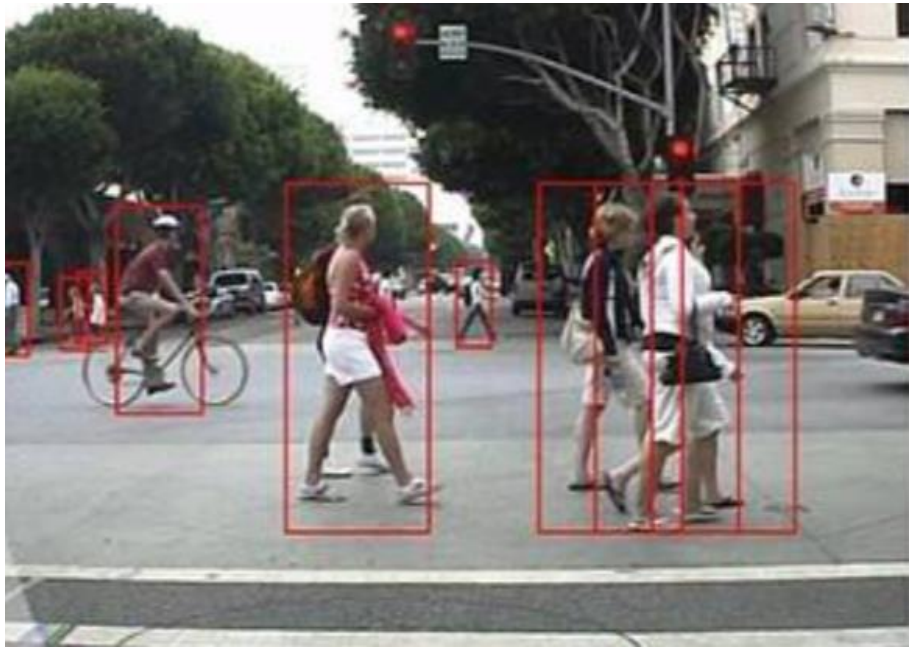
Feature Pyramid Networks for Object Detection

李自豪

Motivation

图像识别任务中目前存在的两大难点：

角度变化敏感

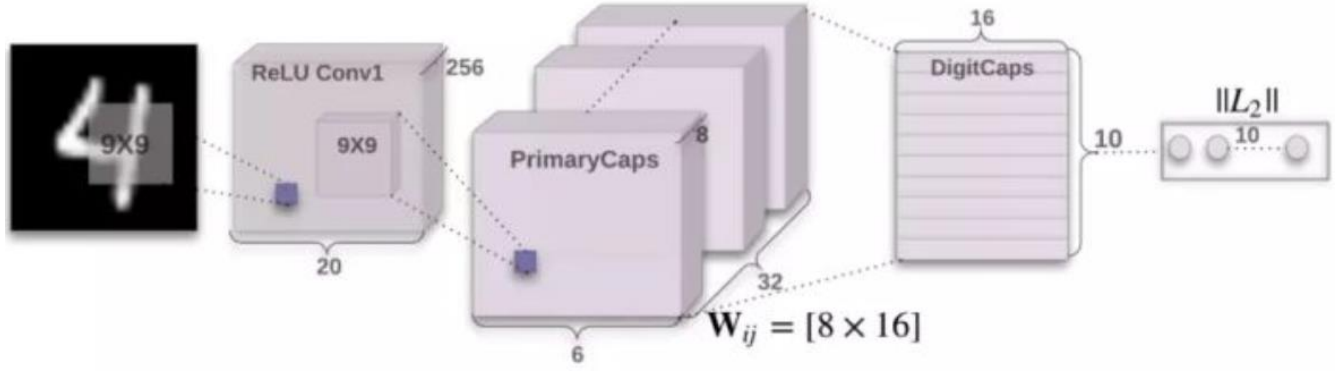


小尺寸物体检测困难

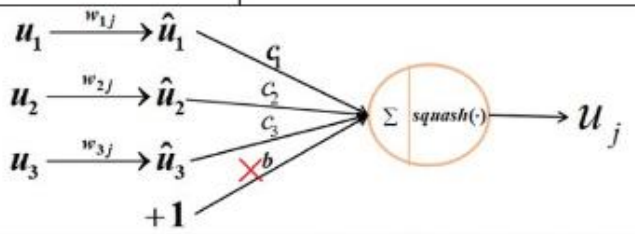
中国科学院大学
University of Chinese Academy of Sciences

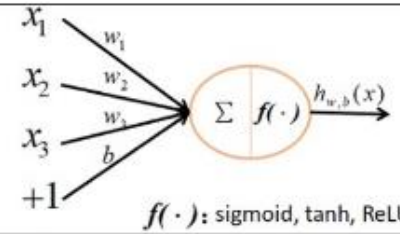
Perspectives Challenge

CapsuleNet从计算结图形学出发，考虑对象部件间的分层、位置关系将图像的角度、空间等信息编码为向量，同时计算物体的存在的概率。



		capsule	vs.	traditional neuron
Input from low-level neuron/capsule		vector(u_i)		scalar(x_i)
Operation	Affine Transformation	$\hat{u}_{j\|i} = W_{ij} u_i$ (Eq. 2)		–
	Weighting	$s_j = \sum_i c_{ij} \hat{u}_{j\|i}$ (Eq. 2)		$a_j = \sum_{i=1}^{3} W_i x_i + b$
	Sum			
	Non-linearity activation fun	$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2} \frac{s_j}{\|s_j\|}$ (Eq. 1)		$h_{w,b}(x) = f(a_j)$
output		vector(v_i)		scalar(h)





