总结-CNN中的目标多尺度处理

作者：yyfyan
授权转载自知乎专栏
深度学习和图像分割

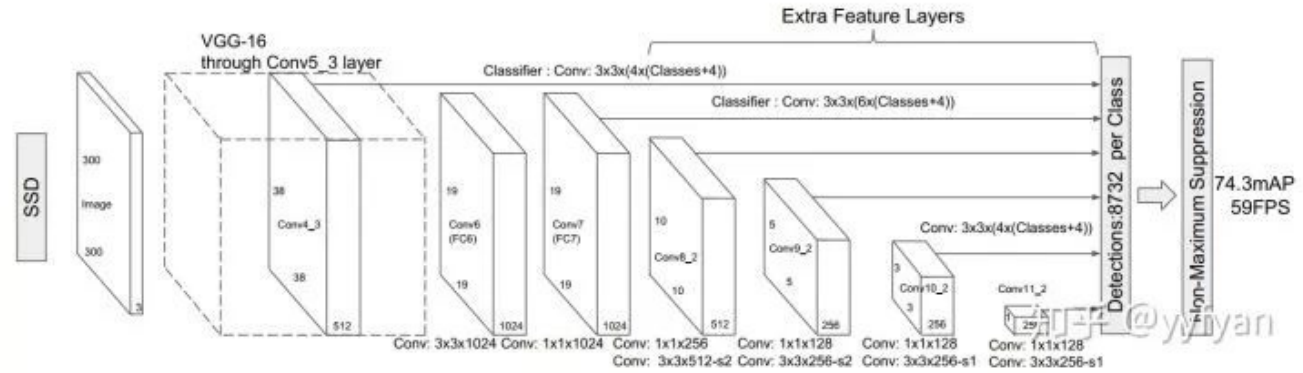
作者语：

1. 后面实习要解决实例分割中的目标多尺度问题(当然不只是这个问题，还有其他的)，为此对CNN中这几年的多尺度处理方法进行简要总结^_^，时间紧任务重，只记录了一点点东西，核心的还是要去看论文读代码。
2. 最近在准备开题的东西，其中也讨论了该问题，这里又扩展了一些论文；

视觉任务中处理目标多尺度主要分为两大类：

- **图像金字塔：经典的基于简单矩形特征(Haar)+级联Adaboost与Hog特征+SVM的DPM目标识别框架，均使用图像金字塔的方式处理多尺度目标，早期的CNN目标识别框架同样采用该方式，但对图像金字塔中的每一层分别进行CNN提取特征，耗时与内存消耗均无法满足需求。但该方式毫无疑问仍然是最优的。值得一提的是，其实目前大多数深度学习算法提交结果进行排名的时候，大多使用多尺度测试。同时类似于SNIP使用多尺度训练，均是图像金字塔的多尺度处理。**
- **特征金字塔：这个概念早在ACF目标识别框架的时候已经被提出(PS: ACF系列这个我前两年入过一段时间的坑，后来发现他对CPU内存要求太大，不过确实是前几年论文灌水利器，效果也还不错，但还是不能落地的，我已果断弃坑)。而在CNN网络中应用更为广泛，现在也是CNN中处理多尺度的标配。目前特征提取部分基本是FCN，FCN本质上等效为密集滑窗，因此不需要显示地移动滑动窗口以处理不同位置的目标。而FCN的每一层的感受野不同，使得看到原图中的范围大小不同，也即可以处理不同尺度的目标。因此，分析CNN中的多尺度问题，其实本质上还是去分析CNN的感受野，一般认为感受野越大越好，一方面，感受野大了才能关注到大目标，另一方面，小目标可以获得更丰富的上下文信息，降低误检。**

