**摘要：**

我们提出了一个使用卷积网络进行分类，定位和检测的综合框架。我们展示了如何在ConvNet中有效地实现多尺度和滑动窗口方法。我们还通过学习预测物体边界，介绍一种新颖的深入学习方法来进行定位。通过累加而不是抑制边界框以增加检测置信度。我们显示使用单个共享网络可以同时学习不同的任务。该集成框架是ImageNet大型视觉识别挑战2013（ILSVRC2013）的定位任务的获胜者，并为检测和分类任务获得了非常有竞争力的结果。在竞赛后的工作中，我们建立了新的检测任务。最后，我们从我们最好的模型“OverFeat”中发布一个特征提取器。

**1.简介：**

识别图像中主要对象的类别是卷积网络（ConvNets）[17]已应用多年的任务，无论对象是手写字符[16]，房屋号码[24]，无纹理玩具[18]]，交通标志[3,26]，Caltech-101数据集[14]中的对象，或1000类别ImageNet数据集[15]中的对象。小数据集（如Caltech-101）上的ConvNets的准确性虽然体面，但尚未创纪录。然而，较大数据集的出现使ConvNets能够显着提升数据集上的最新技术，例如1000类ImageNet [5].

ConvNets对于许多这样的任务的主要优点在于，整个系统是从端到端进行的，从原始像素到最终类别，从而减轻了手动设计合适的特征提取器的要求。主要的缺点是他们对贴标训练样本的贪心胃口。

本文的主要内容是展示了对图像中对象进行分类，定位和检测的卷积网络训练，可以提高所有任务的分类精度和检测定位精度。 本文提出了一种新的ConvNet对象检测，识别和定位的综合方法。 我们也介绍一种新的方法通过积累预测的边界框来定位和检测。 我们建议通过组合定位和预测，可以在不对负样本进行训练的情况下执行检测，并且可以避免耗时且复杂的自举训练通行证。 没有对负样本进行训练也使得网络只关注正样本，以获得更高的准确性。实验将在ImageNet ILSVRC 2012和2013数据集上进行，并在ILSVRC 2013本地化和检测任务上建立最新的结果。虽然ImageNet分类数据集的图像被大量地选择为包含大部分图像的粗略对象，但是感兴趣的对象在图像中的大小和位置上有显着差异。解决这个问题的第一个想法是在图像中的多个位置，滑动窗口方式和多个尺度上应用ConvNet。

 然而，即使这样，许多查看窗口可能包含对象（例如，狗的头部）的完全可识别的部分，但不包括整个对象，甚至是对象的中心。这产生了恰当的分类，但是局部化和检测差。因此，第二种想法是训练系统不仅产生每个窗口的类别分布，而且还产生相对于窗口包含对象的边界框的位置和大小的预测。第三个想法是在每个位置和大小上累积每个类别的证据。

许多作者提出使用ConvNets进行检测和定位，在多个尺度上有一个滑动窗口，可以追溯到20世纪90年代初的多字符串[20]，面[30]和手[22]。最近，ConvNets已被证明在自然图像[4]，面部检测[8,23]和行人检测[25]中产生文本检测的最新性能。

几位作者还提出训练ConvNets来直接预测要定位的对象的实例化参数，例如相对于观察窗口的位置或对象的姿态。例如Osadchy等人[23]描述了ConvNet用于同时进行脸部检测和姿态估计。脸部由九维输出空间中的3D模型表示。模型上的位置表示姿势（俯仰，偏航和滚动）。当训练图像是脸部时，训练网络以在已知姿势的位置处在模型上产生点。如果图像不是脸部，则输出原理模型。在测试时间，到模型的距离表示图像是否包含一个脸部，并且模型上最近点的位置表示姿势。泰勒等人[27,28]使用ConvNet来估计身体部位（手，头等）的位置，以便导出人体姿势。他们使用度量学习标准来训练网络以在身体姿势多元组上产生点。Hinton等还提出训练网络来计算特征的显式实例化参数，作为识别过程的一部分[12]。

其他作者提出通过基于ConvNet的分割来执行对象定位。最简单的方法是训练ConvNet将其视窗的中心像素（或体积图像的体素）分类为区域之间的边界[13]。但是，当区域必须分类时，最好进行语义分割。主要思想是训练ConvNet，使用窗口作为决策的上下文，将观察窗口的中心像素与其所属对象的类别进行分类。应用范围从生物图像分析[21]到移动机器人的障碍物标签[10]来标记照片[7]。这种方法的优点是边界轮廓不需要是矩形，并且这些区域不需要被良好的外接对象。缺点是需要密集的像素级标签进行培训。这种分割预处理或对象建议步骤最近在传统的计算机视觉中得到普及，以减少检测位置，尺度和纵横比的搜索空间[19,2,6,9]。因此，可以在搜索空间中的最佳位置应用昂贵的分类方法，从而增加识别精度。此外，[29,1]表明，这些方法通过大大减少不太可能的物体区域来提高准确度，从而减少潜在的假阳性。然而，我们的密集滑动窗口方法能够超越ILSVRC13检测数据集上的对象提案方法。