University of Chinese Academy of Sciences

Scales challenge

a.Featurized image pyramid

对图像进行缩放，提取Sift、HOG等特征。传统手工设计特征描述子时代的“万金油”。

b.Single feature map

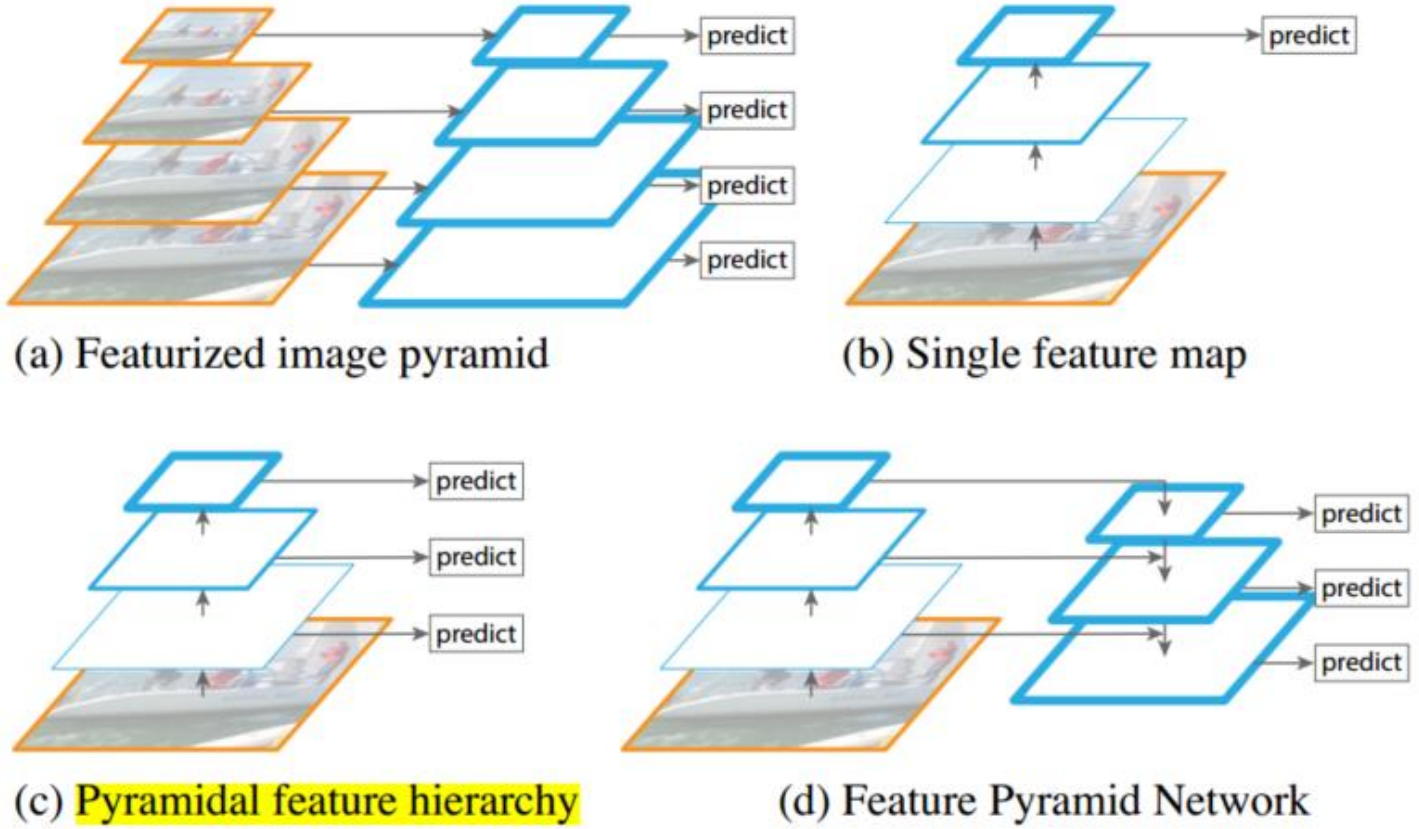
深度网对图像进行卷积，使用最顶层feature map进行分类、检测。

c.Pyramidal feature hierarchy

SSD，利用多层feature map同时进行目标检测，shallow layer检测小目标，deep layer检测大目标。

d.Feature Pyramid Network

融合深层的语义信息以及浅层的细节信息，提高微小目标的检测效果。

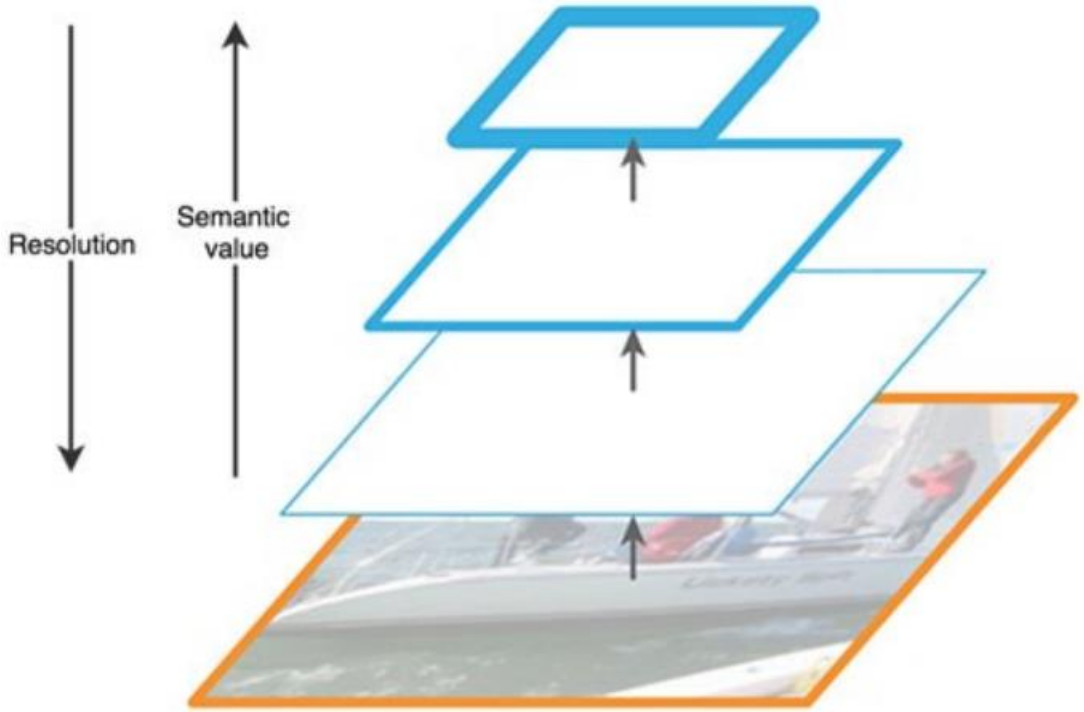


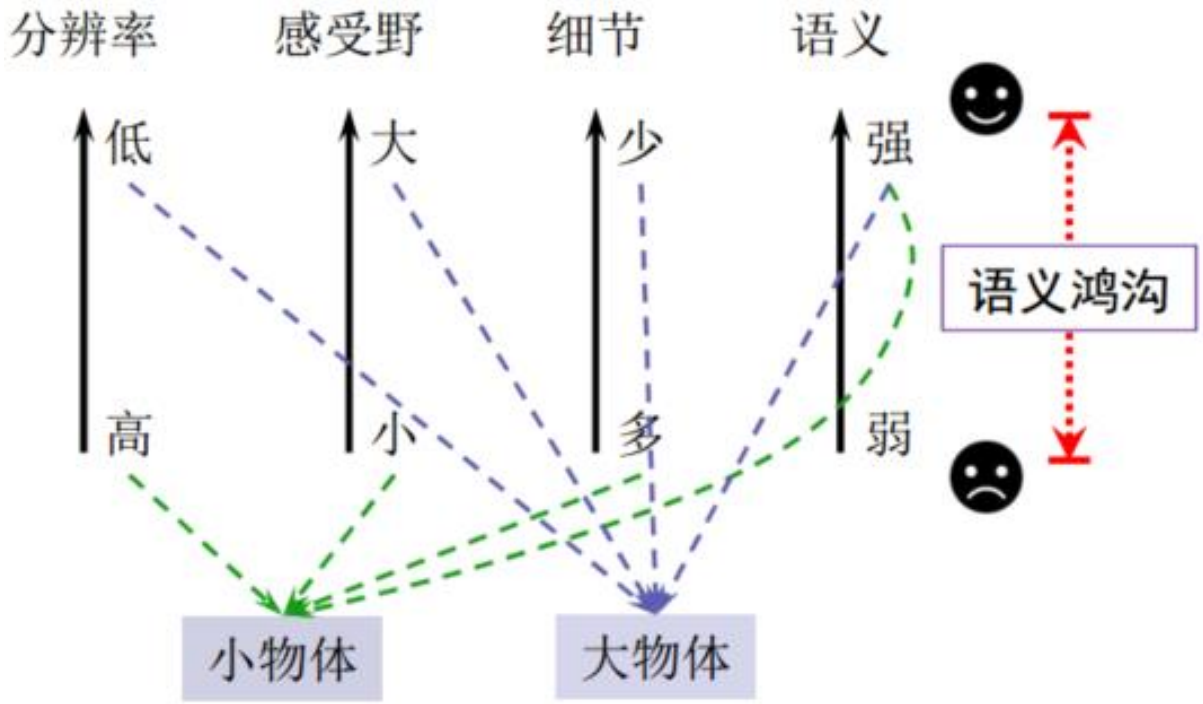
(a) Featurized image pyramid (b) Single feature map

(c) Pyramidal feature hierarchy (d) Feature Pyramid Network

University of Chinese Academy of Sciences

Shallow level vs Deep level