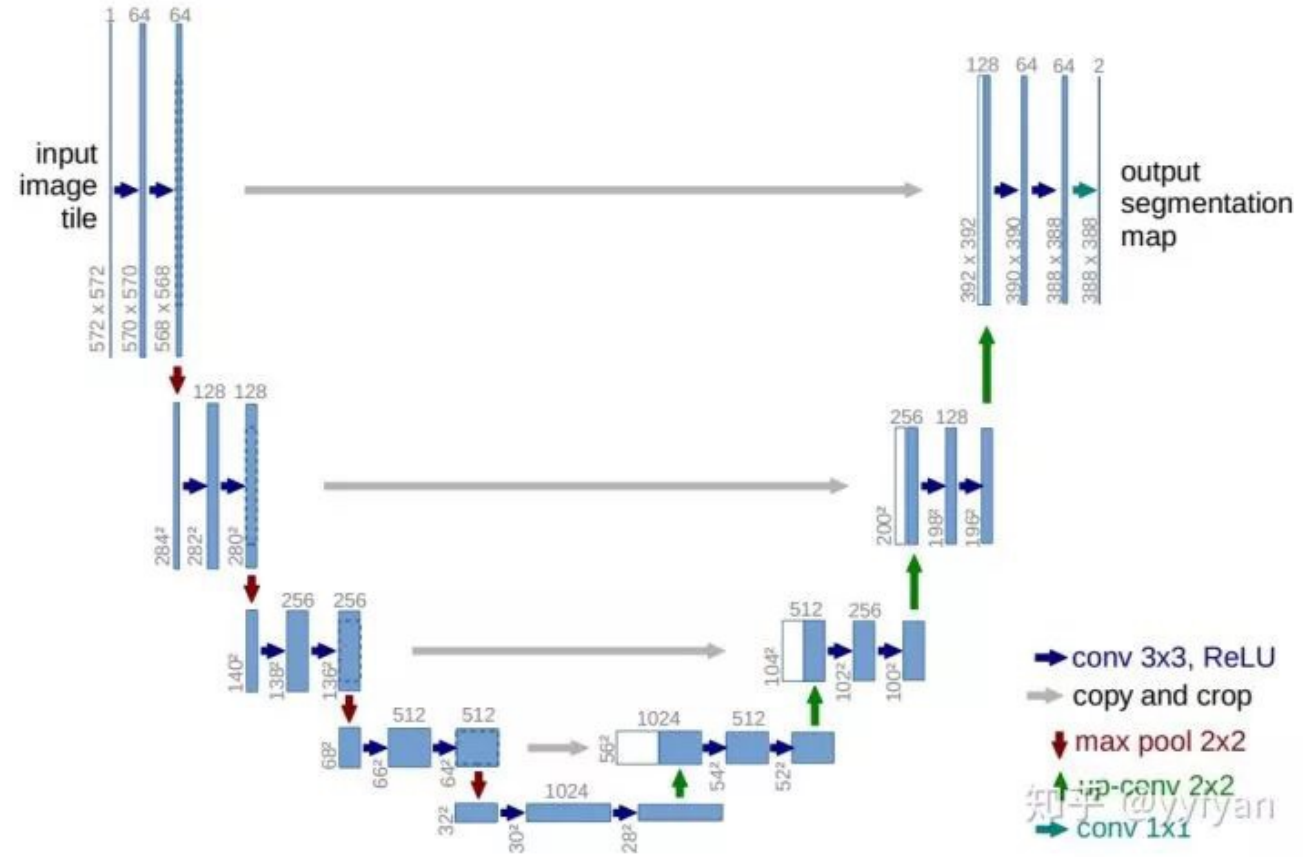
SSD中的多尺度处理



SSD以不同stride的feature map作为检测层分别检测不同尺度的目标，用户可以根据自己的任务的目标尺度制定方案。该方式尺度处理简单有效，但存在一些缺陷：

- 一般使用低层检测小目标，但低层感受野小，上下文信息缺乏，容易引入误检；
- 使用简单的单一检测层多尺度信息略显缺乏，很多任务目标尺度变化范围十分明显；
- 高层虽然感受野较大，但毕竟经过了很多次降采样，大目标的语义信息是否已经丢失；
- 多层特征结构，是非连续的尺度表达，是非最优的结果；

U-shape/V-shape型多尺度处理

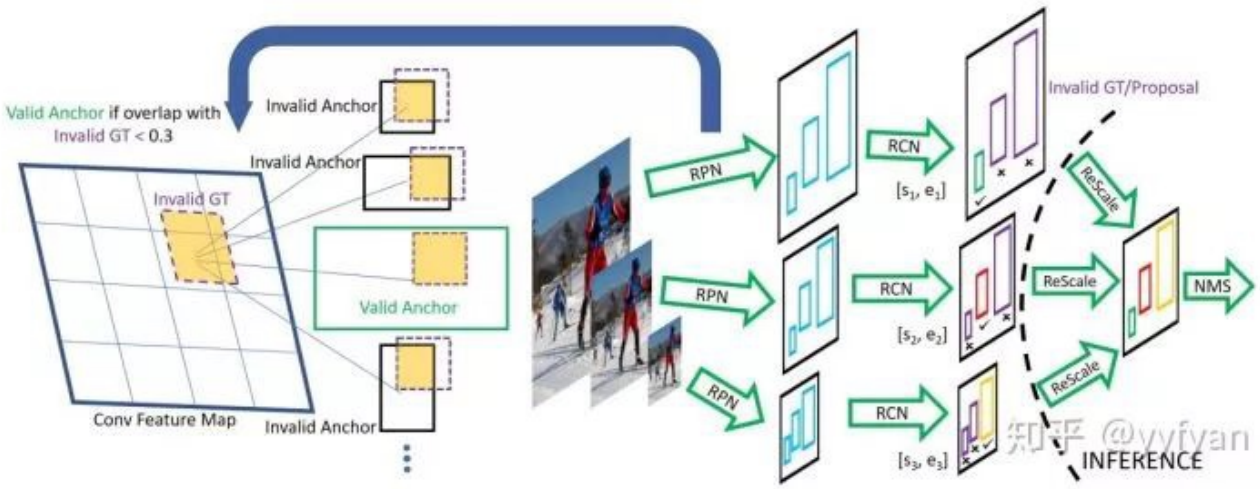


这种方式源于U-Net(不确定是不是~_~)，采用对称的encoder-decoder结构，将高层特征逐渐与低层特征融合，这样的操作类似于将多个感受野进行混合，使得输出的多尺度信息更为丰富；Face++团队在去年COCO(cocodataset.org/worksho)比赛上，在backbone最后加入gpooling操作，获得理论上最

