An Overview of Deep Learning Based Object

Detection Techniques

Abstract：最近几年，我们见证了深度学习领域的快速发展。随着对深度学习领域的偏爱，目标检测的任务变得越来越有趣和越来越有挑战性。目标检测是用来再给出一张对象的情况下对整张图像中存在的所有对象进行检测。基于深度学习的目标检测技术在直接从数据中直接学习对象特征的方面表现出了有效性。这篇论文主要是聚焦于在各种各样的基于目标检测技术的最新的深度学习技术提供一个概述。工作聚焦于当不同的目标检测技术面临机会和障碍的反应提供一个广泛的比较。这篇论文包括未来这些领域的黄金研究范围。

I. INTRODUCTION

计算机视觉作为一种感知和掌握知识的工具。在计算机视觉算法中，有三项基础工作包括分类，检测和局部化。这三种基础的工作在计算机视觉领域提供了机会和挑战。他们已经开始寻找研究的关注而且这些领域中的工作已经有了巨大的增长。图像分类致力于给一个图像一个标签。对象检测的过程在计算机视觉领域是另一个重要的工作，智在对一张图像中的存在的所有对象进行检测。例如，无人驾驶汽车（又称为自主驾驶汽车）结合了软件和传感器，不仅可以检测其他车辆的存在，也可以对道路上的树，人，动物和其他的车辆进行检测。

计算机视觉中的第三项主要的任务是对象定位。对象定位任务与对象检测技术是很相似的。对象检测任务处理对象类别的数量是变化的，而对象定位是解决在对象类别数量固定的情况下的分类问题。对象定位认为被考虑为一个在计算机视觉领域里面的超级图像分类任务。有了这个，我们不仅需要知道一个对象在给出的图像中的类别也需要只要这个对象究竟在图像中的哪里。目标检测被看为在大多数计算机视觉系统中的第一个过程。他有几个应用场景，而且他们中的一些包括安全系统，人机交互，机器人技术，在工业制造中的产品检测，人数统计等。

从过去十年开始，机器学习领域开始有目的的前进，结果，研究开始更多地聚焦于技术方面尤其是在物体检测的方面。基于目标检测的机器学习由两阶段组成，训练阶段和测试阶段。随着图像数量的增加使得计算机可以正确的学习物体的分类和并且有助于鉴别属于不同于类别的对象。测试阶段涉及在测试任务中再给出不同的输入或者测试用例的情况下，机器是否可以给出正确的响应。通过将图像输入到计算机中，他将通过一个名为特征提取的过程去尝试对图片内的对象特征的每个比特进行学习。

人们以不同的方式来学习和回应。人们可以通过经验来进行学习，但是对于机器来说，我们需要对机器进行合适的训练。因此，这些任务是通过训练一个分类器使得其能够提取对象外观上的微小差异。一组被称为候选人和提议的区域将被作为分类器。 这些区域提案被认为是重要的和不正确的建议可能会损害系统在工作中的性能。因此在机器学习的子集中出现了深度学习，深度学习聚焦于以更有效的方式来处理这些问题。

深度学习只是简单的是一种算法，可以更好地完成预测任务。随着近几年基于对象检测的深度学习得到了巨大的发展，因此我们决定为各种最新的基于深度学习的对象检测技术提供一个survey。这篇论文聚焦于为现实世界中具有高价值的对象检测的应用提供一个概述。他还提供了一个表格来对各种基于对象检测的深度学习网络进行比较及在遥感影像中的目标检测。

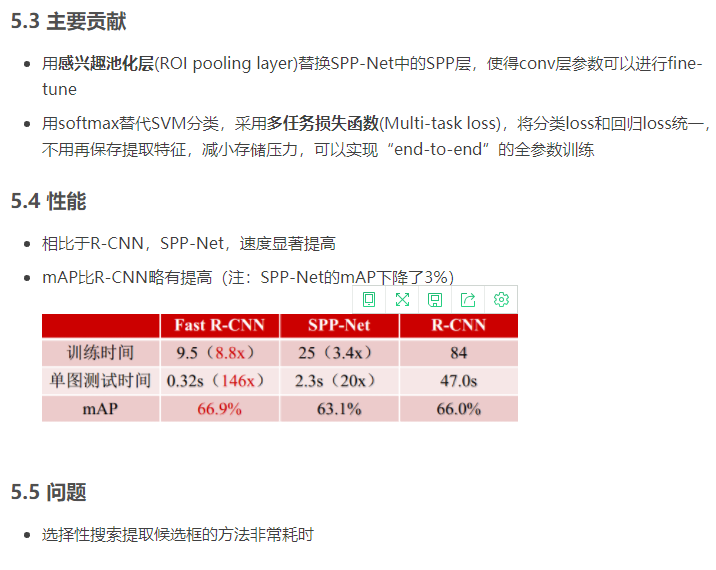
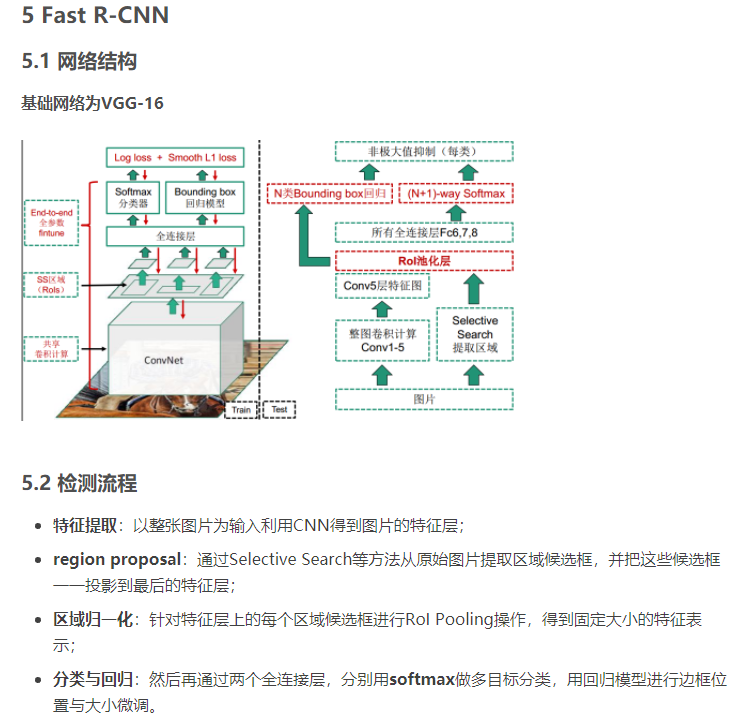
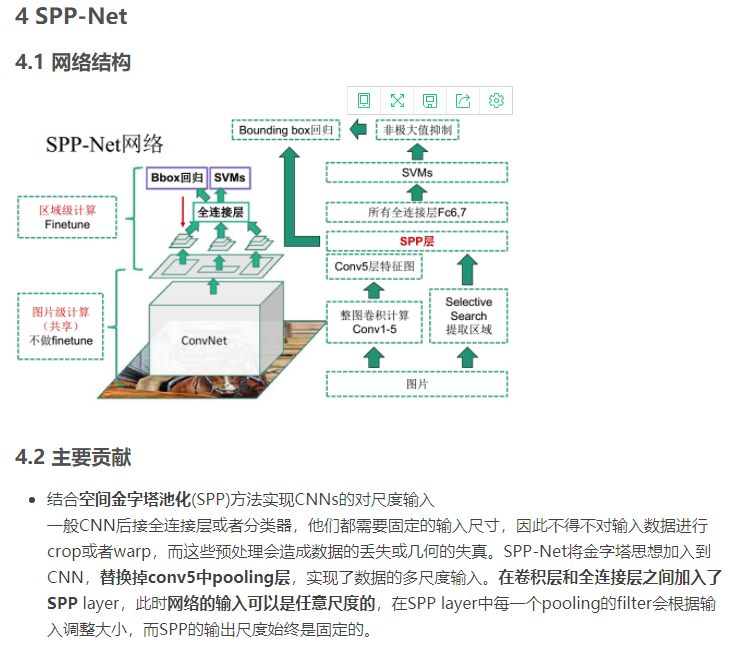
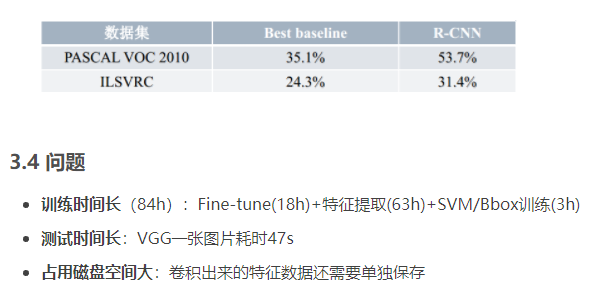
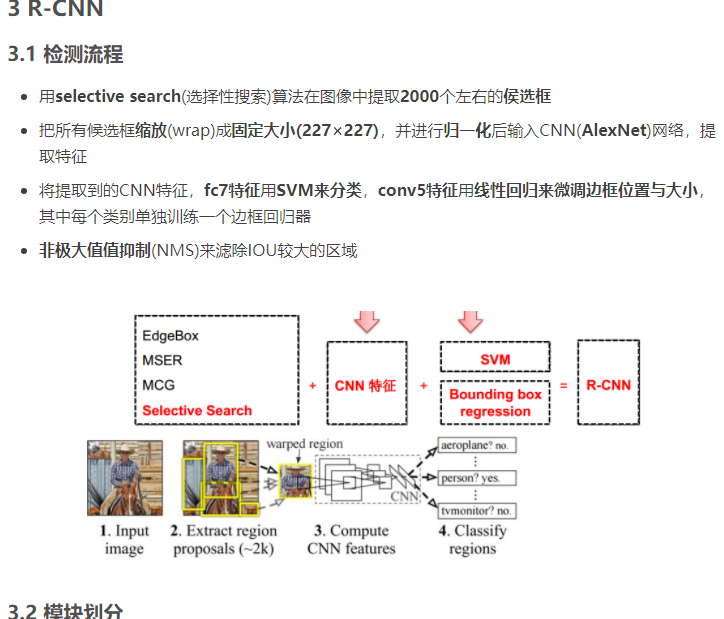
II. TECHNIQUES EMPLOYED