Krizhevsky等人[15]最近使用大型ConvNet展示了令人印象深刻的分类性能。作者还进入了ImageNet 2012比赛，赢得了分类和定位的挑战。虽然网络在定位方面给人留下了深刻的印象，但还没有发表任何描述他们的方法的作品。因此，我们的论文是第一个提供ConvNets清晰的解释并且能够用于Imagenet数据的定位和检测。

在本文中，我们使用的目标定位和检测与ImageNet 2013竞赛中的使用一致，并且都涉及预测图像中每个对象的边界框预测，即唯一的区别是使用的评价标准。



**2.Vision Tasks：**

在本文中，我们以增加难度的顺序探索三种计算机视觉任务：（i）分类，（ii）定位和（iii）检测。每个任务都是下一个任务。虽然使用单个框架和共享功能学习基础来处理所有任务，但我们将在以下部分中单独描述它们。

在整篇文章中，我们报告了2013 ImageNet大型视觉识别挑战赛（ILSVRC2013）的成果。在该挑战的分类任务中，每个图像被分配与图像中的主要对象相对应的单个标签。五个猜测被允许找到正确的答案（这是因为图像也可以包含多个未标记的对象）。定位任务类似于每个图像允许5个猜测，但是除此之外，每个猜测都必须返回预测对象的边界框。要被认为是正确的，预测的框必须匹配地面真实度至少为50％（使用PASCAL标准的联合对交叉点），以及用正确的类标记（即每个预测是一起关联在一起的标签和边界框）。检测任务与定位不同之处在于，每个图像中可以存在任何数量的对象（包括零），并且假阳性将被平均精度（mAP）度量所惩罚。本地化任务是分类和检测之间的便利中间步骤，并且允许我们独立于检测特定的挑战（例如学习背景类）来评估我们的定位方法。在图1中，我们示出了具有我们的定位/检测预测以及相应的基准的图像的示例。请注意，分类和定位共享相同的数据集，而检测还具有其中对象可以更小的附加数据。检测数据还包含一组不存在某些对象的图像。这可以用于引导，但我们在这项工作中没有使用它。

**3.分类：**

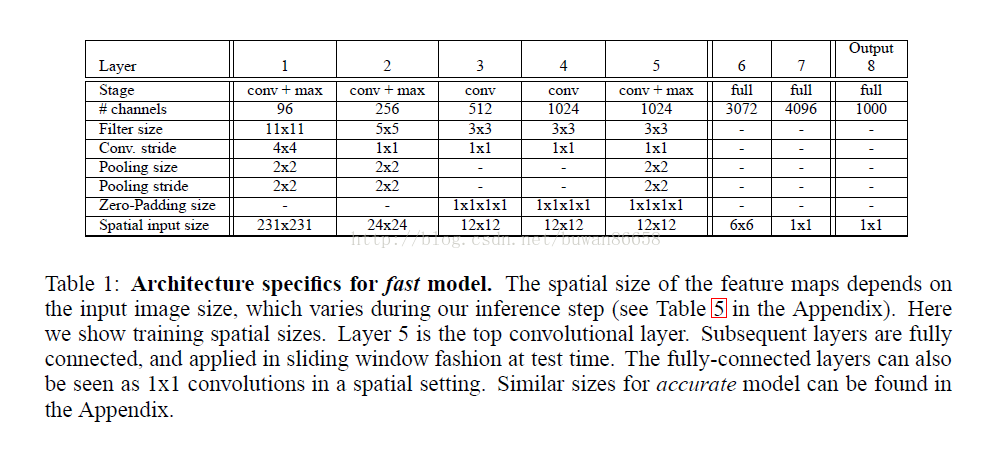
我们的分类架构类似于Krizhevsky等人最好的ILSVRC12架构。[15]。然而，我们改进了网络设计和推理步骤。由于时间限制，Krizhevsky模式的一些训练特征没有被探索，所以我们期望我们的结果能够进一步提高。这些在未来的工作部分[6]