大的感受野，类似于V-shape结构，结果证明确实有效。该方法虽然比SSD的单层输出多尺度信息相比更好，但其也存在问题：

- 由于decoder使用的通道数与encoder相同，导致了大量的计算量；
- 上采样结构不可能完全恢复已经丢失的信息；

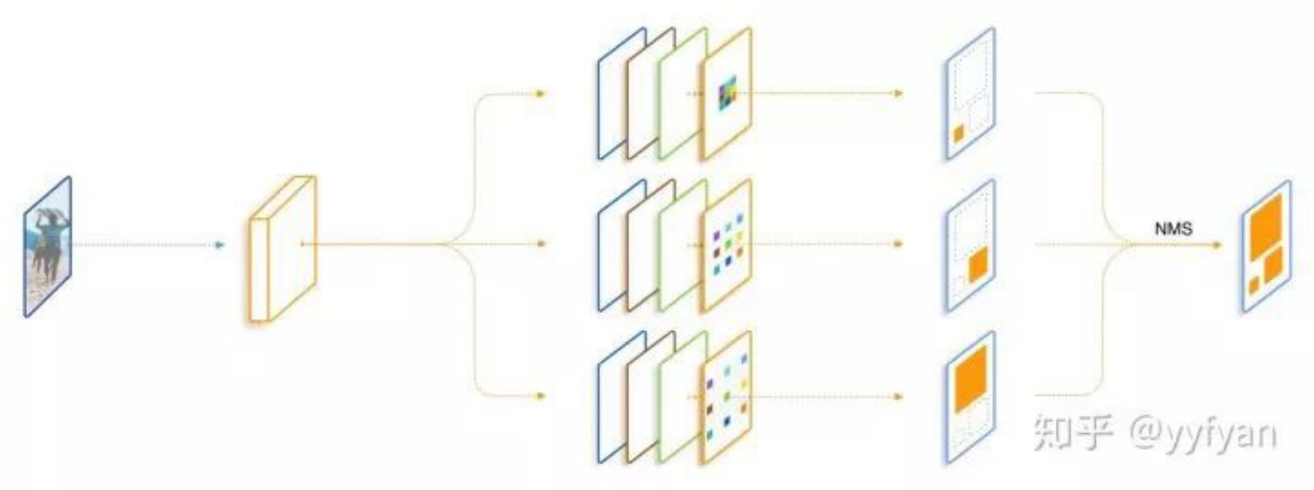
SNIP/SNIPER中的多尺度处理



- 训练与测试分辨率从不一致的时候性能会下降；
- 大分辨率输入图像虽然能提升小目标检测性能，但同时使得大目标过大导致其很难分类，此消彼长，最终精度提升并不明显；
- 多尺度训练(Mutil-Scale training)，采样到的图像分辨率很大（1400x2000），导致大目标更大，而图像分辨率过小时（480x640），导致小目标更小，这些均产生了非最优的结果；
- SNIP针对不同分辨率挑选不同的proposal进行梯度传播，然后将其他的设置为0。即针对每一个图像金字塔的每一个尺度进行正则化表示；

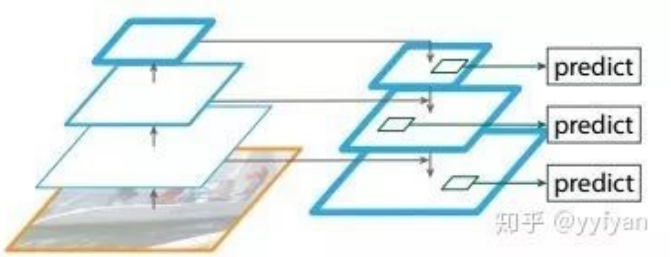
空洞卷积处理多尺度

空洞卷积本身可以控制不同大小的感受野，也即可以处理多尺度；一般空洞率设计得越大，感受野越大（但一般空洞率不能无限扩大，网格效应问题会加剧）。重点分析TridentNet~~



- 控制实验证明了**感受野大小与目标尺度呈现正相关；**
- 设计三个并行分支获取不同大小的感受野，以分别处理不同尺度的目标，感受野使用空洞卷积表征；每个分支采用Trident block构建，取代ResNet-res4中的多个原始的Block；
- 训练类似于SNIP，三个分支分别采用不同尺度的目标训练。