对于微小目标的检测我们既需要较高的分辨率即更多的细节信息，同时，也希望融入深层的语义信息。因此很自然的想法是进行特征融合，并对小尺度目标使用Shallow level feature map，而大尺寸目标则可以使用Deep levelfeature map。

中国科学院大学
University of Chinese Academy of Sciences

Feature Pyramid Networks

- **Feature pyramid**

利用深度网融合“从下至上”和“从上至下”的特征，使其既具有深层抽象的语义信息以及浅层高分辨率的细节信息。

- **Region Proposal Network(RPN)**

引入Anchor Box，利用RPN子网络生成Region Proposals，同时对其进行、分类回归。

- **Fast R-CNN**

利用RoI Pooling操作提取固定维度特征，此后送入全连接层对目标进行分类Bbox进行回归。

FPN = Feature pyramid + RPN + Fast R-CNN

University of Chinese Academy of Sciences

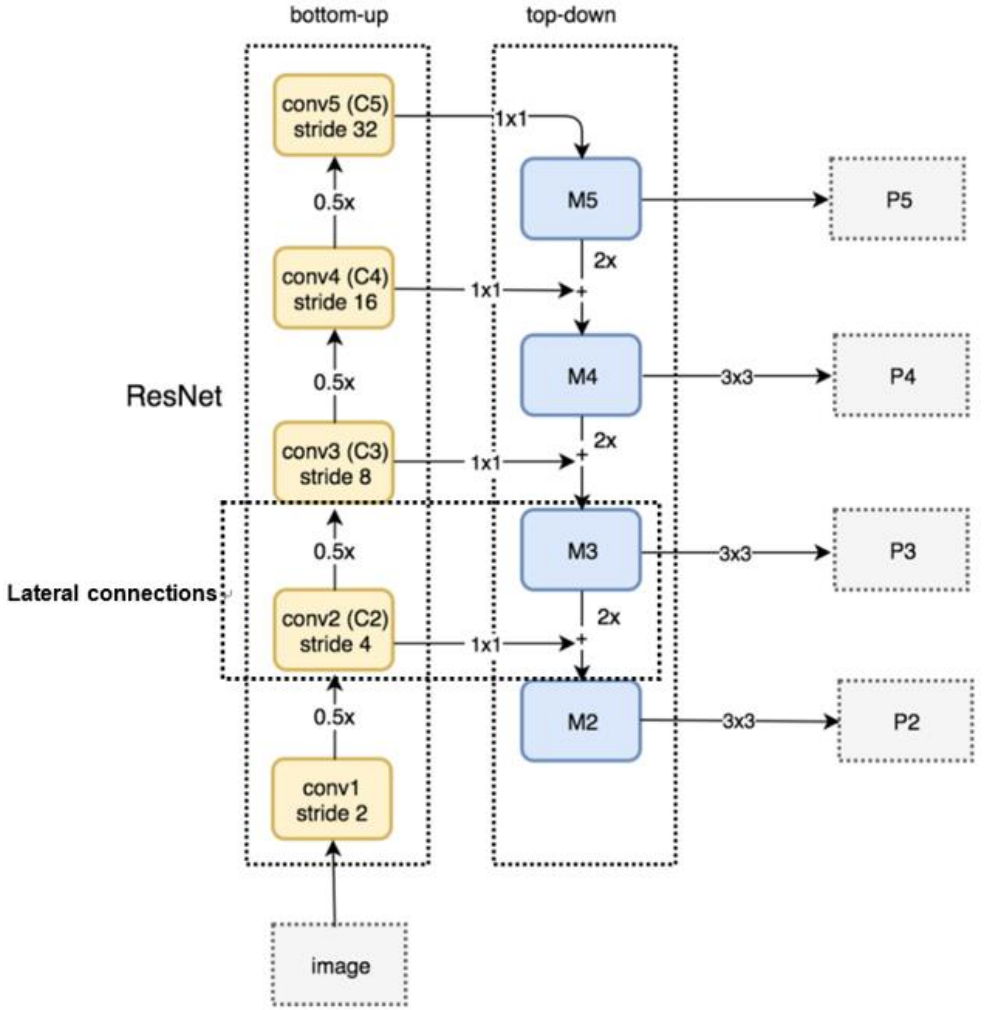
Feature Pyramid

- **Bottom-up pathway**

网络以ResNet50 or 101为backbone，取ResNet后4个block的顶层feature map为输出，记为$\{C_2, C_3, C_4, C_5\}$。其中，原始图像尺寸分别为$\{C_2, C_3, C_4, C_5\}$的4,8,16,32倍。