一个非常流行和简单的对象检测技术被称为滑动窗口。就像滑动窗口这个名字所表示的那样，滑动窗口就是一个小的窗口从图片的左上角开始滑动，穿过整张图片。这是一个简单的技术而且滑动窗口的大小通常远远小于实际图片的大小。在对图像的所有内容进行过检查之后，窗口的大小通常会增加一些固定的值。这个过程将会一直持续直到新的窗口达到了停止条件。在滑动窗口技术中，区域的数量在考虑中有非常高的地位。滑动窗口技术在计算上很昂贵而且给出的边界盒子不是很准确。

为了克服滑动窗口技术的约束， R Girshick提出了一种使用R-CNN的对象检测方法(Regional Convolutional Neural Networks)（区域卷积神经网络）。它着重于选择那些可以再CNN上运行的区域。这里使用的方法是区域提议的选择性搜索算法来基于相似性的目的来对整张图片进行划分，并将其划分为几组。选择性搜索的算法比传统的方法例如滑动窗口法的工作效果更好。即使RCNN给出较少数量的候选人地区提案，这种方法的缺点是他需要一个昂贵的训练阶段和该技术的检测速度比较慢。为了解决这些问题，P Sermanet等为对象识别，定位和检测介绍了一个有趣的方法，这种策略成为OverFeat。（抽空可以去寻找OverFeat的论文进行查看，简单来说通常CNN是使用一个CNN来完成一个任务，但是OverFeat实现了使用一个CNN模型就实现了三种任务）

以下为对部分区域神经网络进行概括：





其中对SPP-Net进行详细介绍的网址为https://blog.csdn.net/qq\_31442743/article/details/80770619

对Fast R-Cnn进行介绍的网址为https://blog.csdn.net/u011501388/article/details/81031780

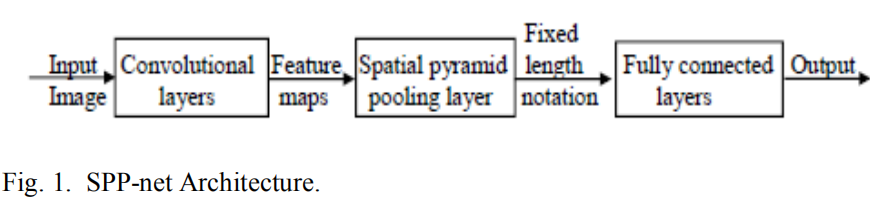
Faster R-cnn去寻找论文原著进行查看 CSDN上的讲解不是很清楚。

对区域卷积神经网络进行概括介绍的网址为https://blog.csdn.net/u013187057/article/details/84023468

ILSVRC2013定位方法比赛宣告了这种方法（OverFeat）是可以兴旺对象检测的策略之一。OverFeat的主要思想是训练一个CNN来同时完成计算机视觉中的三种基础任务。（对象的识别，定位和检测）。OverFeat被视为一种通过收集预测任务的边界框的方法来处理定位和检测任务的独特的方法。边界框只是一种简单的长方形被用来标记在图像中的确定的物体。绝大多数这些方法使用了滑动窗口的概念因此需要为每个窗口计算输入图像。这种低效率（级为每个滑动窗口计算输入图像）可以通过使用ConvNet来进行解决，因为它们共享重叠区域的计算。这导致了新的对象检测技术的到来，新技术称为Spatial Pyramid Pooling or SPP.（即上图中所提及的SPP-Net）

Spatial Pyramid Pooling为Kaiming He等人提出，这是一种不同的池化策略，整个的网络结构被称为SPP-Net。特征提取过程主要是从图像中的每个区域提议中提取一个大的特征集合。这是另一个将spatial pyramid pooling 引入到CNN结构中的主要原因。管理不同尺度，大小和纵横比的图像的能力使得SPP成为一种自适应技术。最开始的步骤是使用一定数量的卷积层来生成输入图像的特征图像，特征图像在SPP-Net上只生成一次。特征图像被允许通过输出n个多维向量的SPP层。这里n是在最后的卷积层中的filters的数量。图1描述了patial

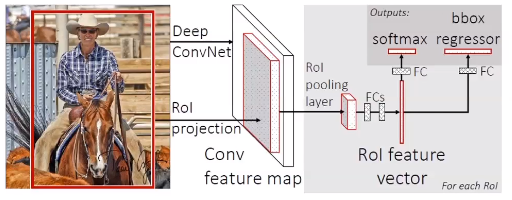
pyramid pooling的方法。我们发现SPP的速度比RCNN的速度快，但是它与其他的深度神经网络相比精度会下降。因此需要有一个可以概括的对象检测技术。