FPN中的多尺度处理



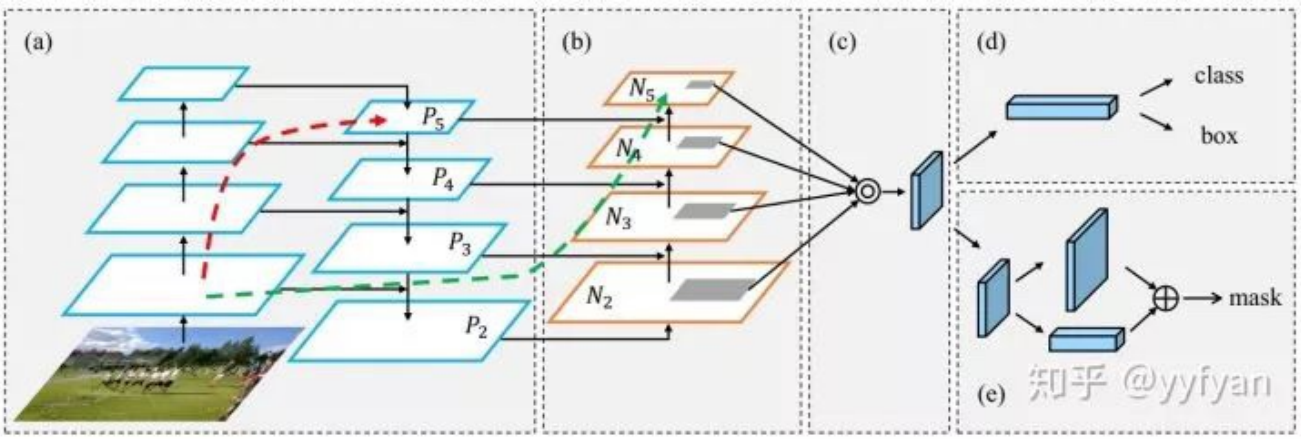
自从2016年FPN网络出来后，目前各大视觉任务的baseline基本都是以backbone-FPN。FPN以更为轻量的最近邻插值结合侧向连接实现了将高层的语义信息逐渐传播到低层的功能，使得尺度更为平滑，同时它可以看做是轻量级的decoder结构。FPN看起来很完美，但仍然有一些缺陷：

- 在上采样时使用了比较粗糙的最近邻插值，使得高层的语义信息不一定能有效传播；
- 由于经过多次下采样，最高层的感受野虽然很丰富，但可能已经丢失了小目标的语义信息，这样的传播是否还合适；
- FPN的构建只使用了backbone的4个stage的输出，其输出的多尺度信息不一定足够；
- FPN中虽然传播了强的语义信息到其他层，但对于不同尺度的表达能力仍然是不一样的，因为本身就提取了不同backbone的输出。