**3.1模型设计和训练：**

我们在ImageNet 2012训练集上训练网络（120万张图像和C = 1000类）[5]。我们的模型使用与Krizhevsky等人提出的相同的固定输入尺寸方法。[15]在训练期间，转向多尺度进行分类，如下一节所述。每个图像进行下采样，使最小尺寸为256像素。然后我们提取大小为221x221像素的5个随机作物（和它们的水平翻转），并以128的小批量呈现给网络。网络中的权重用（μ，σ）=（0,1，10-2）。然后随机梯度下降更新，动量项为0.6，ℓ2重量衰减为1×10-5。学习率最初为5×10-2，并且在（30,50,60,70,80）时期之后依次减少0.5。在分类器中完全连接的层（第6和第7）采用速率为0.5的DropOut [11]。

我们详细说明了表1和表3中的体系结构尺寸。注意，在训练过程中，我们将这种架构视为非空间（大小为1x1的输出图），而不是产生空间输出的推理步骤。层1-5类似于Krizhevsky等人[15]，使用整流（“relu”）非线性和最大池化，但具有以下差异：（i）不使用对比度归一化;（ii）池化区域不重叠，（iii）由于步幅较小（2而不是4），我们的模型具有较大的第1和第2层特征图。更大的步幅有利于速度，但会损失精度。



在图2中，我们显示了来自前两个卷积层的滤波器系数。第一层过滤器捕获定向的边缘，图案和斑点。在第二层中，过滤器具有各种形式，一些扩散，其它具有强线结构或定向边缘。

**3.2特征提取**

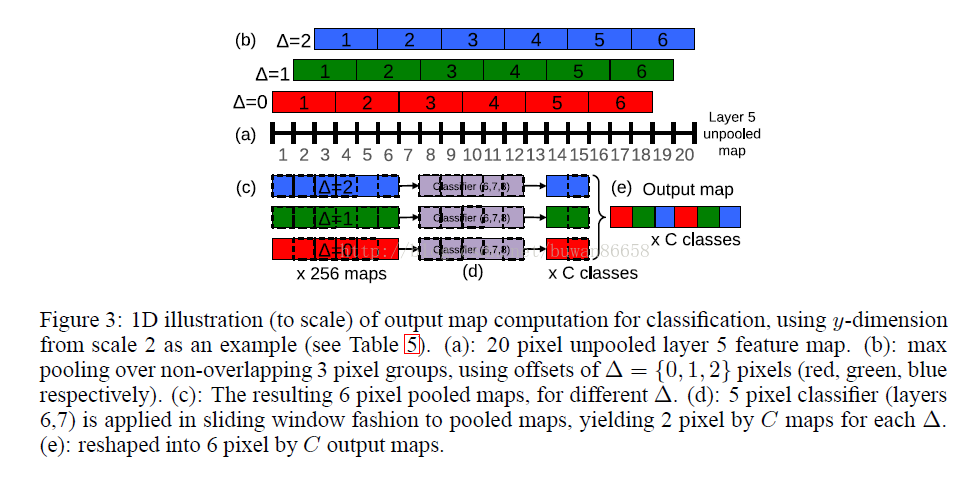
随着这篇文章，我们发布了一个名为“OverFeat”的功能提取器1，以便为计算机视觉研究提供强大的功能。提供两种模型，快速和准确。每个架构在表1和表3中进行了描述。我们还在表4中比较了它们在参数和连接方面的大小。准确的模型比快速模型更准确（14.18％的分类误差，而不是表2中的16.39％），但是它需要几乎两倍的连接数。使用7个精确模型的委员会达到13.6％的分类误差，如图4所示。

**3.3多尺度分类**

在[15]中，多视图投票用于提升性能：固定的10个视图（4corners和center，水平翻转）的平均值。然而，这种方法可能忽略图像的许多区域，并且在视图重叠时计算冗余。此外，它仅适用于单一尺度，ConvNet可能不会以最佳置信度进行多尺度响应。  
  
相反，我们通过在每个位置和多个尺度上密集地运行网络来探索整个图像。虽然滑动窗口方法对于某些类型的模型可能是计算上禁止的，但在ConvNets的情况下固有的效率（参见第3.5节）。这种方法产生了更多的投票视图，这增强了鲁棒性，同时保持了有效性。ConvNet在任意大小的图像上卷积的结果是每个尺度上的C维向量的空间图。  
  
然而，上述网络中的总子采样比是2x3x2x3或36.因此，当密集应用时，该架构只能沿着每个轴在输入维度中每36个像素产生分类矢量。与10视图方案相比，输出的这种粗略分布降低了性能，因为网络窗口与图像中的对象不是很好地对齐。更好地对齐网络窗口和对象，网络响应的信心最强。为了规避这个问题，我们采取类似于Giusti等人介绍的方法。[9]，并在每个偏移处应用最后一个子采样操作。这消除了该层的分辨率损失，产生x12而不是x36的总次采样比。