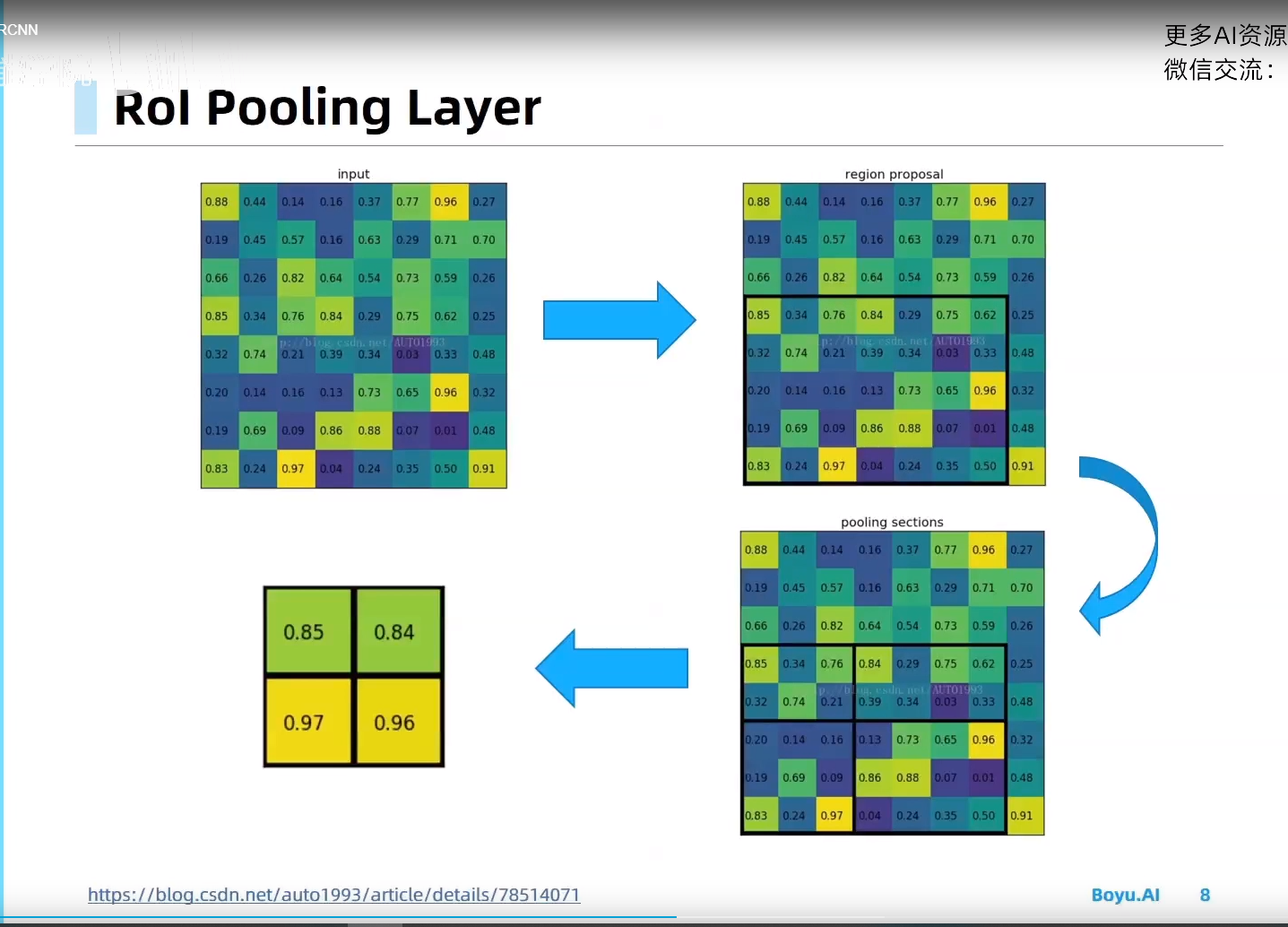
Gidaris S and Komodakis N因此提出了名为MultiRegion CNN 或者简单地称为 MRCNN的用于对象检测的通用方法。中心思想是尽可能多的从区域提议中或者周边提取特征。所有的特征简单的聚合到一起。在考虑各种区域中的对象提案包括一般区域，中心区域，边界区域和上下文区域等。它也整合了语义分割功能用于对象检测，但是它并不适合所有的实时应用。因此，Donggeun Yoo等在对象检测中引入了名为AttentionNet的迭代过程。方针是从图像的边界开始缩小边界框来进行对象定位。但是实验表明AttentionNet适合扩展类别规模的事实，结果也可以降低召回率。（MultiRegion CNN为在实现OverFeat的过程中的所需要使用的模型）

Girshick提出了一种新方法，叫做Fast R-Cnn来克服R-CNN和SPP-Net的限制。事实证明这是fast R-CNN的一个重要的指标。从所有的候选人区域提案的特征图像中学习固定长度的特征向量。这是在一个名为Rol的池化层的帮助下完成的，这就是fast R-CNN的一个诱人的优势。最初，fast R-CNN允许将整张图像输入到ConvNet中来创建感兴趣的区域。它采用一个阶段来从提案区域中去提取特征的方法，并对特征进行分类最后返回边界框和输出类别标签。 与SPP-Net和R-CNN相比，fast R-CNN给出了高质量的对象检测.

Fast R-CNN对输入的数据并不做什么要求，可以输入任意大小的图像，将图像中的多个感兴趣区域输入到ROI（multiple regions of interest）中，每个ROI得到大小固定的特征映射，然后将特征映射输入到全连接网络中。特征映射经过ROI层的处理之后可以得到一个固定大小的特征图像，如SPP方法一样。



Fast R-CNN的优势：1.检测效果更好 2.训练是单阶段的，使用多任务损失函数3.训练时可以更新整个网络层4.特征缓存不需要磁盘存储。允许对CNN中feature map进行重用。

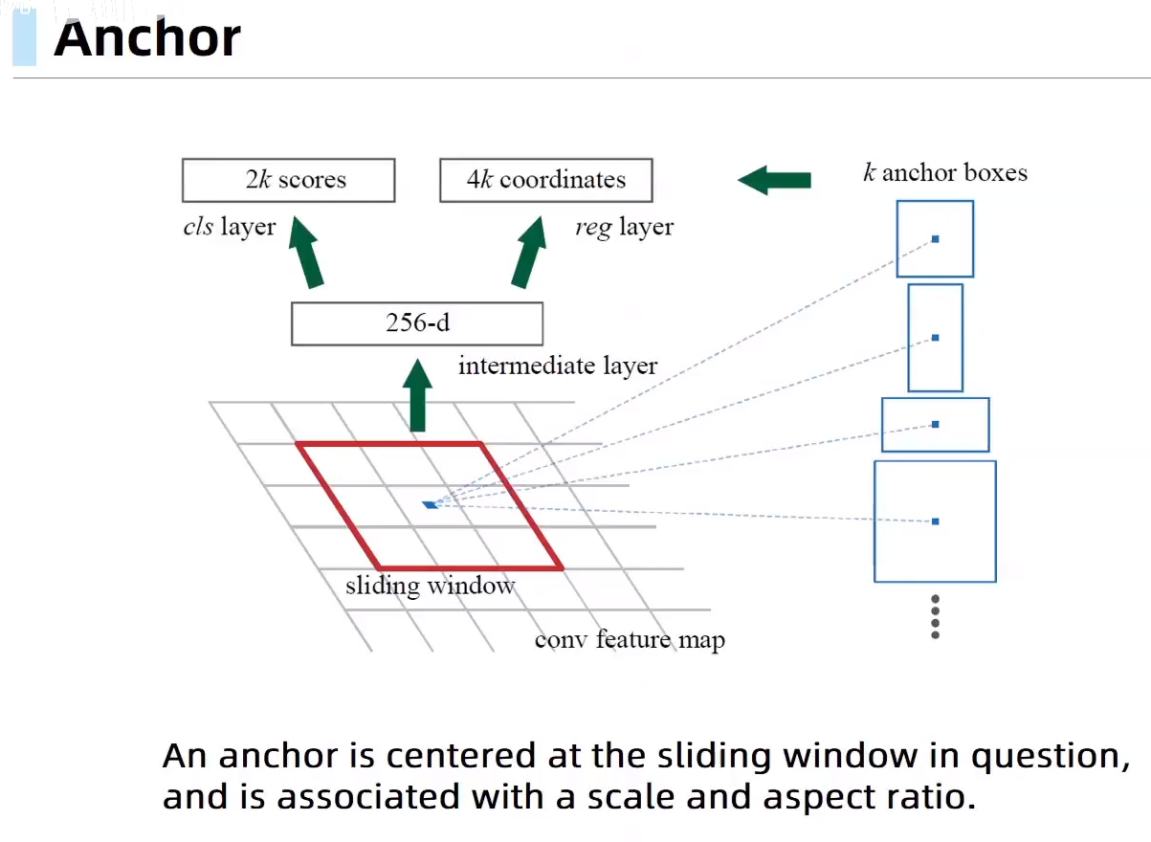


Fast R-CNN的一个缺点是选择性搜索算法的慢速度的区域提案导致了过高的时间消耗。为了推翻这种情况Ren S 等提出了一种名为faster R-CNN的新方法。Faster R-CNN的框架提供了更高的精度和更有效的方法（即RPN结构）来通过区域提案网络来生成区域提案。输入的图像经过ConvNet来获得特征图像然后依靠区域提案网络来生成对象提案。对了进行分类。这些候选人提案将会移至全连接层。Faster R-CNN的优点是它的速度但是却面临很多的挑战例如边框不正确。

Faster R-CNN抛弃了R-CNN的SS方法，使用region proposal network（RPN）方法。RPN先经过3\*3的卷积运算，一路用来判断候选框是前景还是背景，先reshape成1\*1的向量，然后进行前景还是背景的运算，进行判断之后，reshape恢复为二维的feature map。另一路用来确定候选框的位置通过候选框回归实现。在两个运算结束后，挑选出前景候选框，利用前景候选框位置得出我们感兴趣的proposal。然后将proposal输入到ROI中。