- **Top-down pathway**

首先对C_5过一个1×1的卷积核使其Channel数为256，记为M_5。然后在此基础上进行三次upsampling,放大倍数分别为2,4,8，得到M_4, M_3, M_2。

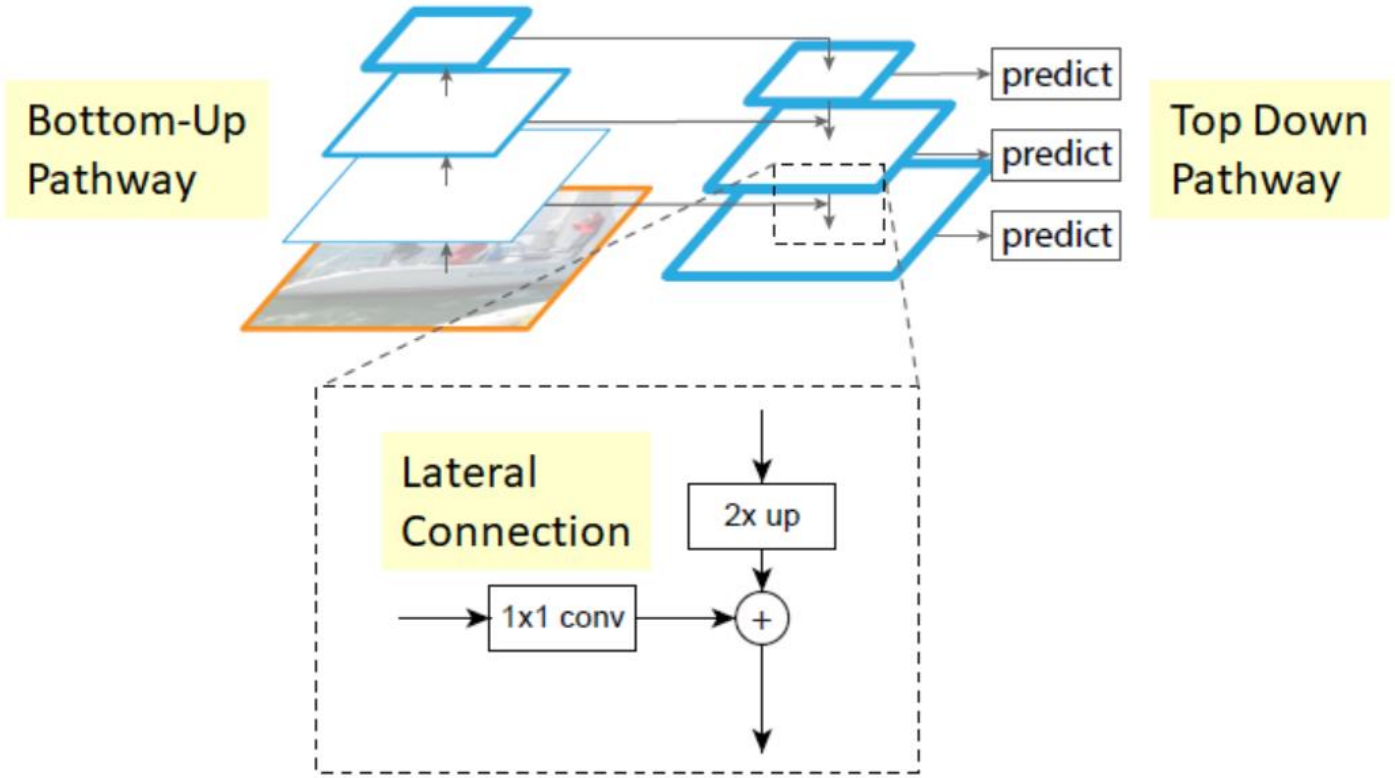


University of Chinese Academy of Sciences

Feature Pyramid

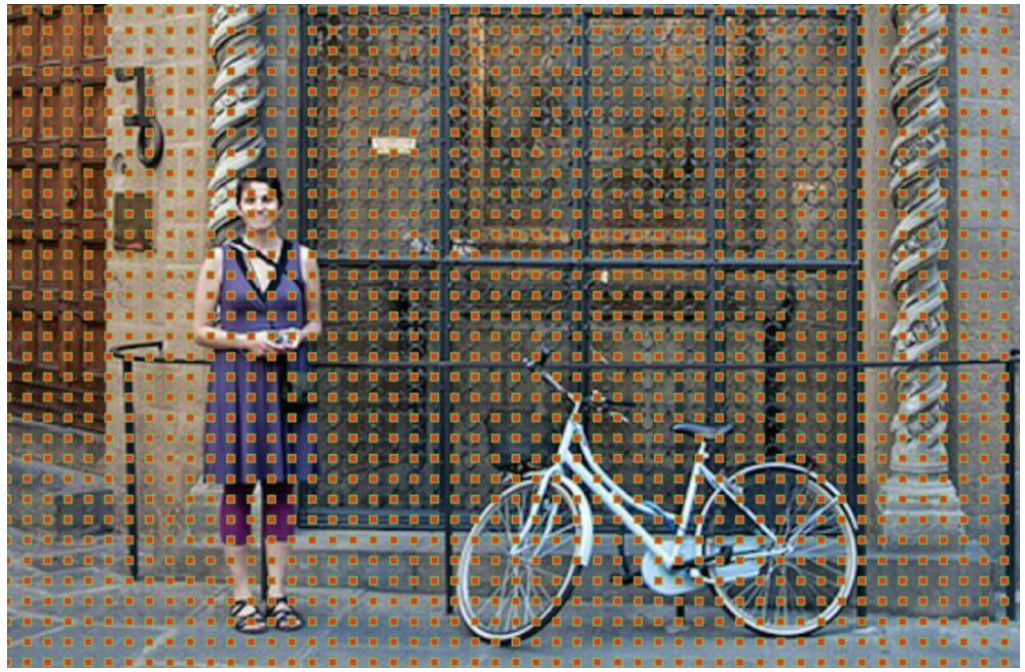
- **Lateral Connection**

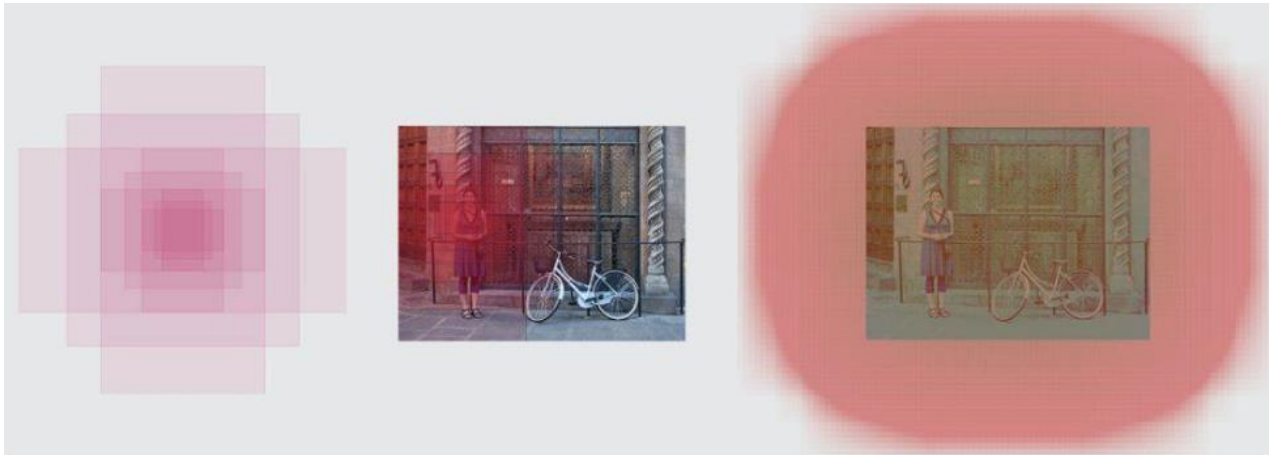
为保证Channel数目相同，同时融合多个Channel的信息，首先利用1×1的卷积核对C_2, C_3, C_4进行卷积，然后将其与M_2, M_3, M_4对应相加，最后在进行3×3的卷积处理，已消除混叠效应，得到P_2, P_3, P_4，其中$P_5 = M_5$。因此$\{P_2, P_3, P_4, P_5\}$即为融合了细节信息与抽象信息的Feature Map，接下来将其送入RPN网络中生成RoIs。



University of Chinese Academy of Sciences

Region Proposal Network

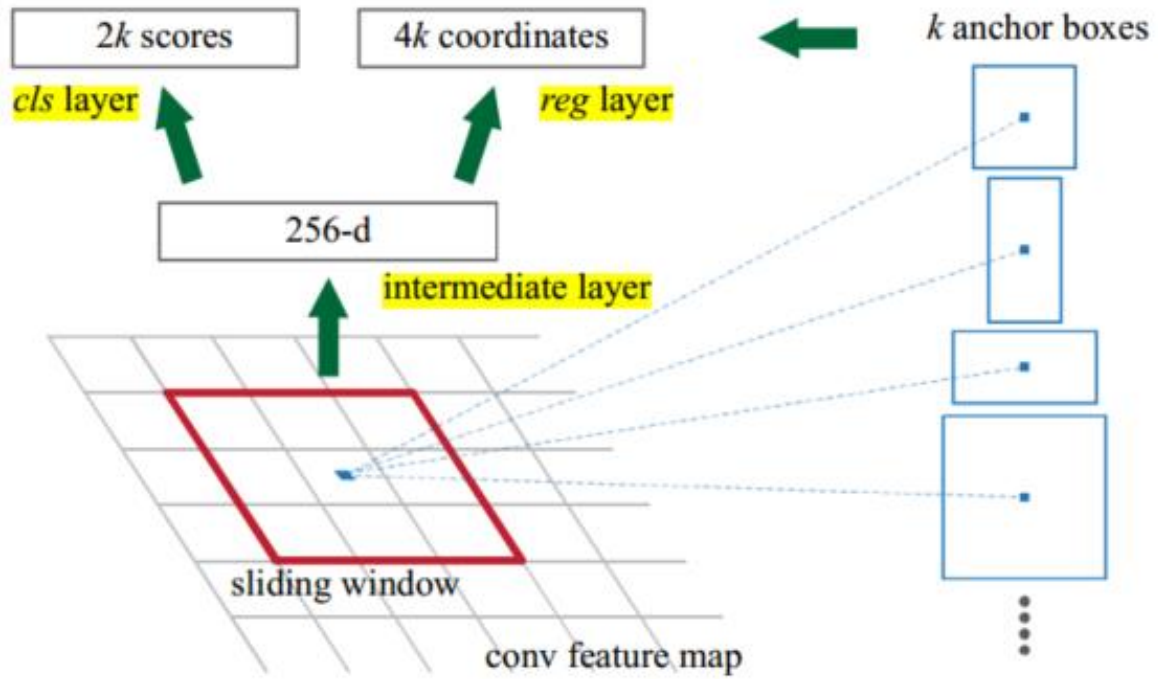




