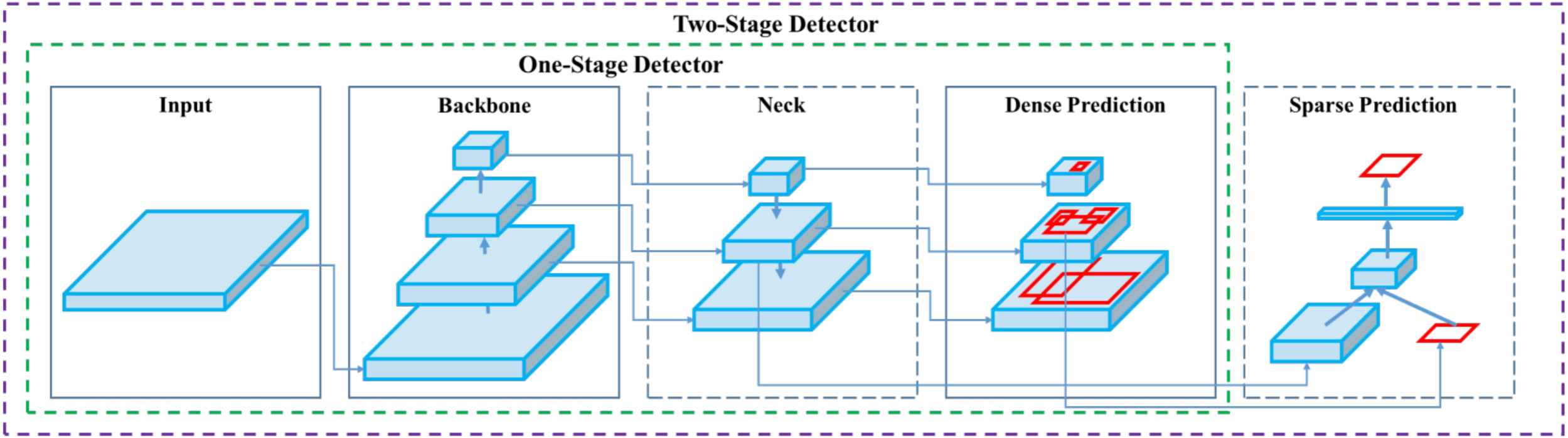
YOLOv5网络架构与组件



Input: { Image, Patches, Image Pyramid, … }

Backbone: { VGG16 [68], ResNet-50 [26], ResNeXt-101 [86], Darknet53 [63], … }

Neck: { FPN [44], PANet [49], Bi-FPN [77], … }

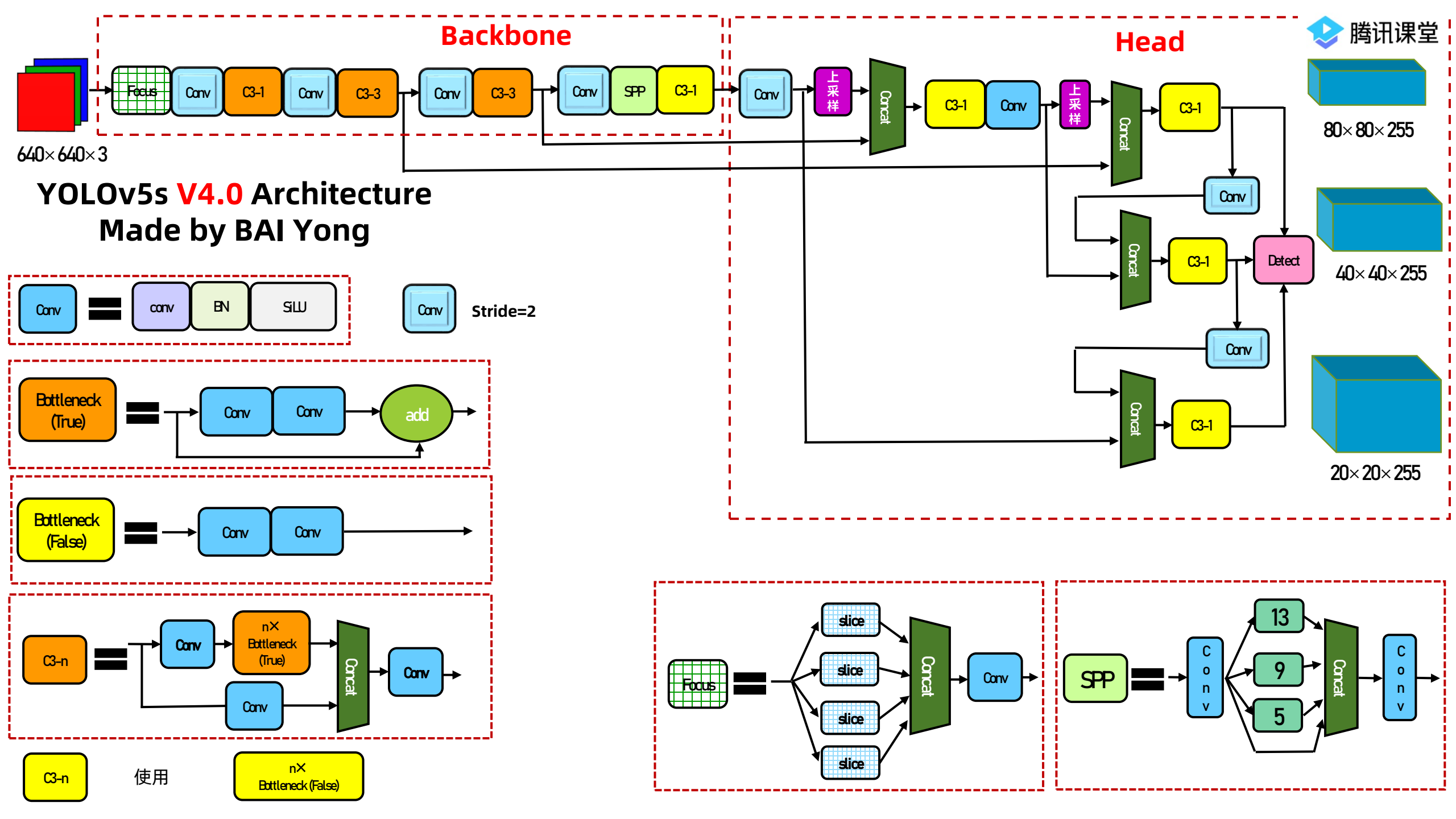
Head:

