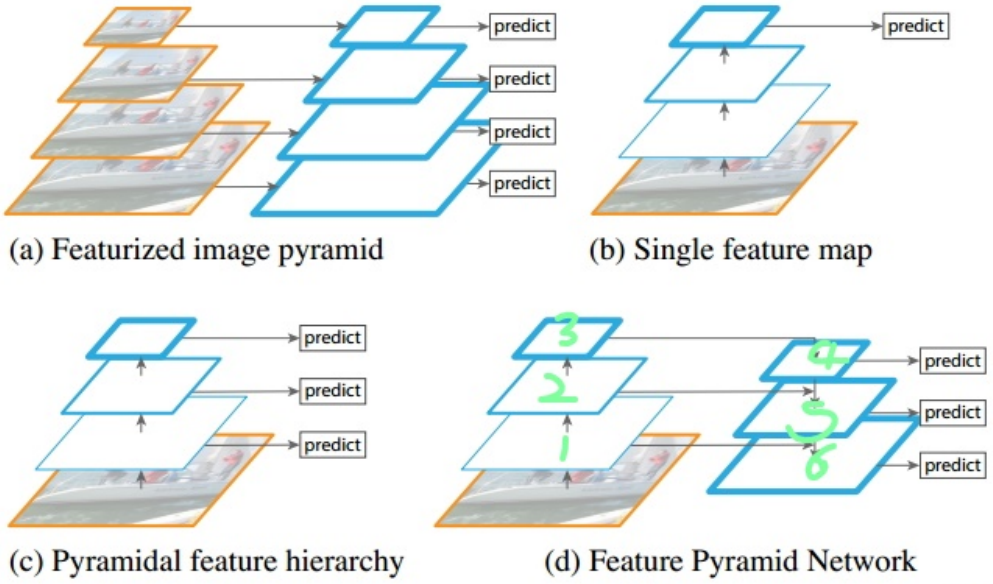

(a) Featurized image pyramid (b) Single feature map

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

Figure 1. (a) Using an image pyramid to build a feature pyramid. Features are computed on each of the image scales independently, which is slow. (b) Recent detection systems have opted to use only single scale features for faster detection. (c) An alternative is to reuse the pyramidal feature hierarchy computed by a ConvNet as if it were a featurized image pyramid. (d) Our proposed Feature Pyramid Network (FPN) is fast like (b) and (c), but more accurate. In this figure, feature maps are indicate by blue outlines and thicker outlines denote semantically stronger features.

图4 不同方案的金字塔

识别不同大小的物体是计算机视觉中的一个基本挑战，我们常用的解决方案是构造多尺度金字塔。

如上图a所示，这是一个**特征图像金字塔**，整个过程是先对原始图像构造图像金字塔，然后在图像金字塔的每一层提出不同的特征，然后进行相应的预测（BB的位置）。这种方法的**缺点是计算量大，需要大量的内存；优点是可以获得较好的检测精度**。它通常会成为整个算法的性能瓶颈，由于这些原因，当前很少使用这种算法。

如上图b所示，这是一种改进的思路，学者们发现我们可以利用卷积网络本身的特性，即对原始图像进行卷积和池化操作，通过这种操作我们可以获得不同尺寸的feature map，这样其实就类似于在图像的特征空间中构造金字塔。实验表明，浅层的网络更关注于细节信息，高层的网络更关注于语义信息，而高层的

语义信息能够帮助我们准确的检测出目标，因此我们可以利用最后一个卷积层上的feature map来进行预测。这种方法存在于大多数深度网络中，比如VGG、ResNet、Inception，它们都是利用深度网络的最后一层特征来进行分类。**这种方法的优点是速度快、需要内存少。它的缺点是我们仅仅关注深层网络中最后一层的特征，却忽略了其它层的特征，但是细节信息可以在一定程度上提升检测的精度。**

因此有了图c所示的架构，它的设计思想就是**同时利用低层特征和高层特征，分别在不同的层同时进行预测，**这是因为我的一幅图像中可能具有多个不同大小的目标，区分不同的目标可能需要不同的特征，对于简单的目标我们仅仅需要浅层的特征就可以检测到它，对于复杂的目标我们就需要利用复杂的特征来检测它。整个过程就是首先在原始图像上面进行深度卷积，然后分别在不同的特征层上面进行预测。**它的优点是在不同的层上面输出对应的目标，不需要经过所有的层才输出对应的目标（即对于有些目标来说，不需要进行多余的前向操作），这样可以在一定程度上对网络进行加速操作，同时可以提高算法的检测性能。它的缺点是获得的特征不鲁棒，都是一些弱特征（因为很多的特征都是从较浅的层获得的）。**

讲了这么多终于轮到我们的FPN啦，它的架构如图d所示，整个过程如下所示，首先我们在输入的图像上进行深度卷积，然后对Layer2上面的特征进行降维操作（即添加一层1x1的卷积层），对Layer4上面的特征就行上采样操作，使得它们具有相应的尺寸，然后对处理后的Layer2和处理后的Layer4执行加法操作（对应元素相加），将获得的结果输入到Layer5中去。其背后的思路是为了获得一个强语义信息，这样可以提高检测性能。认真的你可能观察到了，这次我们使用了更深的层来构造特征金字塔，这样做是为了使用更加鲁棒的信息；除此之外，我们将处理过的低层特征和处理过的高层特征进行累加，这样做的目的是因为低层特征可以提供更加准确的位置信息，而多次的降采样和上采样操作使得深层网络的定位信息存在误差，因此我们将其结合其起来使用，这样我们就构建了一个更深的特征金字塔，融合了多层特征信息，并在不同的特征进行输出。这就是上图的详细解释。