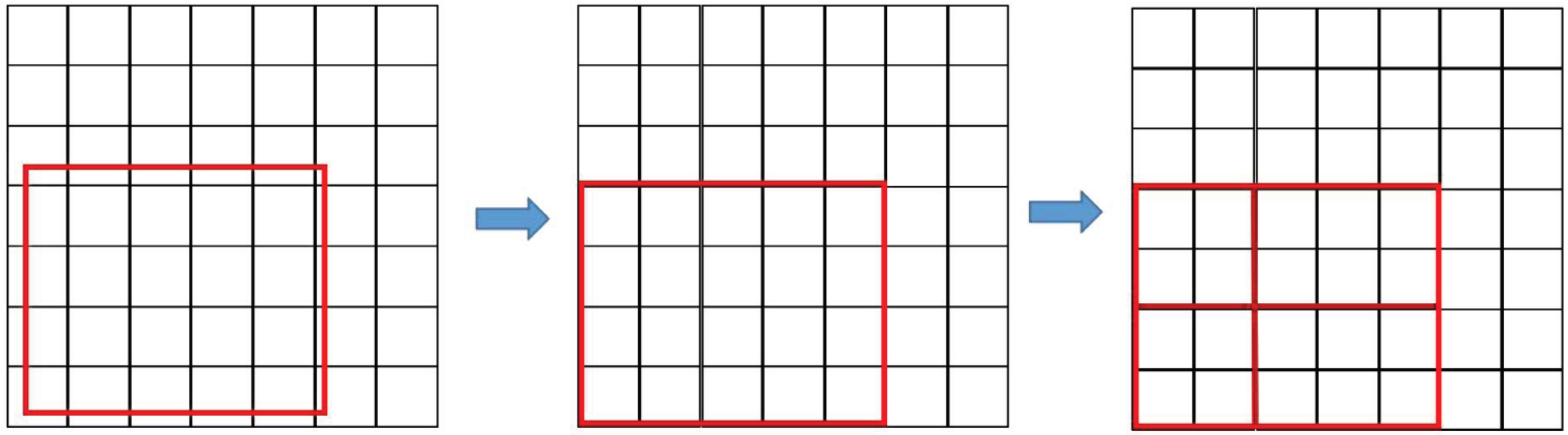
Mask-RCNN Head Architecture



- All convs are 3x3, except the output conv which is 1x1, deconvs are 2x2 with stride 2, and use ReLU in hidden layers.

RoI Align & RoI pooling

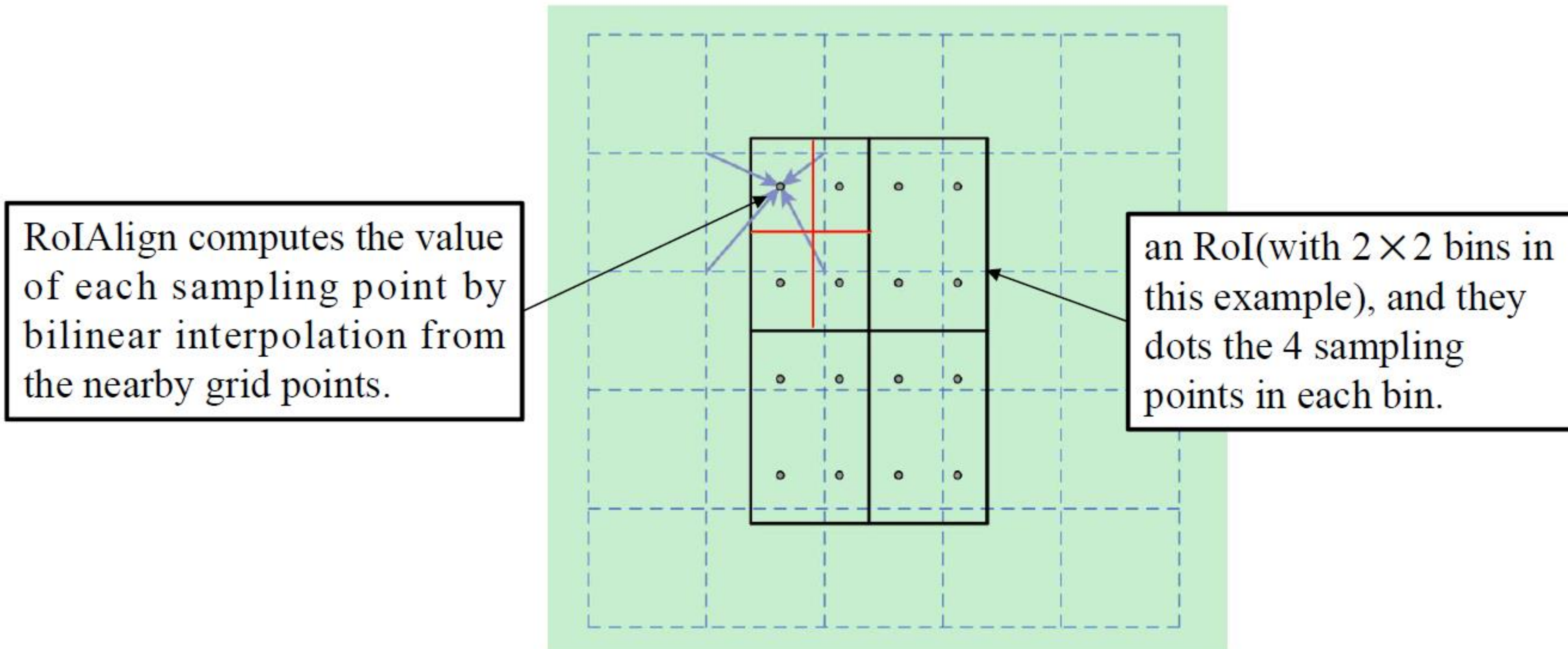
The misalign problem introduced by RoI pooling！