上图中的过程即为RPN

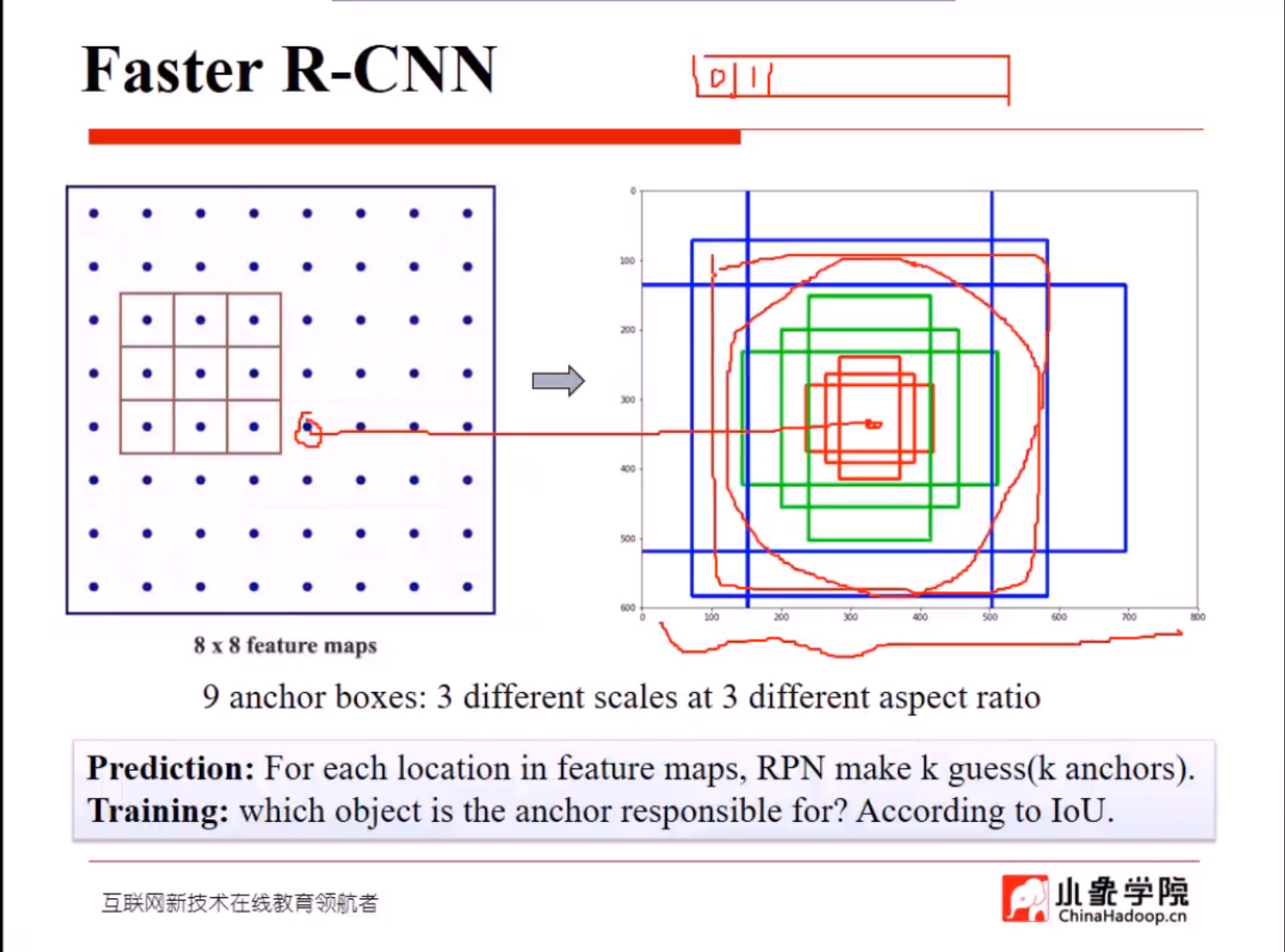


acnchor锚点是 感受域 或者 滑动窗口（原文 3\* 3）其实就可以想象成 的3\* 3卷积核，这个窗口滑动过程中每一个中心就是一个锚点.ROI。K anchor boxes的值为锚点坐标x,y和所选框的大小w,h。

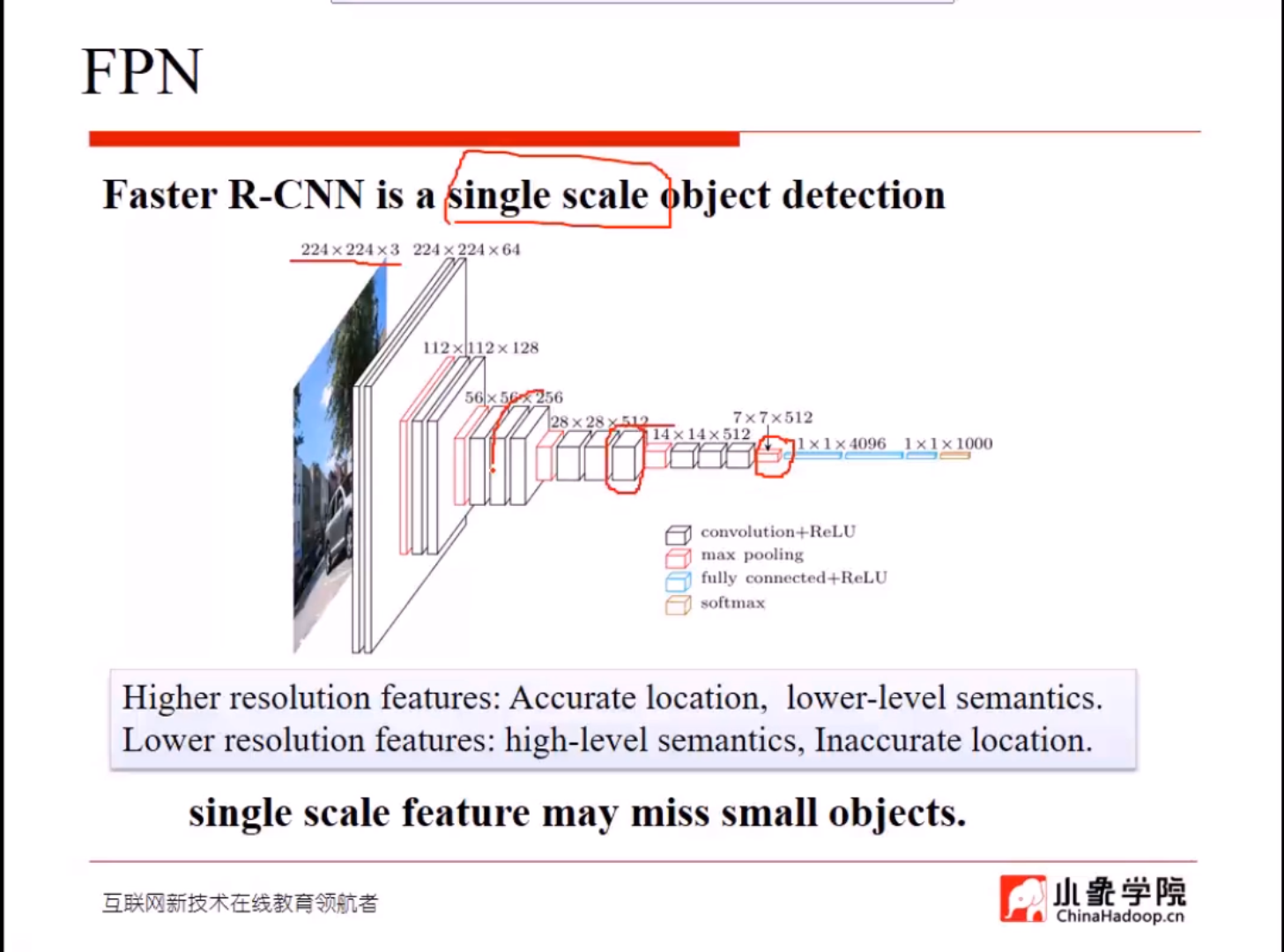
顺便说一下由于只是卷积和池化所以所有锚点是可以按比例对应到原图的，所以可以与真实的物体的框（ground true box）进行比较（即后面要提到的ROI也就是我们的box和ground true box的重合率）

av66083121（讲解fast R-CNN与faster R-CNN）

讲解的更好anchor就是对图像中的每个点选用三组3个的大小相同比例不同不同的框（如下图所示），使用与目标物体大小相似的框来对物体预测。使用这种anchor来对物体进行预测就可以解决当物体的大小发生变化时使用单一类型的anchor就会导致可能无法进行有效预测。（即大的anchor预测大的物体，小的anchor预测小的物体）即使用IOU方法来选择合适的anchor。（根据物体与anchor的交集的大小（比值）来确定到底选择哪一个anchor）