我们现在详细解释如何执行分辨率增强。我们使用6个输入尺度，导致不同分辨率的未压缩层5映射（详见表5）。然后将它们按照以下步骤进行分类并提供给分类器，如图3所示：

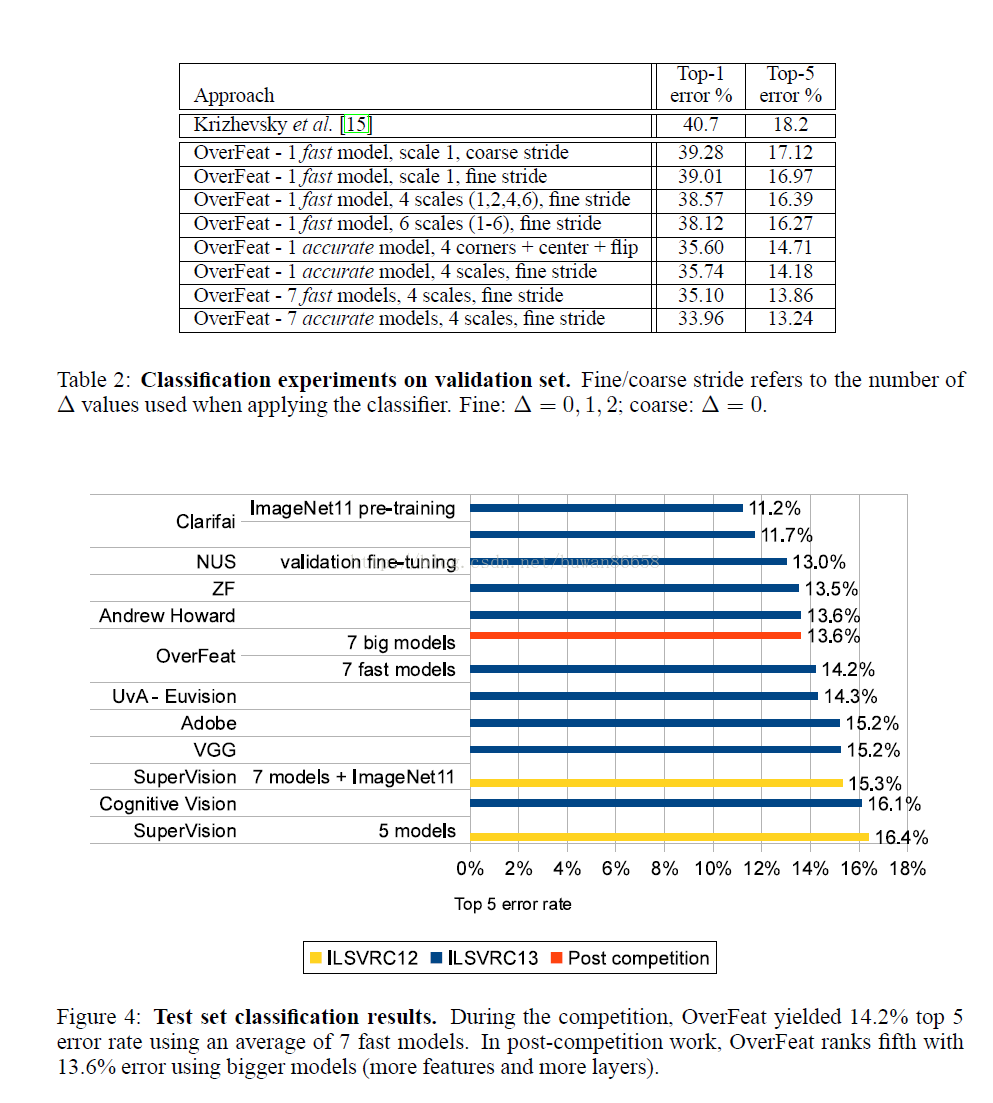
（a）对于一个图像，在给定的比例，我们先从UNPOOLED层5的特征图。  
（b）每个unpool特征图经过3×3的最大池化（非重叠区域），重复3×3次（Δx，Δy）像素偏移{0,1,2}。  
（c）这产生一组合并的特征图，对于不同的（Δx，Δy）组合复制（3×3）次。  
（d）分类器（层6,7,8）具有5×5的固定输入大小，并且为汇池化特征内的每个位置产生C维输出向量。 分类器以滑动窗口方式应用于池化特征，产生C维输出图（对于给定的（Δx，Δy）组合）。  
（e）不同（Δx，Δy）组合的输出图重新整形为单个3D输出图（两个空间维度x C类）。



这些操作可以被看作是通过池化层将分类器的观察窗口移动1个像素，而不需要在二次采样中使用跳跃内核，并且在邻近的值不相邻的地方使用跳过内核。或者等效地，在每个可能的偏移处应用最终的合并层和完全连接的堆栈，并且通过交织输出来组合结果。  
  