First quantization　　　Second quantization

These quantizations introduce misalignments between the RoI and the extracted features.
It has a large negative effect on predicting pixel-accurate masks!

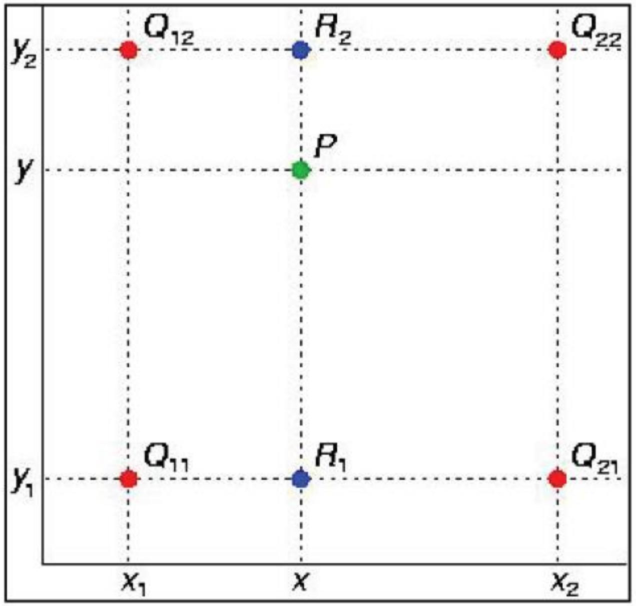
RoI Align & RoI pooling



RoIAlign is used to remove the harsh quantization of RoIPooling

RoI Align & RoI pooling

Bilinear interpolation



$$f(R_1) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}) \quad \text{Where} \quad R_1 = (x, y_1),$$