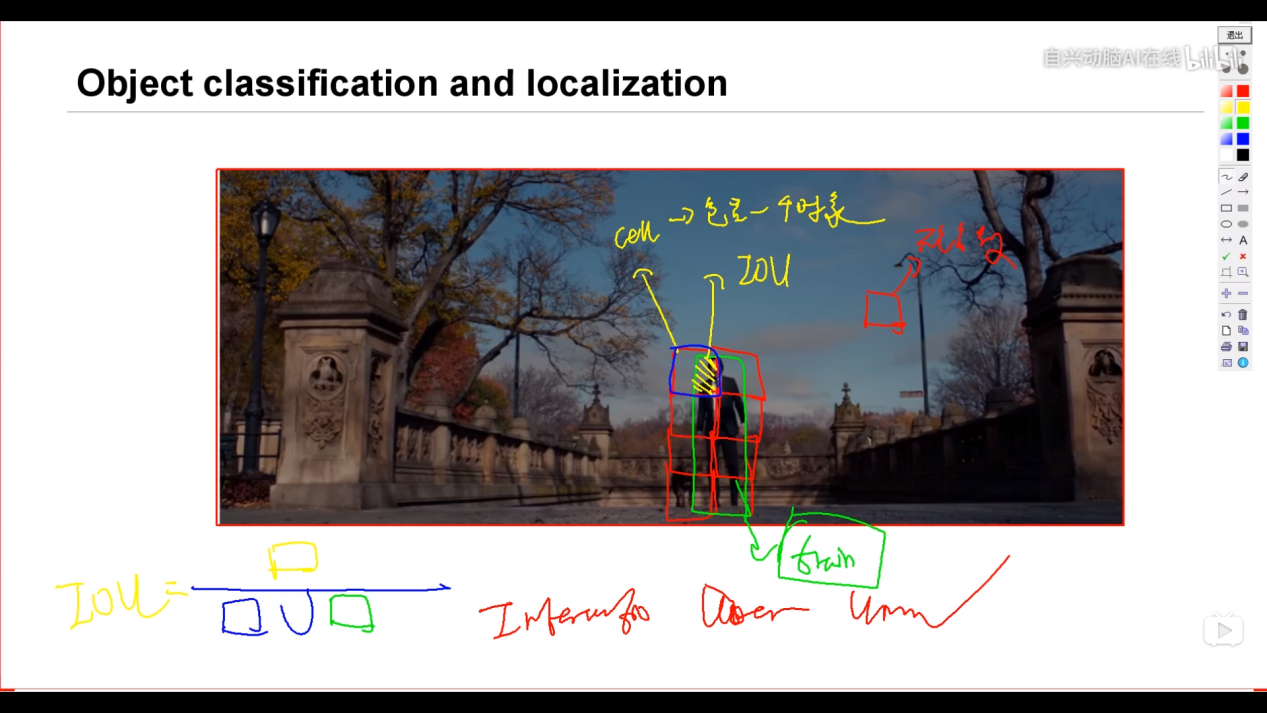


Faster R-CNN是一种单尺度的对象检测，如下图所示，Faster R-CNN使用VGG-16对图像特征进行提取。下图使用VGG-16对图像进行特征提取是单向的。最后只使用了7\*7\*512来进行预测，而在VGG-16过程中的数据信息如56\*56\*256或者28\*28\*512并没有使用，而原图是224\*224\*3，会导致损失太多的数据，如果是预测大的数据还好，但是预测小的数据的话可能会无能为力。FPN即对Faster

的这个缺点进行了改进，即使用VGG-16的浅层特征来进行判断。

为了克服计算费用的昂贵和提供精度更高的边界框，Redmon等提出了YOLO (You Only Look Once) 的方法.YOLO的主要思想是将图像分为很多数量的网格。分类和定位的算法是依靠这些网格进行的。类标签或者网格中当前存在的每个对象的标签是在物体的中心。对于每个类实例，YOLO预测固定数量的边界框和分数或者属于多个不同类别的对象的可能性。YOLO的速度很快提供较少的背景错误。但是YOLO不能使用一个网格来检测多个对象，但是可以多次检测一个对象。下面进入YOLO的损失率。

Tips:YOLO算法是将图片分为不同数量的网格，分成的网格数量不固定，可能是13\*13或者19\*19 如果需要的粒度大的话则选用13\*13（精准度会下降），如果所选用的粒度小的话则选用19\*19的。在YOLO中检测对象主要就是对网格中的IOU值进行计算，IOU值得大小就是如下图所示的代表对象的黄色方框的大小与蓝色（代表网格的大小）和绿色（代表训练集中对象的大小）的比值。如果这个比值大于某个预定的阈值，则代表这个网格总存在对象，否者则代表这个网格中不存在对象。将图片中的所有YOLO的网格进行计算，即可得到在图像中识别出物体所在位置的框即下图中包括人的那个大红色框所示。



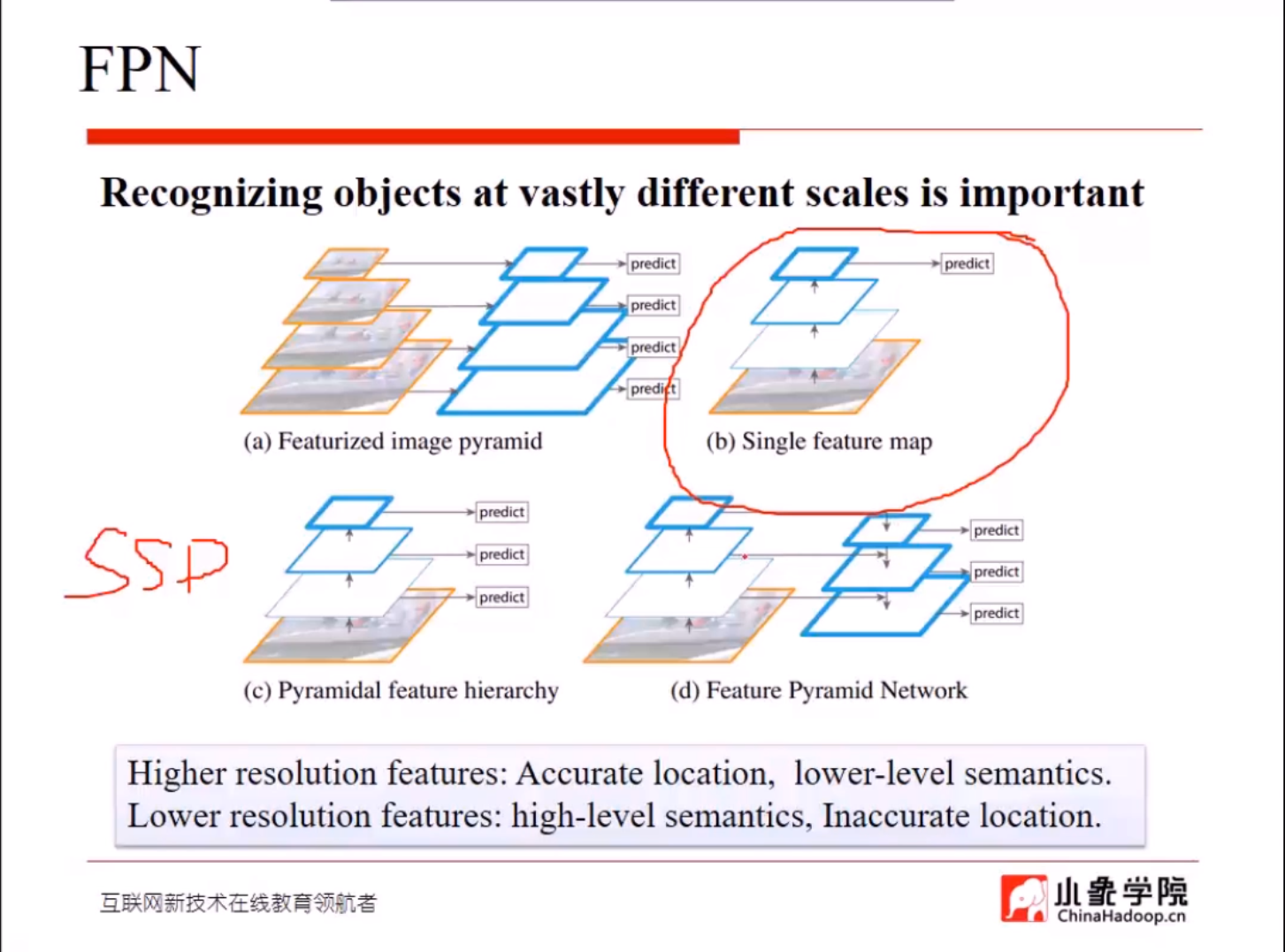
在图的图像中包括两个对象，人和灯，因此在YOLO的输出向量的表示中各单位的意思是 M\*N即使用YOLO方法将图像分为多少个网格 S为一个网格中有多少个边界框（bounding box）,(5+class)表示首先这个对象的类别是什么，然后表示网格中对象的中心点的X,Y的坐标，这个网格的长w和高h以及这个对象属于两个类别的概率是多少。在YOLO中输出的向量就是这个意思。