1. 将原始图片划分为$m \times n$的网格（m,n为feature map尺寸），以网格线交点为锚点生成k个不同尺寸的锚框，以满足不同目标大小的检测需求。（Faster R-CNN中共设置$3 \times 3 = 9$个anchor box，3个不同尺度$128^2, 256^2$，512^2，3个不同比列2:1,1:1,1:2进行组合）

University of Chinese Academy of Sciences

Region Proposal Network



2. 在深度网卷积所得的feature map上首先利用3 × 3的窗口进行滑动卷积；
3. 此后外接两个大小为1×1的并行卷积核，其输出channels分别为2k和4k，在anchor box的基础上进行位置的修正，同时判断是否包含检测目标，以生成region proposal。（k为anchor box的数量；2表示0-1，即判断该BBox是背景或前景；4表示Bbox位置(x,y,w,h)的修正值(t_x,t_y,t_w,t_h)）。

中国科学院大学
University of Chinese Academy of Sciences

RPN Loss function

- p_i^*, t_i^*, p_i, t_i分别代表ground truth box的类别，中心坐标、长、宽以及predict box的类别中心坐标、长、宽；

- $x, x^*, x_a; y, y^*, y_a; w, w^*, w_a; h, h^*, h_a$分别为predict box, ground truth box以及default box的中心坐标、长宽。