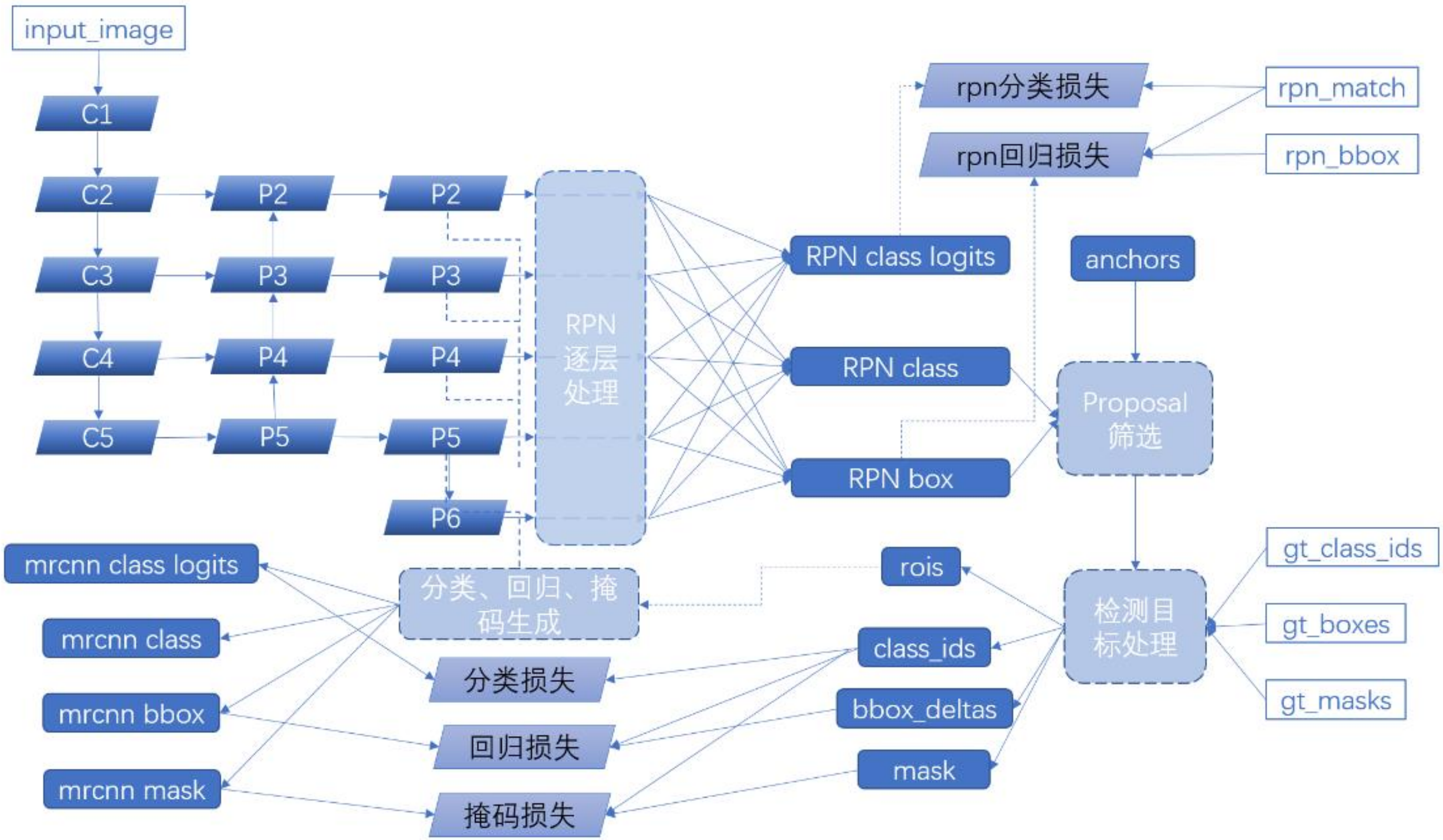
$$f(R_2) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}) \quad \text{Where} \quad R_2 = (x, y_2).$$

$$\Downarrow$$

$$f(P) \approx \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2).$$

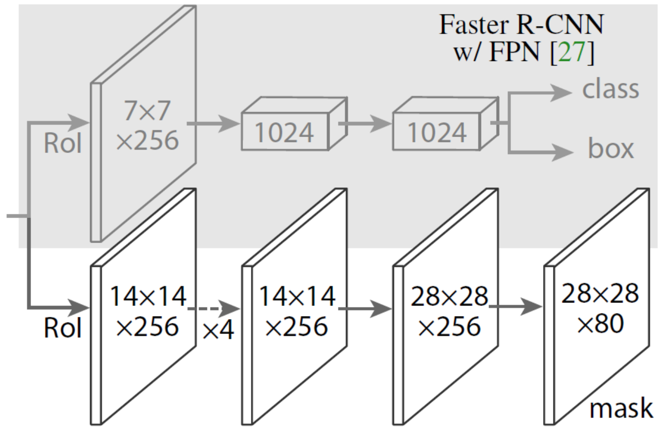
Mask-RCNN

Mask-RCNN training Model



Mask-RCNN

Loss function



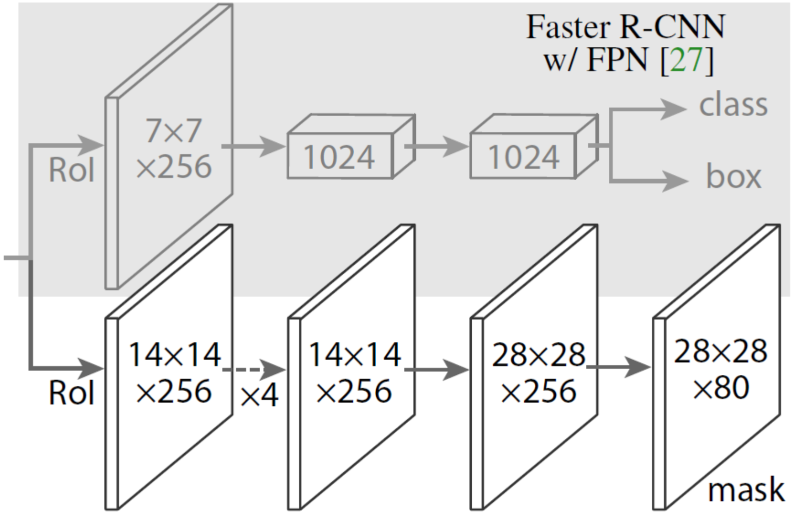
A multi-task loss on each sampled RoI as：

$$L = L_{cls} + L_{box} + L_{mask}$$

- The classification loss L_{cls} and bounding-box loss L_{box} are identical as those defined in Faster R-CNN.