因此，需要对不同尺度的物体进行检测，所以Anguelov D等提出了一种名为单次多盒检测的方法，简称为SSD。它通过不同尺度的边框进入CNN的不同层中，允许每一层基于尺度价值来预测物体。比例值不过是图像的装饰参数。SSD适用于较大规模的对象，对于小规模的对象来说没法生成足够数量的高级特征。可为所有需要提高精度的等级提供丰富的语义。Wanli Ouyang et 等介绍了DeepIDNet来对各种各样的大小和形状的物体的几何特征进行学习。验证问题就是这种方法的一个挑战。为了克服SSD 和DeepIDNet的问题来提高精度， Dai等人提出了基于区域的完全卷积网络，简写为RFCN。相较于R-CNN，fast R-CNN，RFCN的一个优点就是通过共享计算结果从而降低系统的复杂度。它通过使用简单的训练策略来减少复杂性并且提供可接受的精度。

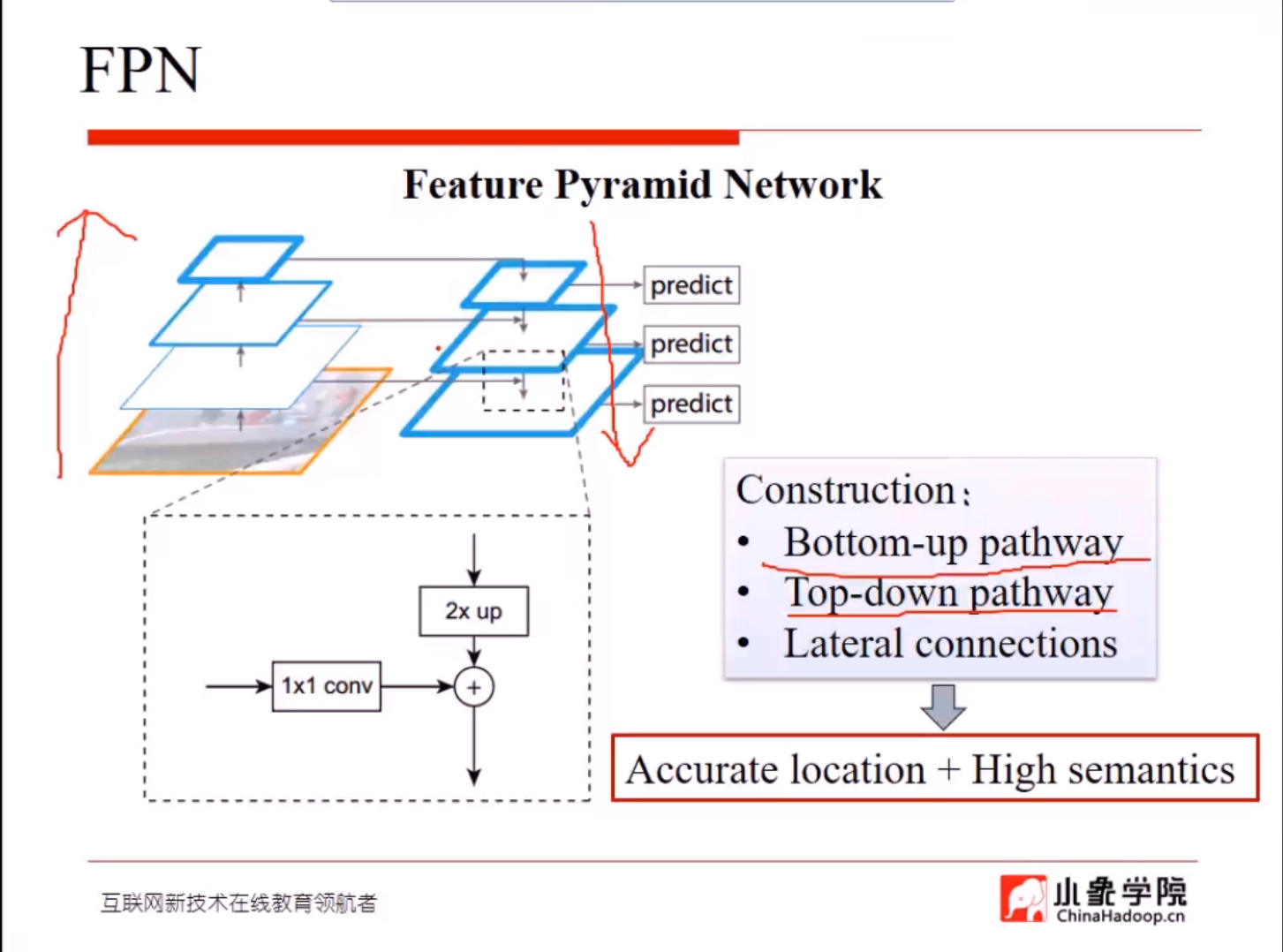
Lin T等提出了金字塔的概念来产生用于物体监测的更高质量的特征，成为特征金字塔网络或者FPN。FPN实际上是一个特征提取器，它是被用来改变其他探测器的提取器。FPN是一种分层的结构，语义特征的值随着每一层的增加而增加。语义丰富的层可以用于生成具有高分辨率的层。语义层对FPN的表现有很高的影响，FPN的一个挑战就是它的连接。例如减少从上到下的连接可以降低整个网络的精度。依旧有一些社会相关应用的工作需要做来解决FPN带来的的挑战。