- RPN Loss主要由分类的交叉熵损失，以及回归的smooth L1损失两部分组成。对于回归损失只考虑前景目标。

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

$$t_x = (x - x_a)/w_a, \quad t_y = (y - y_a)/h_a$$

$$t_w = log(w/w_a), \quad t_h = log(h/h_a)$$

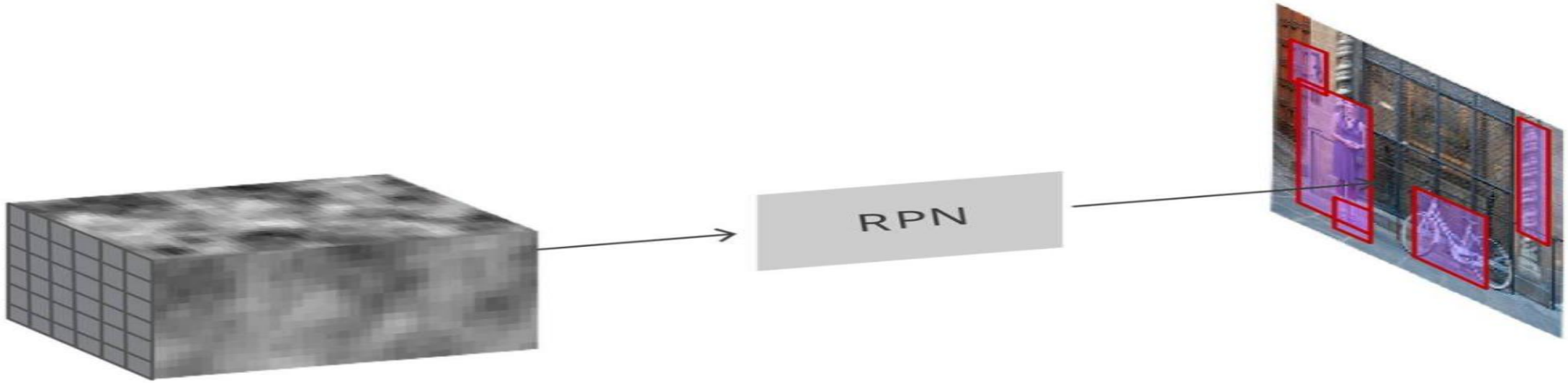
$$t_x^* = (x^* - x_a)/w_a, \quad t_y^* = (y^* - y_a)/h_a$$

$$t_w^* = log(w^*/w_a), \quad t_h^* = log(h^*/h_a)$$

$$smooth_{L_1} = \begin{cases} 0.5x^2 & \text{if } |x| \leq 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}$$

中国科学院大学
University of Chinese Academy of Sciences

RPN for FPN



FPN中的RPN与Faster R-CNN中的RPN有细微的差别，主要体现在：

1. FPN中将RPN称之为predictor head，并将$\{P_2, P_3, P_4, P_5\}$的feature map均送入至RPN网络生成RoIs，且参数共享。
2. 对于Anchor Box尺寸的设计，FPN网络由于其设计初衷即为不同level的feature map检测不同尺度的目标，因此其在不同的level使用不同尺寸的anchor box，相同level使用同一尺寸anchor box，只是进行比例变化。对于$\{P_2, P_3, P_4, P_5\}$，其尺寸分别对应为$\{32^2, 64^2, 128^2, 256^2\}$。

中国科学院大学
University of Chinese Academy of Sciences

Fast R-CNN

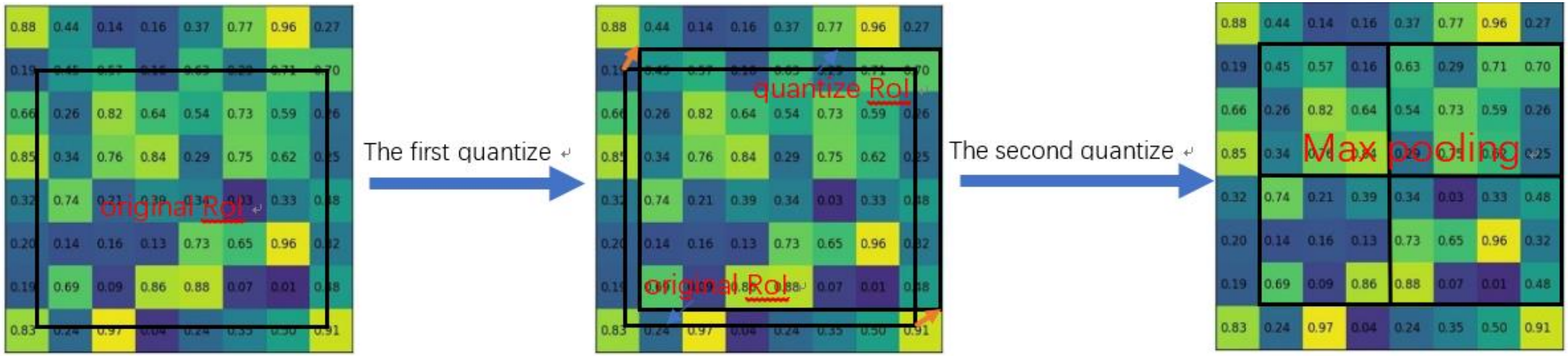
Faster R-CNN中RPN产生的RoIs在同一尺度的Feature map上提取特征，而FPN中产生的不同尺度的RoIs需分配给不同level的feature map提取特征，这里定义：

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

- K为被分配的层级；
- w, h分别为RoIs的宽和长（相对原始图片大小）；
- k_0 =4，224为预训练ImageNet图片大小。
- 若RoI的尺寸为112×112，则$k = 3$。即检测目标尺寸越小，将被分配到更低，分辨率更精细的层级。

中国科学院大学 University of Chinese Academy of Sciences

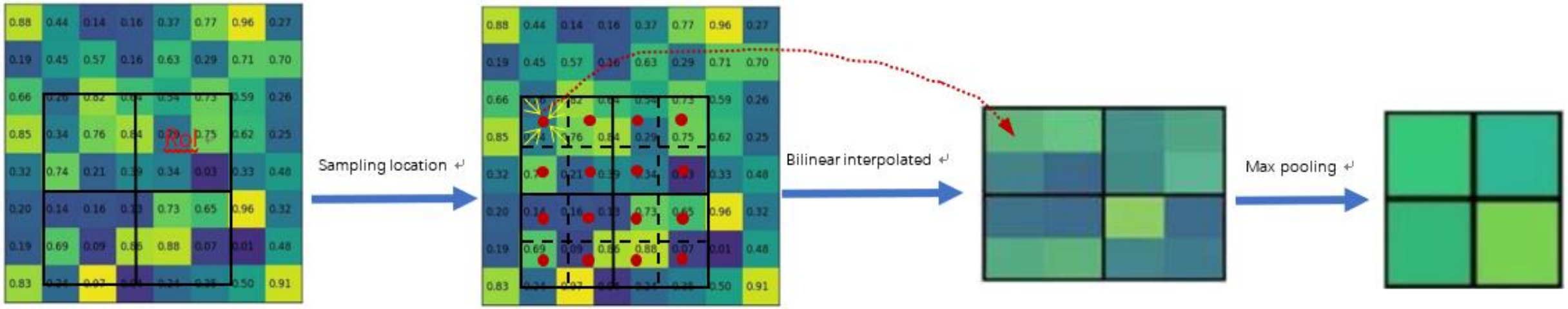
RoI Pooling



- SPPNet中的RoI Pooling操作将存在两次量化的过程：RoIs区域并非整数，需对其坐标进行量化；对feature map的划分并非能取整，需再次量化。特征图上的误差反映到原始图像上将成倍的放大，这样即对BBox位置的预测带来了精度的损失。