对于每个图像的水平翻转版本，重复上述过程。然后，我们通过以下方式产生最终分类：（i）在每个尺度上取每个类别的空间最大值，并且翻转;（ii）对来自不同尺度和翻转的所得到的C维向量进行平均，（iii）从平均类向量中取出前1个或前5个元素（取决于评估标准）。  
  
在直观的层面上，网络的两半 -特征提取层（1-5）和分类器层（6输出） - 以相反的方式使用。在特征提取部分中，滤波器在一次通过中跨越整个图像卷积。从计算的角度来看，这比将固定大小的特征提取器在图像上滑动，然后聚合来自不同位置的结果更有效【2】。然而，对于网络的分类器部分，这些原理是相反的。在这里，我们希望在不同位置和比例的第5层特征图中寻找固定大小的表示。因此，分类器具有固定大小的5x5输入，并且被穷尽地应用于第5层地图。穷尽的汇总方案（单个像素移位（Δx，Δy））确保我们可以在分类器和特征图中对象的表示之间获得精确的对齐。

**3.4结果**

在表2中，我们尝试了不同的方法，并将其与Krizhevsky等人的单一网络模型进行了比较。[15]供参考。上述方法具有6个尺度，达到了13.6％的前5个错误率。可以预期的是，使用较少的尺度会降低性能：单分数模型更糟糕的是top-5错误率为16.97％。图3所示的精细步幅技术在单一尺度方面带来了相当小的改进，但对于这里显示的多尺度增益也是重要的。

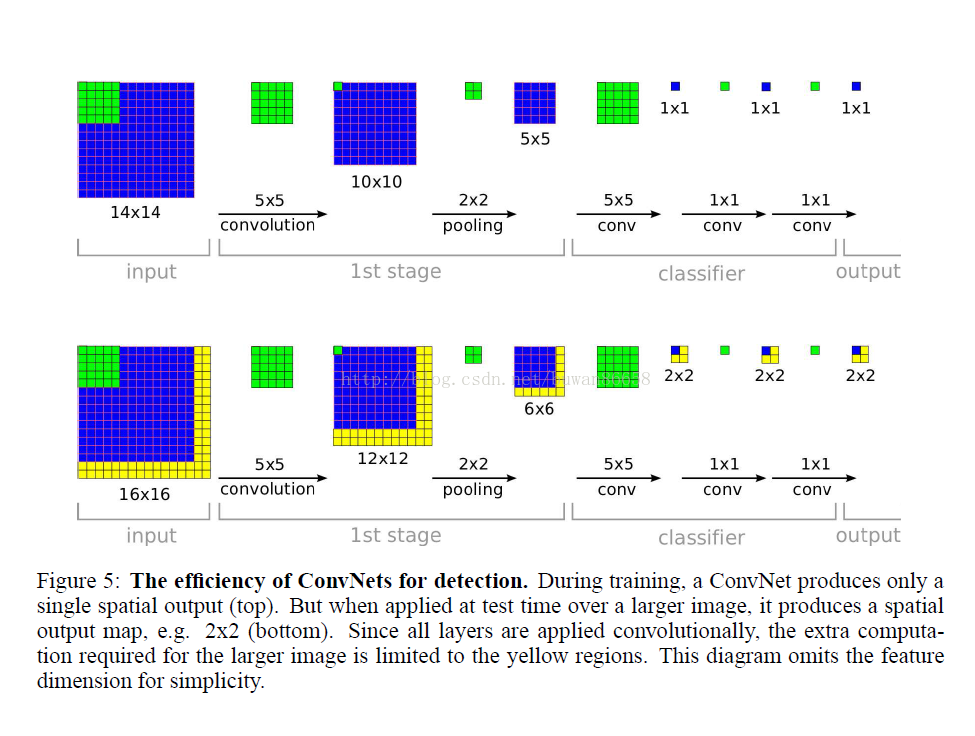


我们报告了图4中的2013年比赛的测试结果，其中我们的模型（OverFeat）通过投票7个ConvNets（每个训练有不同的初始化）获得14.2％的错误率，并在18个队伍中排名第5。仅使用ILSVRC13数据的最佳错误率为11.7％。来自ImageNet Fall11数据集的额外数据的预训练将错误率降到11.2％。在竞赛后的工作中，我们通过使用更大的模型（更多的特征和更多的层），将OverFeat的结果降到13.6％的错误率。由于时间限制，这些较大模型未经过充分训练，在不久的将来就会得到很好的改善。