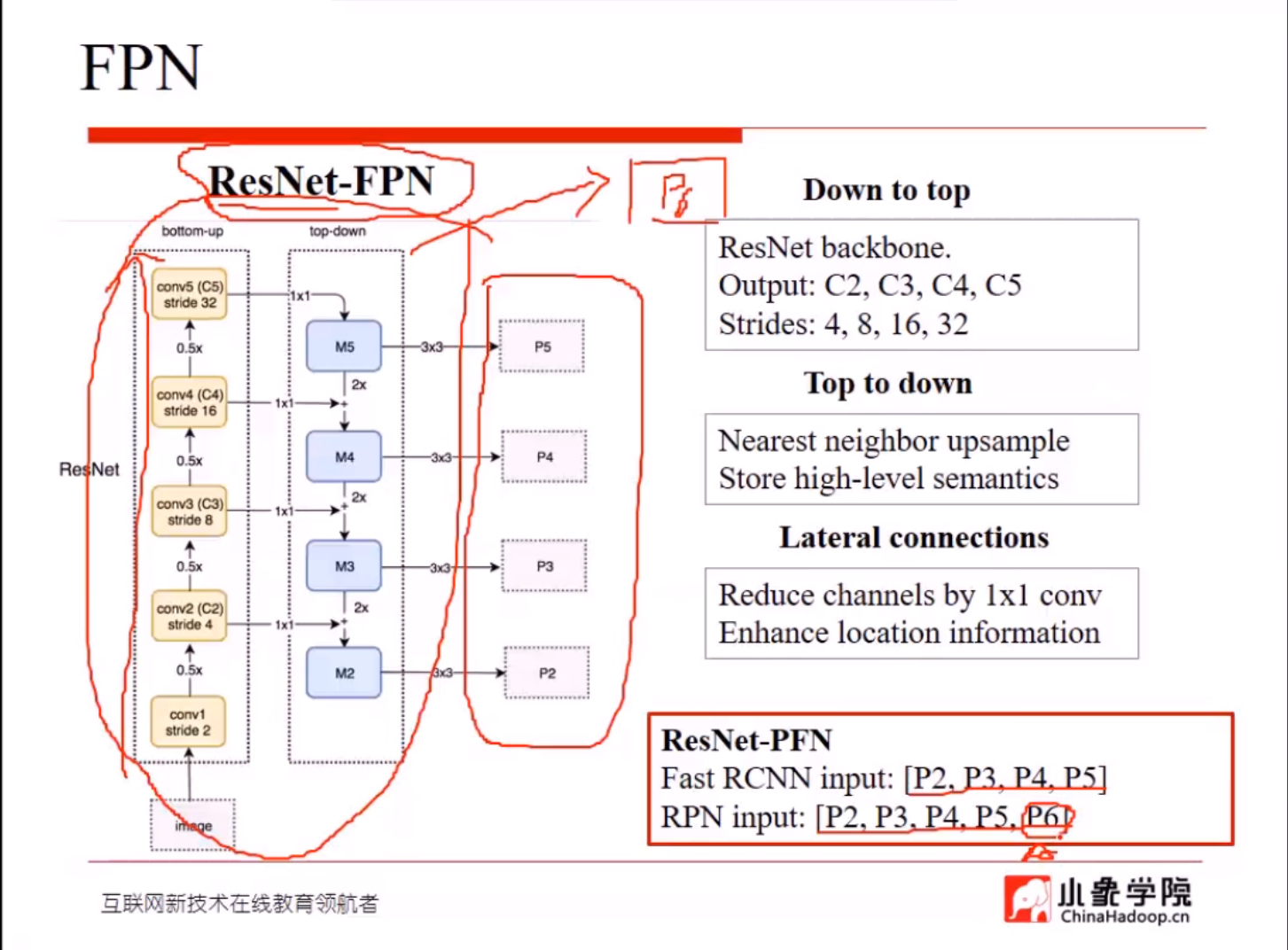
在对图像进行卷积的过程中，特征图像大（如只进行一两次卷积的图像）的图像保留的原图像的特征信息多，譬如原图像中的位置信息之类的信息，但是这种图像中所包含的语义信息很少，语义信息少的话会导致在对图像进行分类的时候很容易分类错误。而在经过几次的卷积之后，图片特征图像会变得很小，但是经过几次卷积之后的特征图像会具有很多的语义信息，这时候对特征图像进行分类就会变得很容易。就如下图的金字塔型结构，图（b）表述的应该就是faster R-CNN的结构，对特征图像使用VGG-16进行卷积，但是最后只使用最后的一个特征图像进行预测，前面的特征图像都丢弃不用。SSD方法就是如图（c）所示，从原图中使用VGG-16提取特征图像，然后使用每个提取的特征图像对对象进行预测，这样的话相较于faster R-CNN会提高对VGG-16中的特征图像的利用率，同时也会提高模型对于比较小的对象的预测能力，不会像faster R-CNN一样无法对图中的小的对象进行预测。但是由于SSD方法是先使用语义信息弱的信息对图中包含的对象进行预测，可能会导致出错。而FPN方法则对SSD方法中的这个缺点进行了改进。FPN方法如图（d）所示，这种方法也是对从原图中提取出来的所有特征图像都进行利用，但是与SSD方法不同的是，SSD方法是对从原图提取出来的特征图像进行从卷积次数少的特征图像到卷积次数多的特征图像进行预测。而FPN方法则是将从卷积图像中提取出来的特征图像从卷积次数多的特征图像（即语义信息强的特征图像）开始进行预测，然后再将语义特征强的特征图像与语义特征弱的特征图像进行融合，对融合之后得到的特征图像在进行预测。重复此过程既可以既对图像中的对象进行准确的预测，同时也不会导致预测对象的位置信息的丢失亦或者是出现像faster R-CNN无法对图像中较小的对象进行预测的情况。



FPN使用以上的方法既保证了有较高的定位精度，同时有保证了对含有丰富语义特征的位于后面的卷积层的特征图像的使用，这就是FPN的优点。

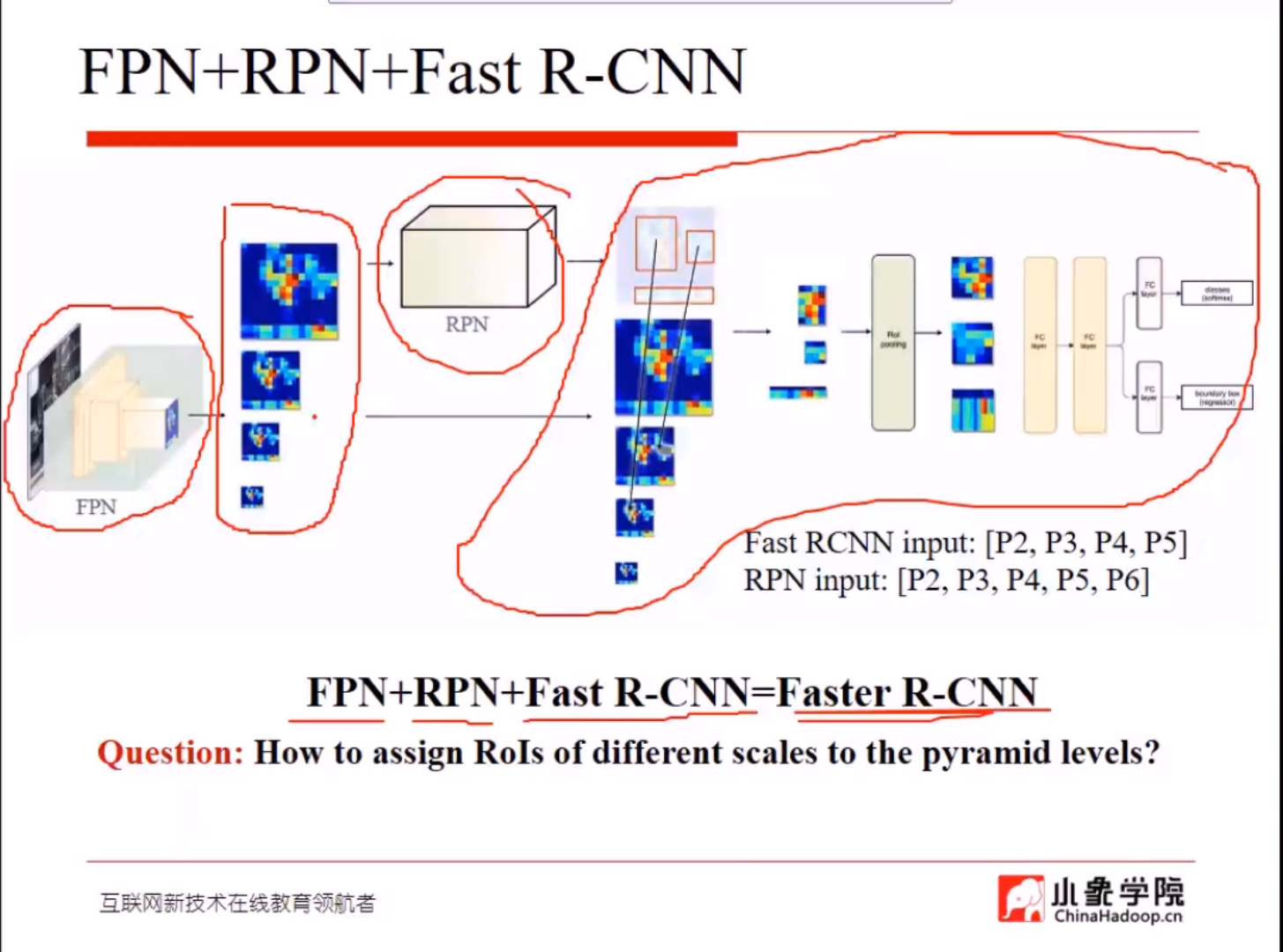
在目标检测，语义分割的任务中很多场景都会使用到ResNet-FPN.