Mask-RCNN

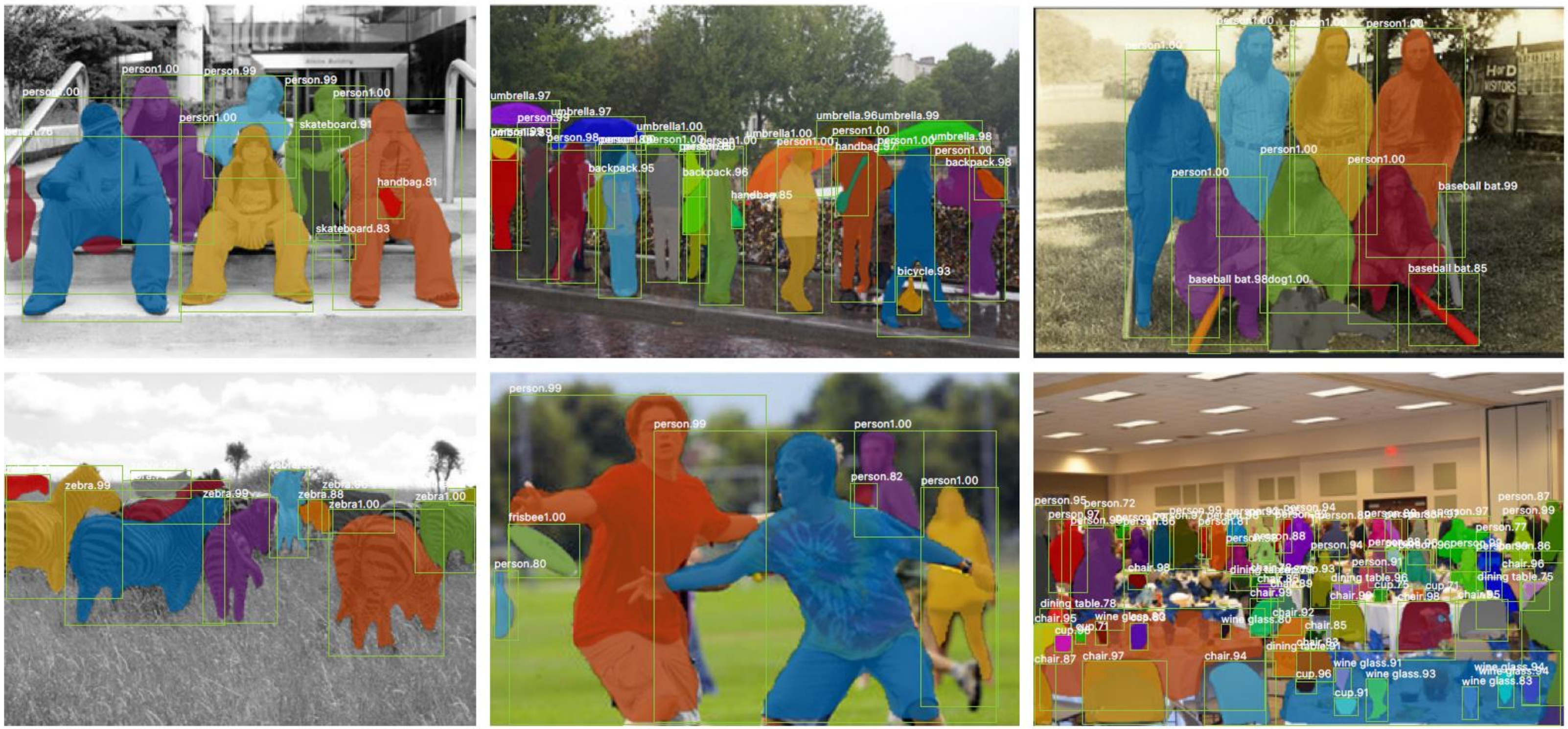
Loss function



- 对于每个RoI，mask 分支产生一个Km^2-d 的输出, 对应K个类别.
- 对每个28x28-d 的输出的每个像素使用sigmoid激活, L_{mask} 定义为所有像素上的平均二分类交叉熵损失.
- RoI 对应的ground-truth为k，则只有第k个mask对L_{mask} 产生贡献(其他mask 输出不对loss产生贡献).

Applications

Instance segmentation



Mask R-CNN results on the COCO test set

Applications

Human pose estimation

Keypoint detection results on COCO test

Thank you!