Dense Prediction: { RPN [64], YOLO [61, 62, 63], SSD [50], RetinaNet [45], FCOS [78], … }

Sparse Prediction: { Faster R-CNN [64], R-FCN [9], … }

YOLOv5 包括

- **Backbone: Focus, C3, SPP**
- **Head: PANet + Detect (YOLOv3/v4 Head)**


Backbone
Head
腾讯课堂
640×640×3
Focus
Conv
C3-1
C3-3
SPP
上采样
Concat
Detect
80×80×255
40×40×255
20×20×255
YOLOv5s V4.0 Architecture
Made by BAI Yong
conv
BN
SiLU
Stride=2
Bottleneck (True)
add
Bottleneck (False)
C3-n
n×
Bottleneck (True)
Bottleneck (False)
使用
slice
SPP
13
9
5

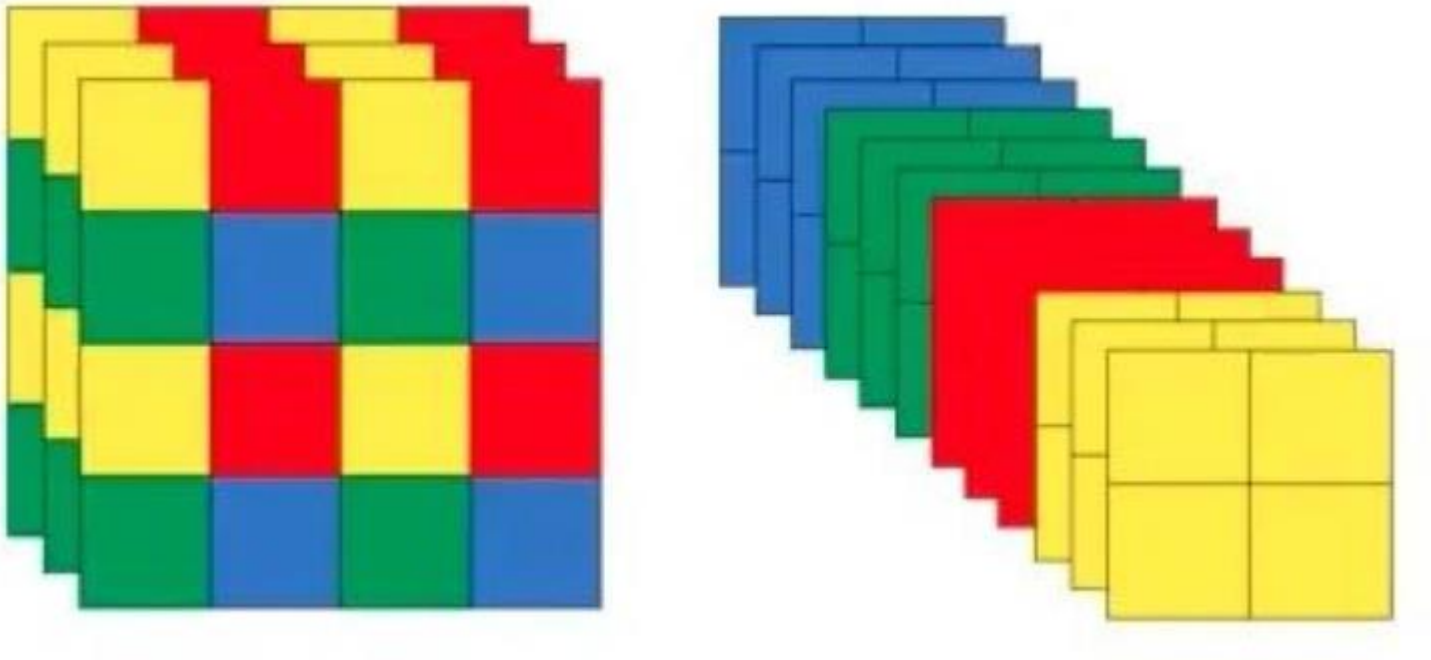
灵活配置不同复杂度的模型

应用类似EfficientNet的channel和layer控制因子

	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
depth_multiple	0.33	0.67	1.0	1.33
width_multiple	0.50	0.75	1.0	1.25
C3中的n Bottleneck(True)	1,3,3	2,6,6	3,9,9	4,12,12
C3中的n Bottleneck(Flase)	1	2	3	4
Conv卷积核数量	32,64,128,256,512	48,96,192,384,768	64,128,256,512,1024	80,160,320,640,1280

YOLOv5的四种网络结构是depth_multiple和width_multiple两个参数，来进行控制网络的**深度**和**宽度**。其中**depth_multiple**控制网络的**深度**（C3中的n），**width_multiple**控制网络的**宽度**（卷积核数量）。

Focus



把数据切分为4份，每份数据都是相当于2倍下采样得到的，然后在channel维度进行拼接，最后进行卷积操作。

Focus() module is designed for FLOPS reduction and speed increase, not mAP increase.

In YOLOv5 the author wants to reduce the cost of Conv2d computation and actually using tensor reshaping to reduce space (resolution) and increase the depth (number of channels).

Input will be transformed likes this [b, c, h, w] -> [b, c*4, h//2, w//2]

把宽度w和高度h的信息整合到c空间中

以Yolov5s的结构为例，原始640*640*3的图像输入Focus结构，采用切片操作，先变成320*320*12的特征图，再经过一次32个卷积核的卷积操作，最终变成320*320*32的特征图。

而yolov5m的Focus结构中的卷积操作使用了48个卷积核，因此Focus结构后的特征图变成320*320*48。yolov5l，yolov5x也是同样的原理