**3.5 convNets 和滑动窗效率**

与许多滑动窗口方法相比，一次为输入的每个窗口计算整个流水线时，ConvNets以滑动方式能够得到更好的效率，因为它们自然地共享重叠区域共有的计算。当在测试时将网络应用于较大的图像时，我们简单地在整个图像的范围内应用每个卷积。这扩展了每个图层的输出以覆盖新的图像大小，最终产生输出类预测的映射，每个“窗口”（视野）的输入具有一个空间位置。这在图5中示出。卷积自底向上应用，使得相邻窗口共同的计算仅需要执行一次。

请注意，我们的架构的最后一层是全连接的线性层。在测试时，这些层被1x1空间范围的内核的卷积运算有效地替代。整个ConvNet只是简单的一系列卷积，最大池和阈值操作。



**4. 定位**

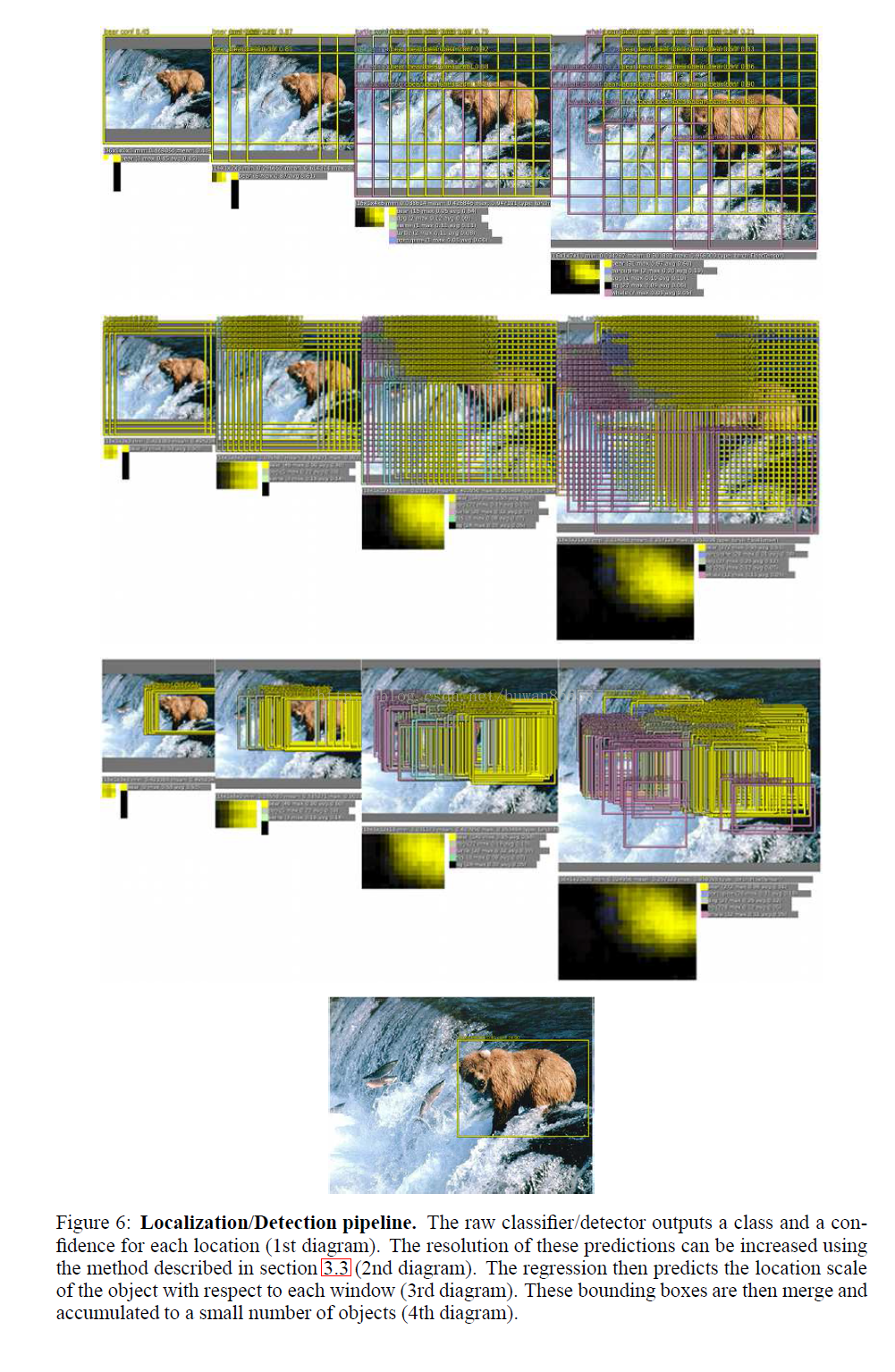
从我们的分类训练网络开始，我们用回归网络代替分类器层，并对其进行训练，以预测每个空间位置和尺度上的对象边界框。然后，我们将回归预测结合在一起，以及每个位置的分类结果，如我们现在所描述的。