中国科学院大学
University of Chinese Academy of Sciences

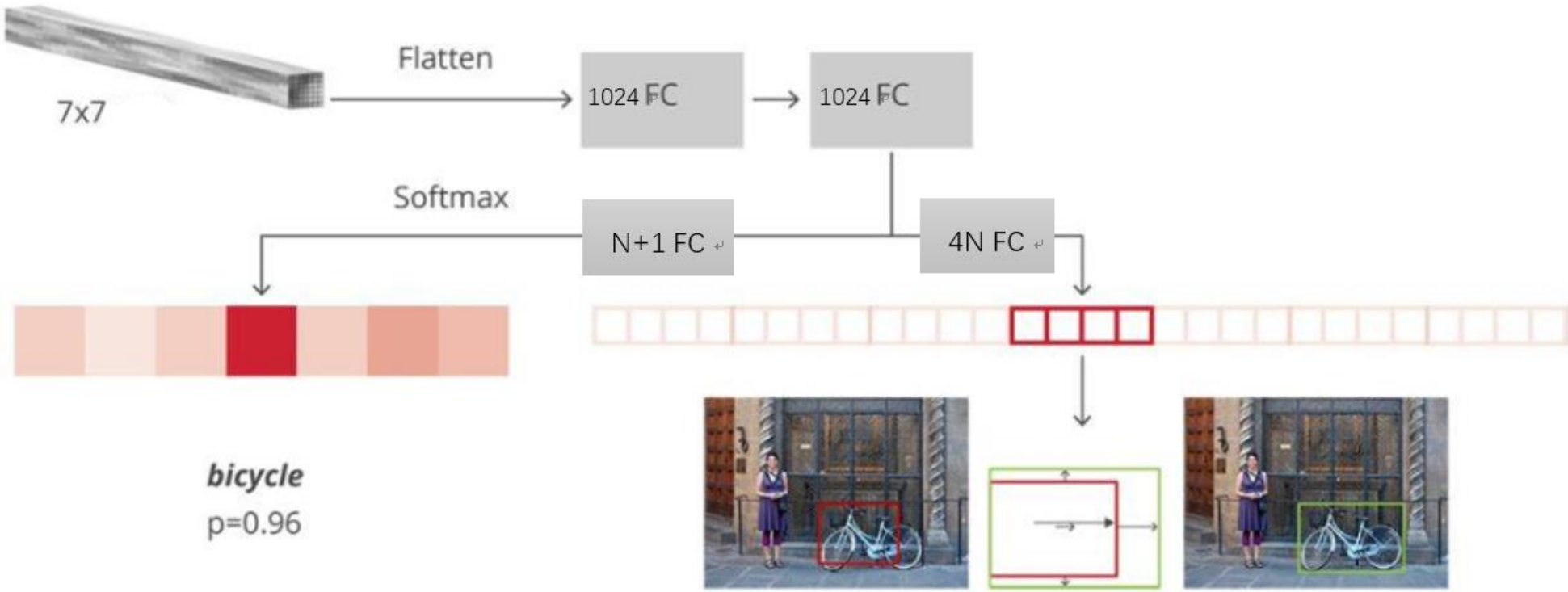
RoI align



- Mask R-CNN中，保持RPN输出的RoI位置坐标为小数，将RoIs区域中的每个bin区域再次划分为4×4的单元，采用双线性内插法计算每一个单元中心特征，然后进行最大池化操作，以提高精度。

University of Chinese Academy of Sciences

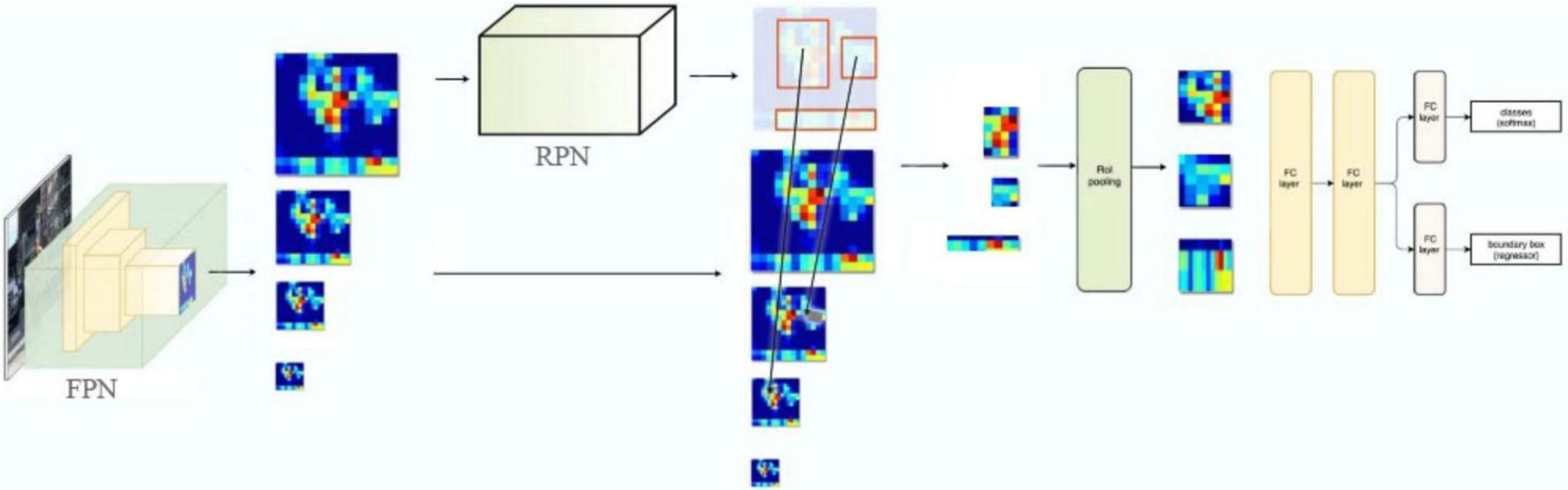
Full Connection



- 将每一个RoI Feature展平，后接两个1024的全连接层；
- 此后，再分别使用N+1FC+Softmax进行分类，4NFC进行BBox回归。其中N为类别数，1表示背景。