CSPNet

CSP （Cross Stage Partial Network） 跨阶段局部网络

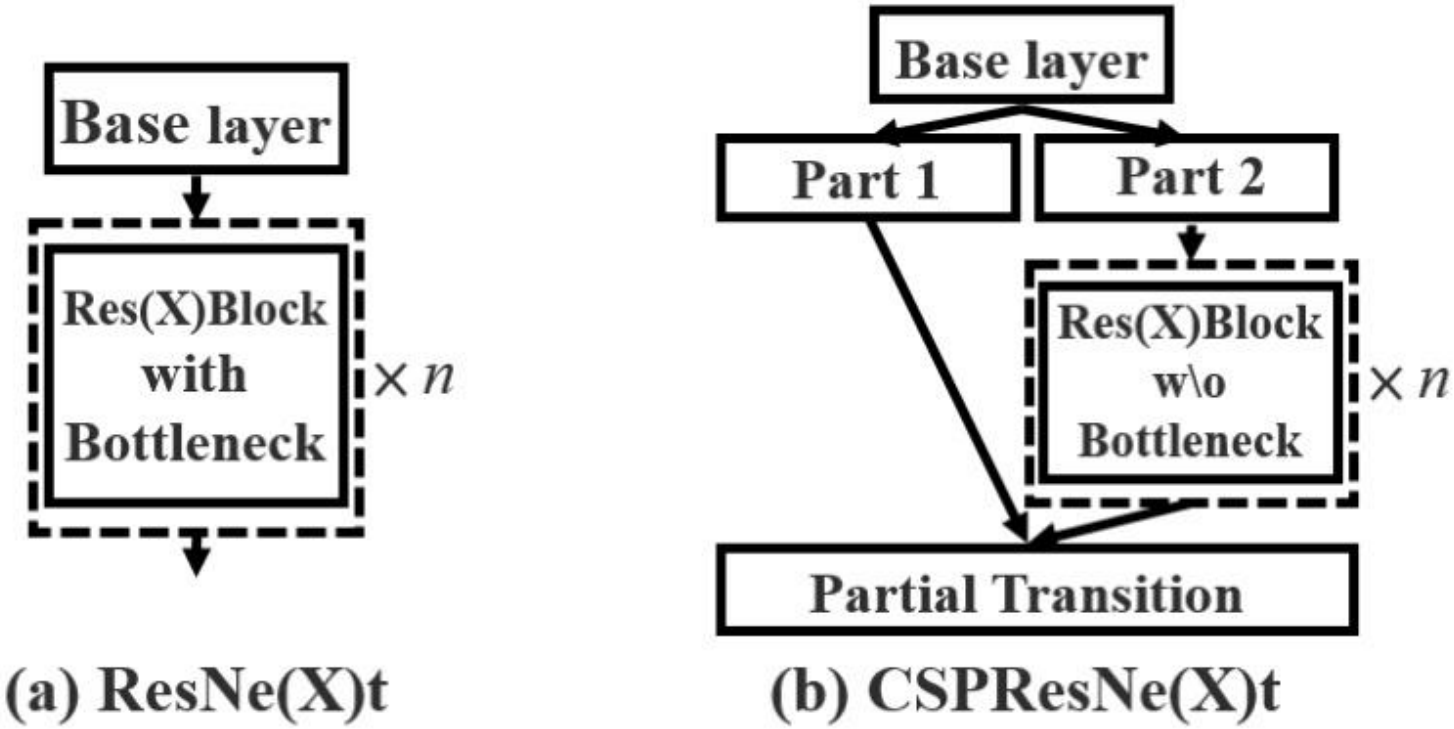


Figure 5: Applying CSPNet to ResNe(X)t.

doubts regarding the CSPNet implementation：

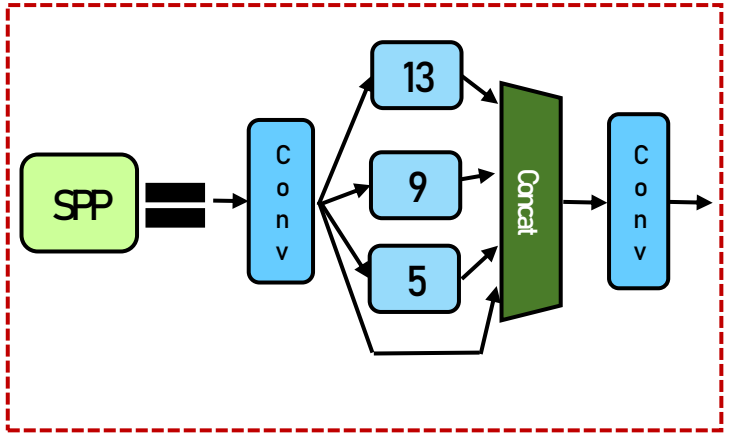
问题：As per the paper the https://openaccess.thecvf.com/content_CVPRW_2020/papers/w28/Wang_CSPNet_A_New_Backbone_That_Can_Enhance_Learning_Capability_of_CVPRW_2020_paper.pdf , the input should be split into 2 parts which should be processed through 2 branches independently. But in your implementation, Both branches are taking the same input and without any spliting.

回答：yes. that is correct. The inputs are not split, they are used in two places here, which I believe aligns with the actual CSPNet implementation.

SPP (Spatial Pyramid Pooling)