FPN的各种改进版本

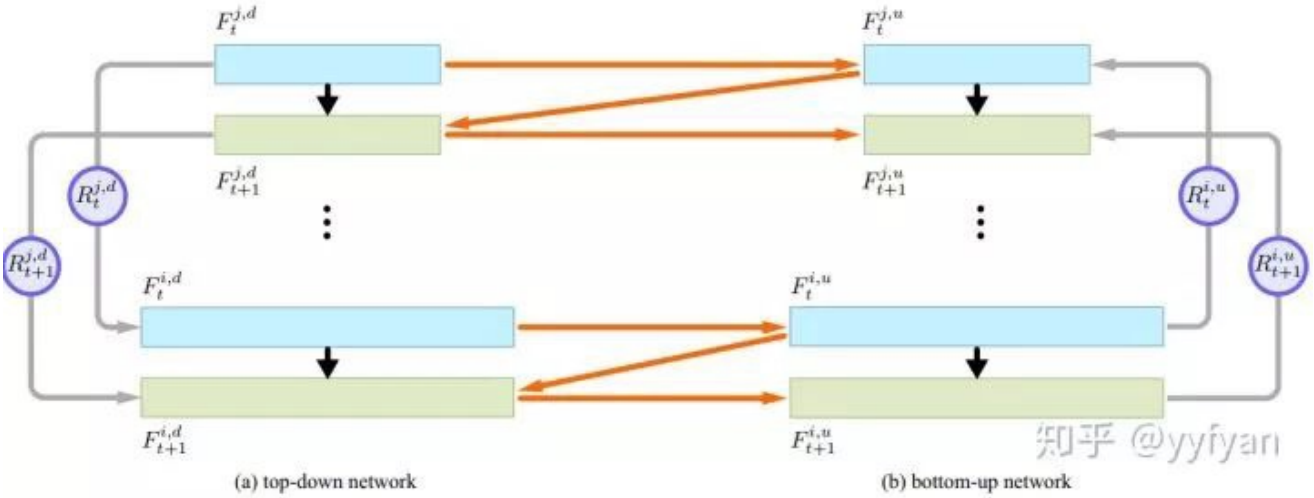
FPN这种有效的黑科技，对其进行魔改也情理之中，用几首歌的时间简要介绍介绍~~

- Shu Liu, et al. Path Aggregation Network for Instance Segmentation.//CVPR 2018



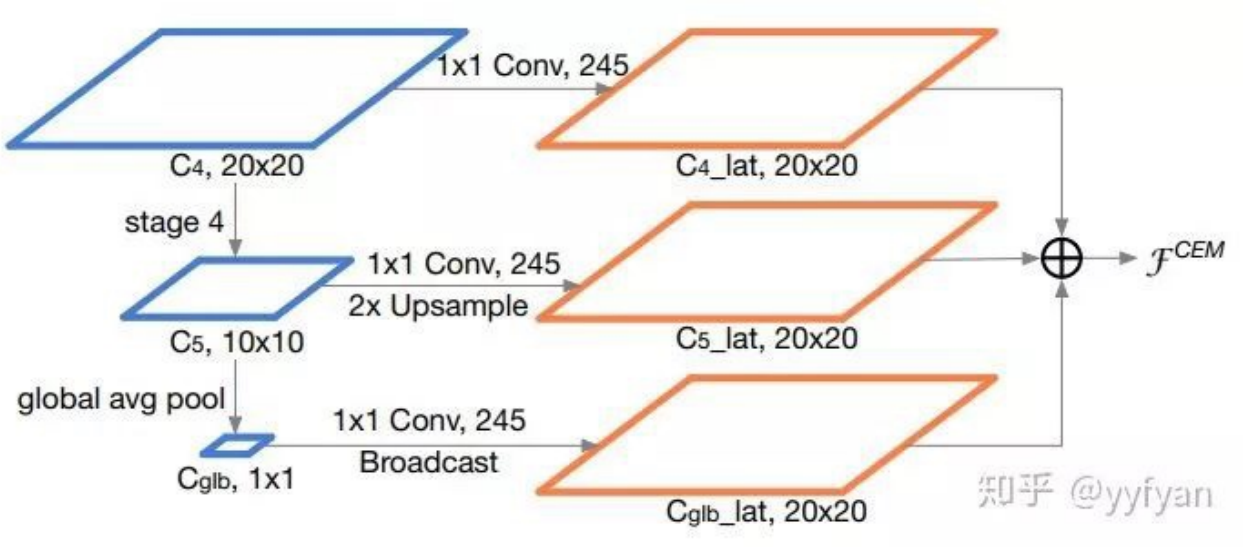
PANet在FPN的基础上加入bottom-up结构有效地传播P2的定位信息到其他层，结构与top-down结构基本一致。

- Di Lin,et al. **ZigZagNet: Fusing Top-Down and Bottom-Up Context for Object Segmentation.**//CVPR 2019



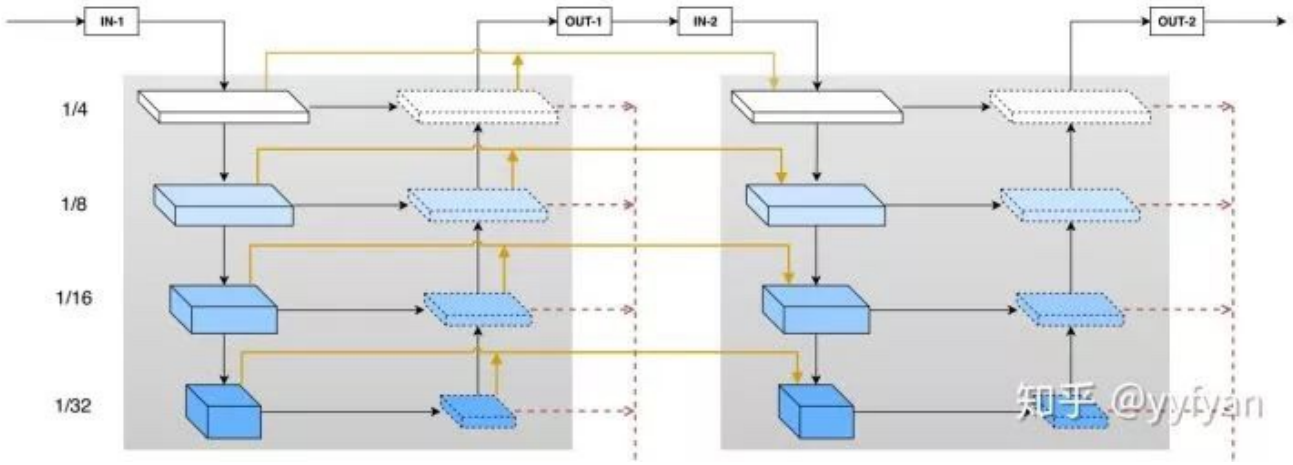
ZigZagNet在PANet上进行改进，使得top-down和bottom-up之间进行交互，同时使top-down和bottom-up的每一层之间也进行信息交互。这样就完成了双方向上的多尺度上下文信息加强。

- Zheng Qin,et al. **ThunderNet: Towards Real-time Generic Object Detection.**//CVPR 2019



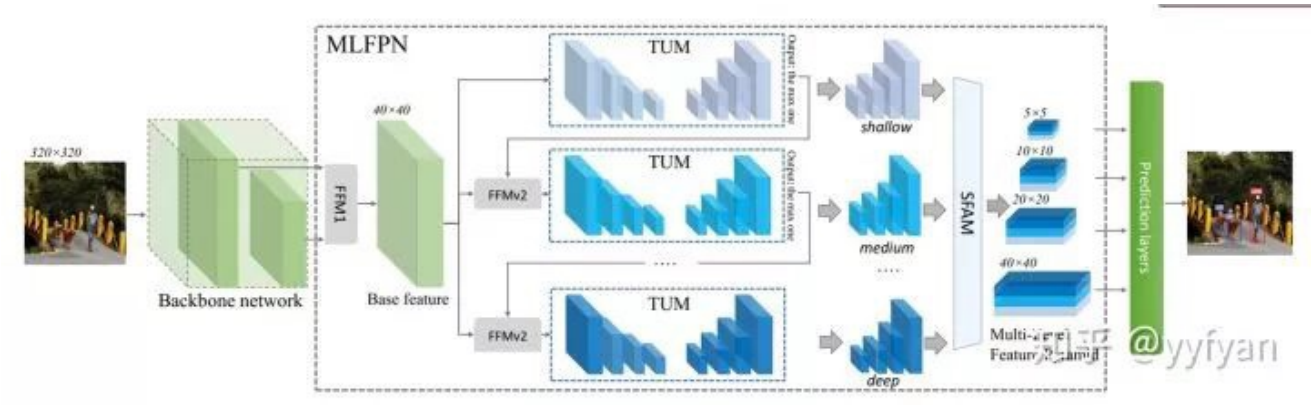
ThunderNet是做ARM上实时的目标检测算法，主要简化了FPN结构，只使用C4/C5，同时引入gpooling操作(Face++论文好多这么用，确实有效)，最终输出C4分辨率大小的累加特征。