下图为Rest-FPN的一个架构bottom-up，top-down两个框为FPN。  


RPN即为region proposal network，位于faster R-CNN中，即为快速生成提案的一个过程。VGG-16网络只会生成一个feature map但是使用FPN则会生成多个feature map。此Fast R-CNN与传统的Fast R-CNN不同。

具体的细节问题等到以后用到的时候在细究 视频地址av66083121



在此之后，TychsenSmith L 和 Petersson L设计了场景适应的目标检测模型，叫做DeNet。这种方法的基本思路是通过提供一个边框角估计的建议来预测对象的候选人提案。这种方法的优点是它不需要对锚进行预先定义而且比其他的对象检测技术要快得多。这种方法的一个问题就是它需要更多的时间来生成跳和评估基础网络。通过解决这个问题，DeNet可以有一个非常重大的性能提升。（可以对DeNet进行深入了解，新方法还没人用）

III. APPLICATIONS OF OBJECT DETECTION

对象检测在不同的领域内有广泛的应用是一个确定的趋势。实时监测应用在自动驾驶汽车上是用来检测道路上的树，其他车辆，行人，动物等。使用物体检测技术的一个流行趋势是通过遥感图像来检测沙漠绿洲，森林大火，坠机事故等，研究者已就开始朝这个方向进行努力。对象检测技术在医学图像检测例如肿瘤检测，CT等方面有重大意义，一些常用的应用如下所述：

1. *Face detection.*

物体检测的一个非常流行的应用就是人脸检测，包括检测在给出的图像中是否存在人脸。卷积神经网络在提取人脸的特征中扮演了一个重要的角色，对特征进行观察并回报一个最终的输出。社交媒体在图片处在上传阶段时将这项技术应用到检测人脸图像，并且在当前的世界中很重要。

*B. Vehicle detection.*

对象检测的另一个实时应用是检测车辆的存在例如汽车，摩托车，公共汽车等车辆，为在这种情况下的检测对象。为了跟踪道路上车辆的移动，速度是保持能否对对象进行跟踪的一个重要的因素，这种方案被证明是成功的。（是不是跟踪对象的速度的快慢是对系统能否检测成功有重要的影响）

*C. People counting.*

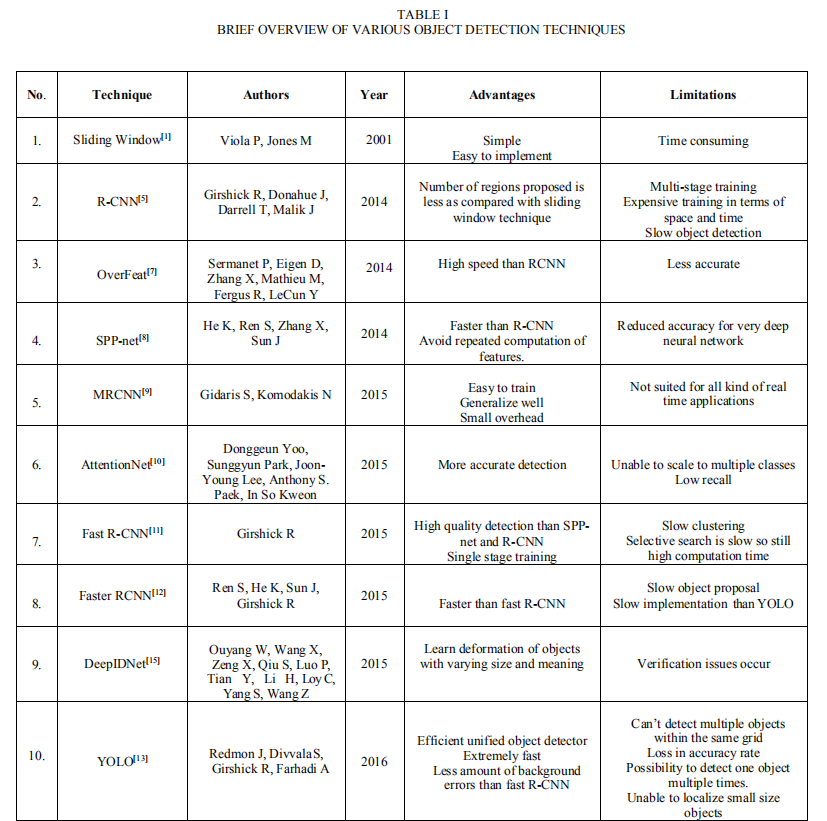
这个应用旨在统计在给定的场景中检测人的数量。统计一段视频中的人的数量是很困难的，但是却非常重要。这项技术主要应用在在一个事件中对人群进行分析。这项技术在很多国家已经有了积极地应用。（在景点门前的显示当前景区内有多少人数的电子屏幕上用的是不是这个技术？亦或者是通过售卖门票的多少来对景区中游客的数量进行统计）

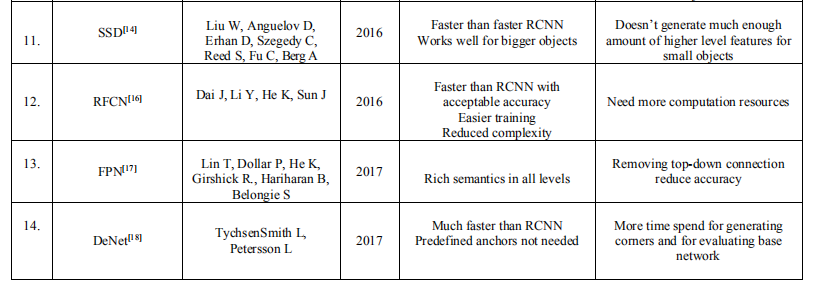
*D. Security and surveillance.*

该技术在当今世界中的应用价值非常高随着各种反社会行为的增加。基于遥感图像来检测入侵者和爆炸物等也属于这个应用。异常检测是另一个应用，每年很多公司都会在上面投很多钱。大量的研究在这个方面进行以使得这些系统可以实现自动化来提供系统的性能。

IV. COMPARATIVE STUDY

表1提供了各种应用于对象检测的技术的主要思想以及其优缺点。





V. SCOPE OF OBJECT DETECTION

最近几年，我们注意到在目标检测领域尤其是在遥感图像方面已经得到了巨大的完善。随着遥感图像的数量和质量的增加，为这个领域带来了机会和挑战。遥感图像是在离地球非常远的高空拍摄的高质量图片，在变化的光照条件下具有各种应用。对象检测的精度可以随着图片数量的增加而增加。即使图片的质量增加了，使用较少数量的高质量图片训练的模型也没有使用大量图片训练出来的模型有效。随着在图像中出现的对象数量的增加，图像得背景的复杂度会增加，结果就是对这些图像进行分析的困难度会得到很大的上升（因为图像中包含的图像数量增加）。所以对象检测在遥感图像中是一个非常重要的任务。对象检测是检测在图像（遥感图像）中包含的所有的对象，例如道路，水源，车辆，建筑，森林火灾等。对图像中存在的所有对象进行检测是很有效的，而且也是一种有效的方法来对这些图像进行分析。

遥感图像在环境的监视，追踪和改变检测等方面很重要而且非常的有用。跟踪一个城市的成长，农业用地的减少，森林用地的减少，水源的消失所有的这些都在变更检测和环境监测的标签下。我们使用超过20年的遥感图像并对这些图像进行分析。从这些图像中，我们可以做出一些预测例如处在任何条件下的河流。这就是对遥感图像的一些独特的应用，并且以研究为导向。例如监测石油的储备，火山喷发，军事监测，监测和控制森林火灾，绘制沙漠中的绿洲，天气预测等这些全部都需要在自然环境下的高分辨率遥感图像。

可以通过遥感图像来进行物体检测，模型可以通过嵌入或者是通过对候选区域提案，特征提取和将这些图像天转换为的最终分类标签进行三个步骤的简单连接。候选区域提案是对感兴趣的区域进行绘制的过程，从某种意义上说，在我们感兴趣的区域上有一个对象有一些偶然的，在如果对象不再感兴趣的区域上就不会产生兴趣。特征的提取伴随着对每个提案中的独特的，高级特征得提取。更加合适和高效的区域提案可以更高的对特征进行提取，反过来也会提高物体检测和分类的准确率。

VI. CONCLUSION

基于深度学习的对象检测技术在过去的几年中得到了很大的发展，在这个过程中，物体检测技术的应用也得到了较大的发展。在这片survey中提供了一个简单明了的关于不同物体检测技术的想法，而且列举了一些突出和具有挑战性的工作。这项工作还专注于对象检测中的各项应用。除此之外，本论文还提到了在对象检测中的具有光明前景的领域———遥感图像的检测。这个survey在对象检测方面提供了有价值的观察并且提供了新的研究方向。