空间金字塔池化

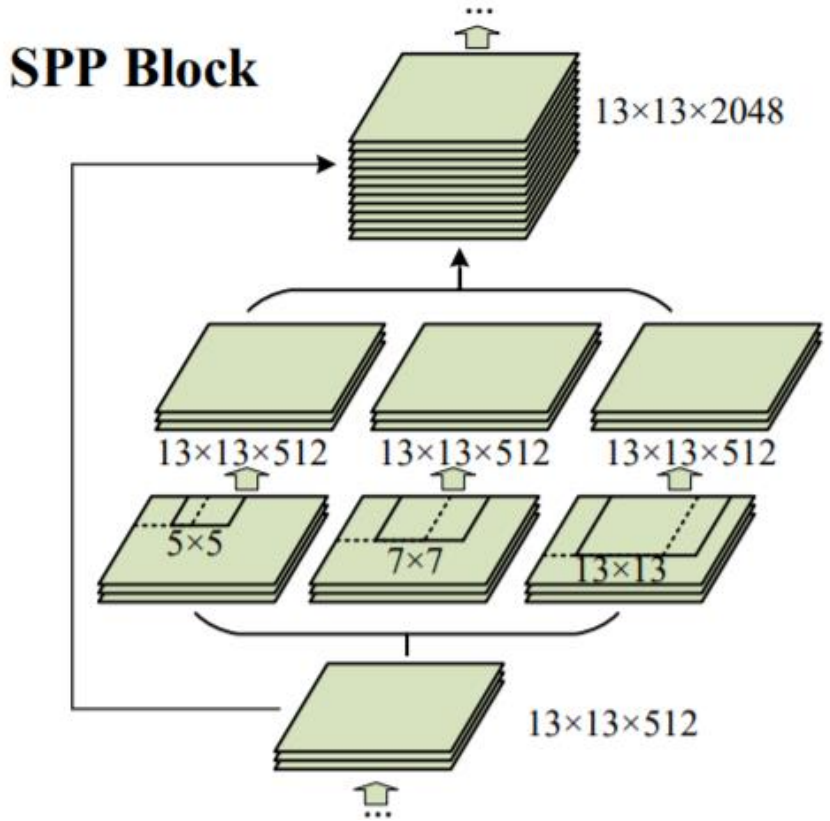
YOLOv5 SPP



Add the SPP block over the CSP, since it significantly increases the receptive field, separates out the most significant context features and causes almost no reduction of the network operation speed.

YOLOv4 SPP

416×416输入



PANet

Path-Aggregation Network 路径聚合网络

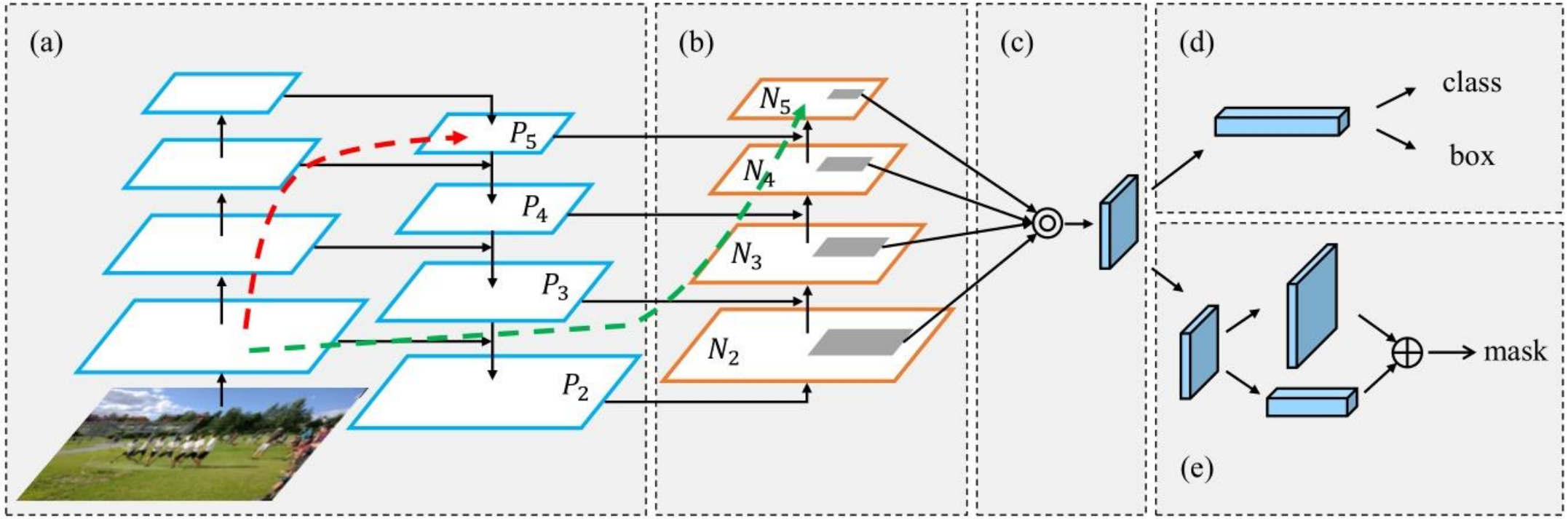


Figure 1. Illustration of our framework. (a) FPN backbone. (b) Bottom-up path augmentation. (c) Adaptive feature pooling. (d) Box branch. (e) Fully-connected fusion. Note that we omit channel dimension of feature maps in (a) and (b) for brevity.