- Wenbo Li, et al. Rethinking on Multi-Stage Networks for Human Pose Estimation.//arxiv 2019



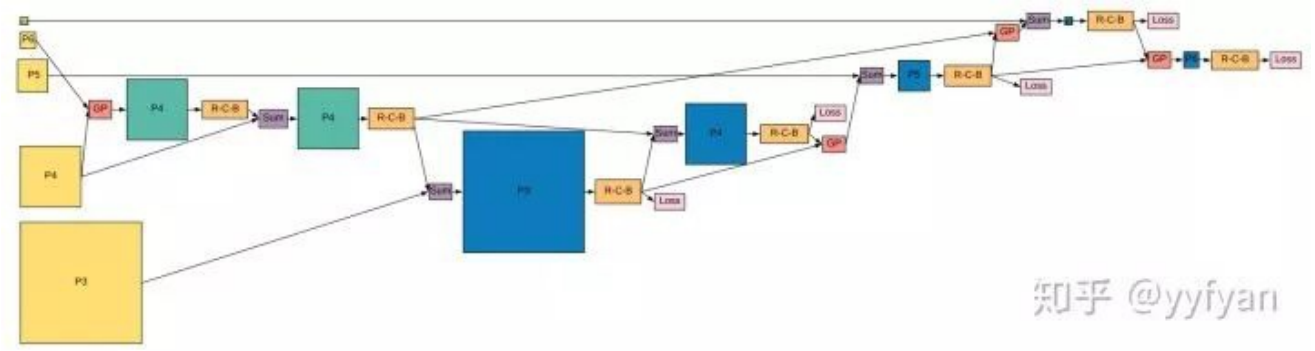
一个FPN有效，2个FPN是否有效，MSPN当然还有其他创新点，不同于FPN固定通道256，而是与backbone一致，同时还有一个特征融合模块，解决梯度消失和特征重利用。这个是做姿态估计的，其实这个结构在其他任务上也是work的。

- Qijie Zhao,et al. M2Det: A Single-Shot Object Detector based on Multi-Level Feature PyramidNetwork.//AAAI 2019



一个FPN有效，那么多个FPN是否有效，M2Det告诉了我们答案，多个FPN结构可以有效的获得多尺度特征分层。是有效果的，但这样单纯的叠加FPN真的那么友好吗！！！：)，两个已经很多了。也许别人都想得到，但去不去做就是完全不同的结局啦，我是在说我自己吗，哈哈。

- Golnaz Ghaisi, et al. NAS-FPN: Learning Scalable Feature Pyramid Architecturefor Object Detection.//CVPR2019