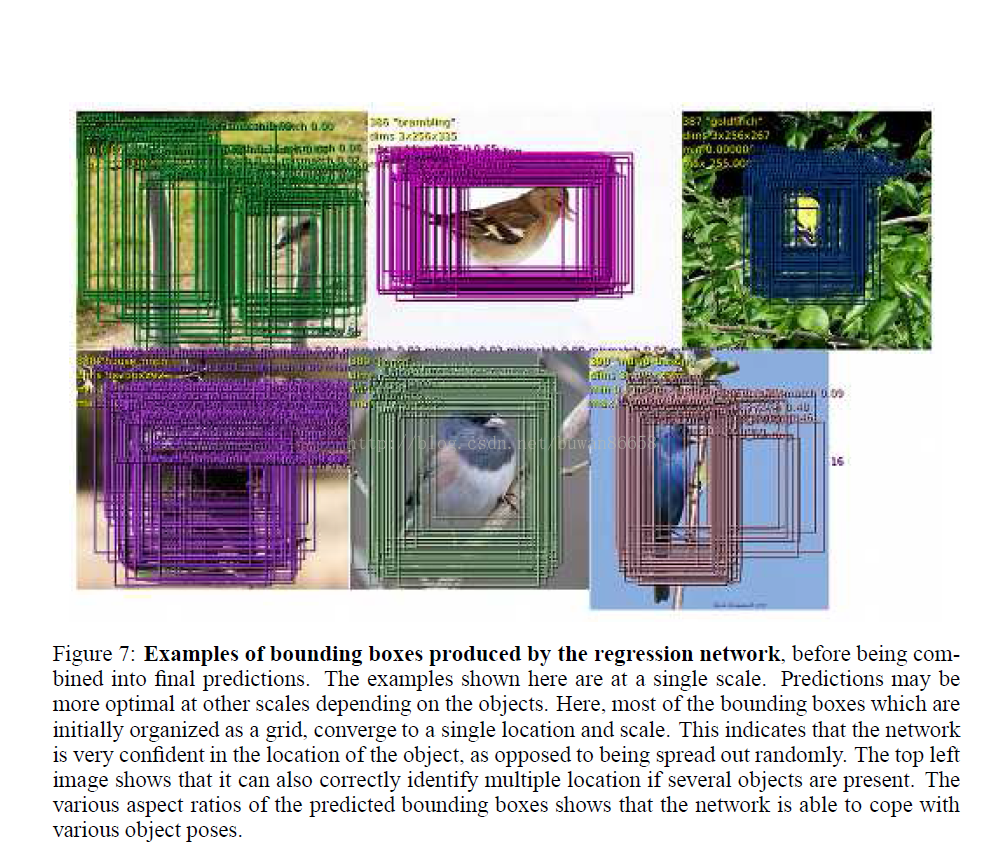
**4.1 生成预测**

为了生成对象边界框预测，我们同时在所有位置和尺度上运行分类器和回归网络。由于这些共享相同的特征提取层，在计算分类网络之后，仅需要重新计算最终的回归层。在每个位置的类c的最终softmax层的输出提供了在相应视野中存在类c的对象（尽管不一定完全包含）的可信度得分。因此，我们可以为每个边界框分配置信度。

**4.2 回归训练**

回归网络将层5的池化特征映射作为输入。它具有2个完全连接的隐藏层，分别为4096个和1024个信道。最终输出层有4个单位，用于指定边界边框的坐标。与分类一样，总共有（3x3）个副本，从Δx，Δy移位所得。该架构如图8所示。



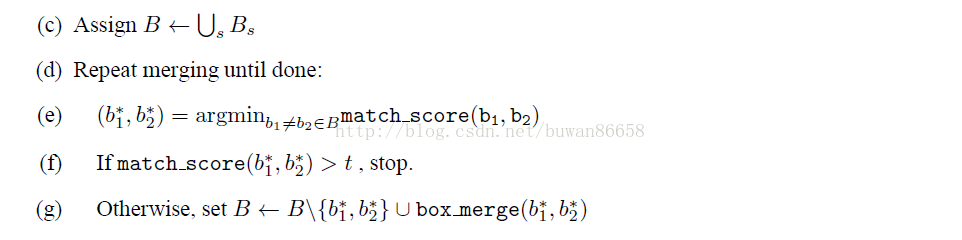
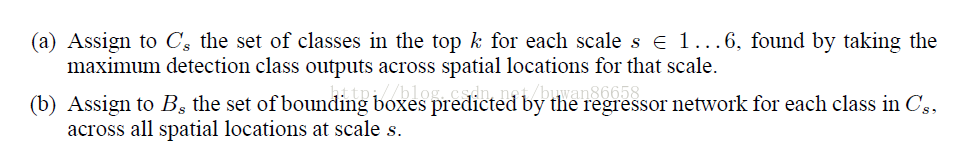