中国科学院大学
University of Chinese Academy of Sciences

FPN Architecture



1. 以ResNet为backbone，融合“自下而上”和“自上而下”的语义信息和细节信息；
2. 利用RPN网络对各个level的Anchor box进行分类、回归，生成RoIs；
3. 将不同尺寸的RoIs对应至不同level的feature map，利用RoI Pooling提取等维特征；
4. 后接两个全连接层，再接两个并行的全连接层对Region proposal进行分类和回归。

中国科学院大学
University of Chinese Academy of Sciences

Experiments （RPN+FPN）

RPN	feature	# anchors	lateral?	top-down?	AR^{100}	AR^{1k}	AR_s^{1k}	AR_m^{1k}	AR_l^{1k}
(a) baseline on conv4	C_4	47k			36.1	48.3	32.0	58.7	62.2
(b) baseline on conv5	C_5	12k			36.3	44.9	25.3	55.5	64.2
(c) **FPN**	$\{P_k\}$	200k	✓	✓	**44.0**	**56.3**	**44.9**	**63.4**	66.2
Ablation experiments follow:									
(d) bottom-up pyramid	$\{P_k\}$	200k	✓		37.4	49.5	30.5	59.9	**68.0**
(e) top-down pyramid, w/o lateral	$\{P_k\}$	200k		✓	34.5	46.1	26.5	57.4	64.7
(f) only finest level	P_2	750k	✓	✓	38.4	51.3	35.1	59.7	67.6

Bounding box proposal results of RPN on the COCO minival set

- 对于RPN，使用conv5(b)与使用conv4(a)相比并没有很大优势；
- 融入FPN(c)后RPN无论在小物体还是大物体上其AR均有明显提升，这说明FPN对尺度鲁棒;
- 除去top-down pathway(d)后（类似SSD）其AR将有所损失，但仍然是次优的，这启示我们即使不融合特征，对不同尺度的目标使用不同level的feature map其AR也会提高;
- 移除lateral connection其AR也会降低;
- 仅使用最底层的feature map $P_2(f)$ 其效果也优于baseline，这说明，当融合特征后对结果的提升也将有帮助。

中国科学院大学 University of Chinese Academy of Sciences

Experiments （Fast R-CNN+FPN）

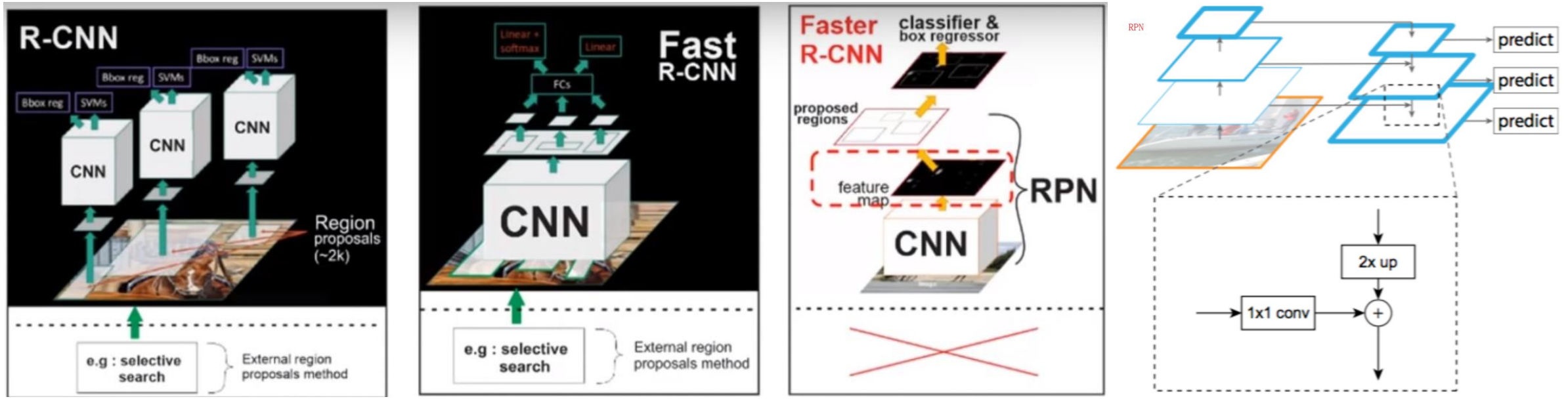
Fast R-CNN	proposals	feature	head	lateral?	top-down?	AP@0.5	AP	AP_s	AP_m	AP_l
(a) baseline on conv4	RPN, $\{P_k\}$	C_4	conv5			54.7	31.9	15.7	36.5	45.5
(b) baseline on conv5	RPN, $\{P_k\}$	C_5	2*fc*			52.9	28.8	11.9	32.4	43.4
(c) **FPN**	RPN, $\{P_k\}$	$\{P_k\}$	2*fc*	✓	✓	**56.9**	**33.9**	**17.8**	**37.7**	**45.8**
Ablation experiments follow:										
(d) bottom-up pyramid	RPN, $\{P_k\}$	$\{P_k\}$	2*fc*	✓		44.9	24.9	10.9	24.4	38.5
(e) top-down pyramid, w/o lateral	RPN, $\{P_k\}$	$\{P_k\}$	2*fc*		✓	54.0	31.3	13.3	35.2	45.3
(f) only finest level	RPN, $\{P_k\}$	P_2	2*fc*	✓	✓	56.3	33.4	17.3	37.3	45.6

Object detection results of detection network on the COCO minival set

- 与baseline on conv4(a)比较，FPN(c)AP提高了2.0，在小物体上 AP提升了2.1；
- 移除top-down connection(d)或移除lateral connection均导致结果变差，且移除top-down connection对小物体检测的影响更大；
- 若仅使用最底层的feature map P_2 ，与FPN相比其精度仅有少量损失。

University of Chinese Academy of Sciences

R-CNN to FPN



"化零为整"，精度、速度不断提高

University of Chinese Academy of Sciences

Reference

- Jonathan Hui. Understanding Feature Pyramid Networks for object detection (FPN). Medium.

- Javier Rey. Faster R-CNN: Down the rabbit hole of modern object detection.

- Tomasz Grel. Region of interest pooling explained. deepsense.ai


中国科学院大学
University of Chinese Academy of Sciences
感谢！