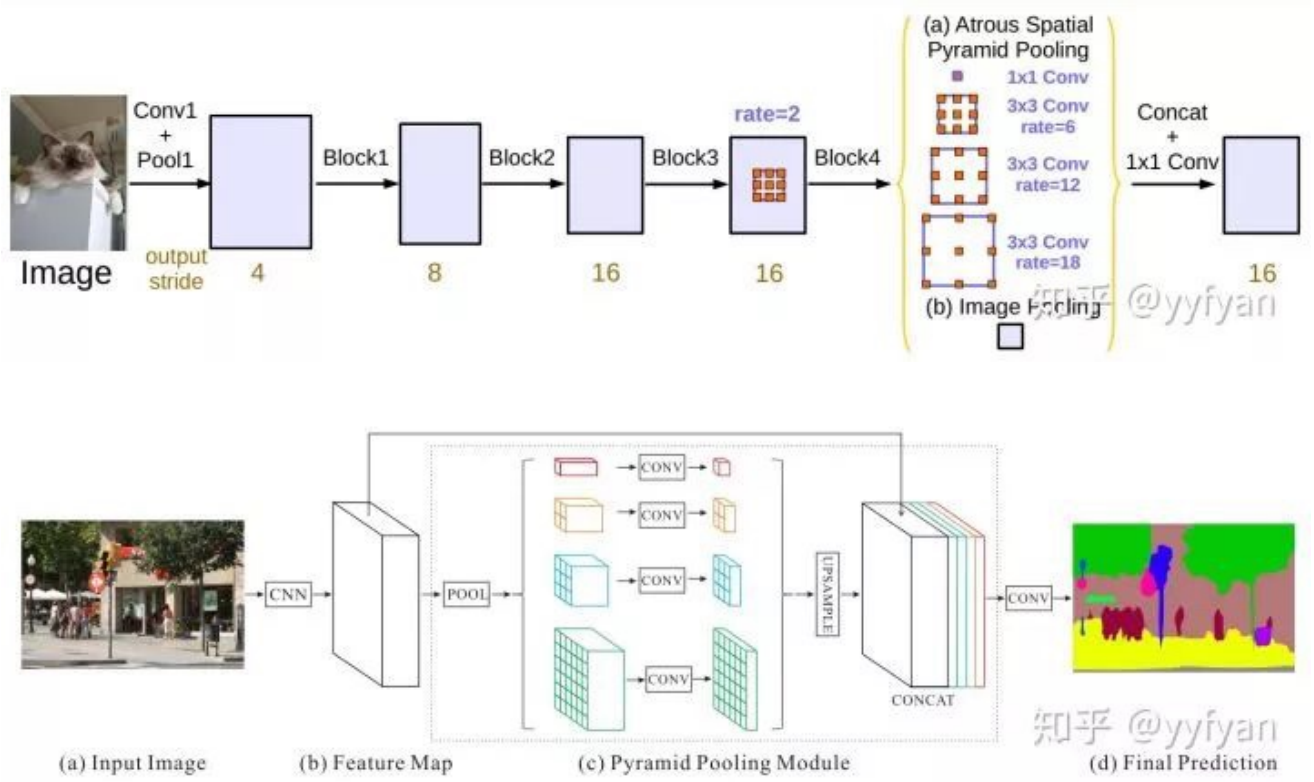
NAS最近在CNN领域刷的飞起，没错，FPN结构也已经被搜索出来了，欢迎大家实验。在目标检测上挺work的。虽然不一定会NAS，但搜索出来的网络结构可以看看的，还是能得出一些结论。

上下文模块加强多尺度信息

各种添加模块确实是CNN论文中的利器！对于语义分割，一般会将这些模块添加到backbone的最后stage，以增强预测时候的多尺度信；对于目标检测，一般会加到检测头，以增强其上下文信息。对于实例分割，可以加到mask预测分支，也可以加到其他地方。当然，现在有些做法是在FPN中的C5后加入这些模块。

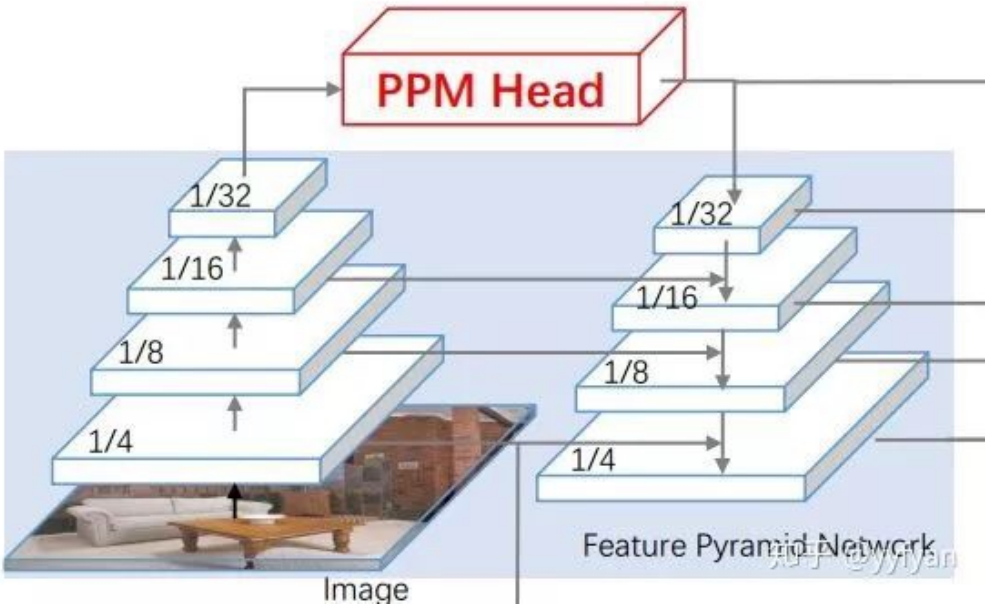
由于实在太多，这里介绍几种典型结构：

- Liang-Chieh Chen, et al. Rethinking Atrous Convolution for Semantic Image Segmentation.//arxiv 2017
- Hengshuang Zhao, et al. Pyramid Scene Parsing Network.//CVPR 2017



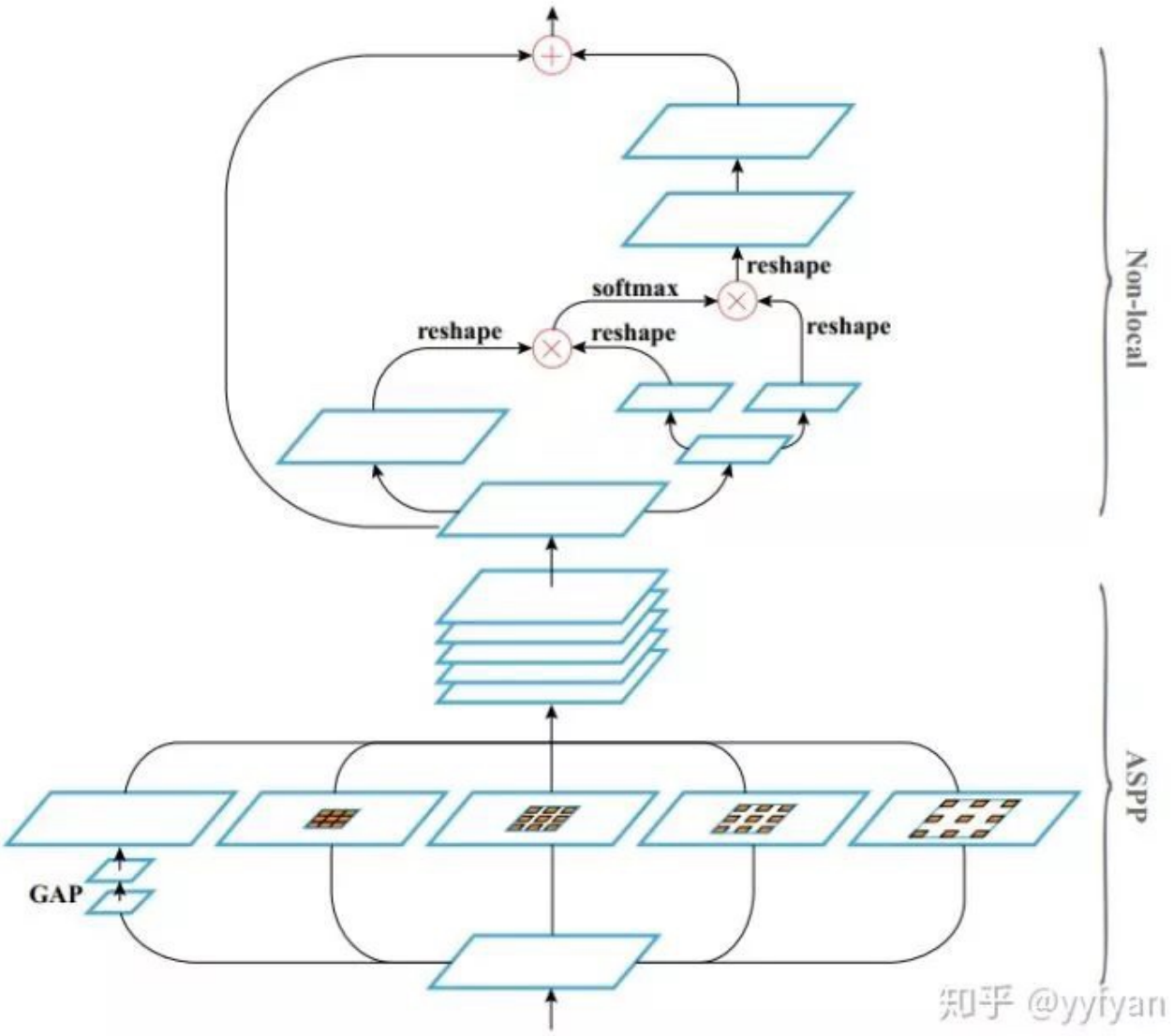
以上ASPP与PSP模型是语义分割中的经典模型，一个是使用空洞卷积，一个是不同尺度的池化。

- Tete Xiao,et al. Unified Perceptual Parsing for Scene Understanding.//ECCV 2018



在FPN上加了PPM模块，获取更为丰富的多尺度上下文信息。

- Lu Yang,et al. Parsing R-CNN for Instance-Level Human Analysis. //CVPR 2019



没错，就是ASPP+Non-local，说了要自己动手，不要说别人没有创新~~

- 这个实在太多了，怎么说都有道理，没了！！

CVPR 2019中的图像分割语义分割/实例分割

语义分割

- Hang Zhang, et al.Co-Occurrent Features in Semantic Segmentation.
- Zhi Tian, et al.Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation.
- Junjun He et al. Adaptive Pyramid Context Network for Semantic Segmentation.
- Hanchao Li, et al. DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation.

- Marin Orsic, et al. In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images.
- Fu Jun, et al. Dual Attention Network for Scene Segmentation.

实例分割

- Kai Chen, et al.Hybrid Task Cascade for Instance Segmentation.
- Di Lin, et al. ZigZagNet: Fusing Top-Down and Bottom-Up Context for Object Segmentation.
- Zhaojin Huang, et al.Mask Scoring R-CNN.
- Lu Yang, et al. Parsing R-CNN for Instance-Level Human Analysis.

github链接：https://github.com/yyfyan/read-paper-list

*延伸阅读

- **即插即用新卷积：提升CNN性能、速度翻倍**
- **重磅！商汤开源Grid R-CNN Plus：相比Grid RCNN，速度更快，精度更高**
- **CVPR2019 Oral | Relation-Shape CNN：以几何关系卷积推理点云3D形状**