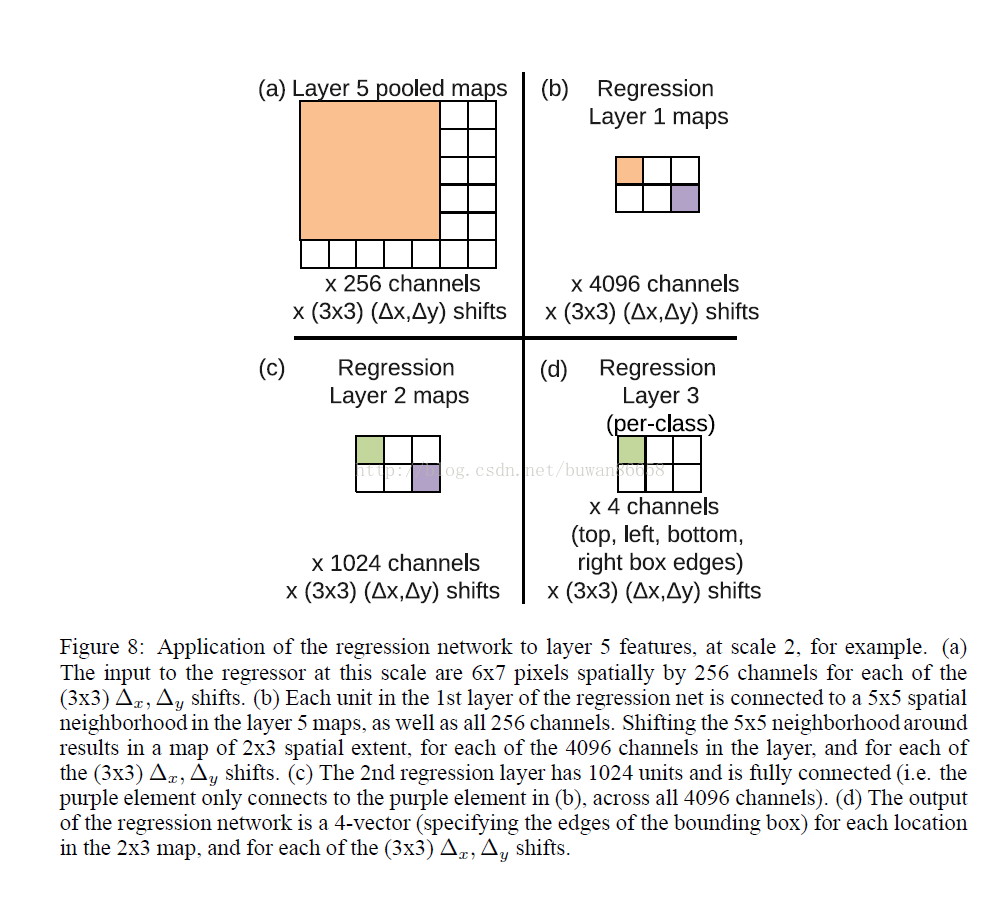
我们固定分类网络中的（1-5）特征提取层，并使用每个示例的预测和真实边界框之间的l2损失来训练回归网络。最后的回归层是类特定的，具有1000个不同的输出，每个类别一个。我们使用与第3节中描述的相同的比例尺来训练该网络。我们将每个空间位置处的回归网的预测与ground-truth边界框进行比较，其转移到卷积内的回归器的平移偏移的参考系（见图8）。然而，我们不会在输入视野下与重叠的小于等于50％的边界框上训练回归器：由于对象大多在这些位置之外，所以通过包含该对象的回归窗口可以更好地处理。

以多尺度方式训练回归分析对于跨尺度预测组合是很重要的。单一规模的训练在一定尺度上表现良好，并且在其他尺度上仍然表现合理。但是训练多尺度模型能够对跨尺度问题进行更好的预测，并成倍地提高合并后的预测的置信度。反过来，这允许我们仅用几个尺度来表现良好，而不是像检测中通常的情况那样的许多尺度。行人检测中的一个比例与另一个比例的典型比例[25]约为1.05到1.1，但是我们使用大约1.4的大比例（每个比例的数字不同，因为尺寸被调整以适合我们网络的步幅）这使我们能够更快地运行我们的系统。

**4.3 联合预测**

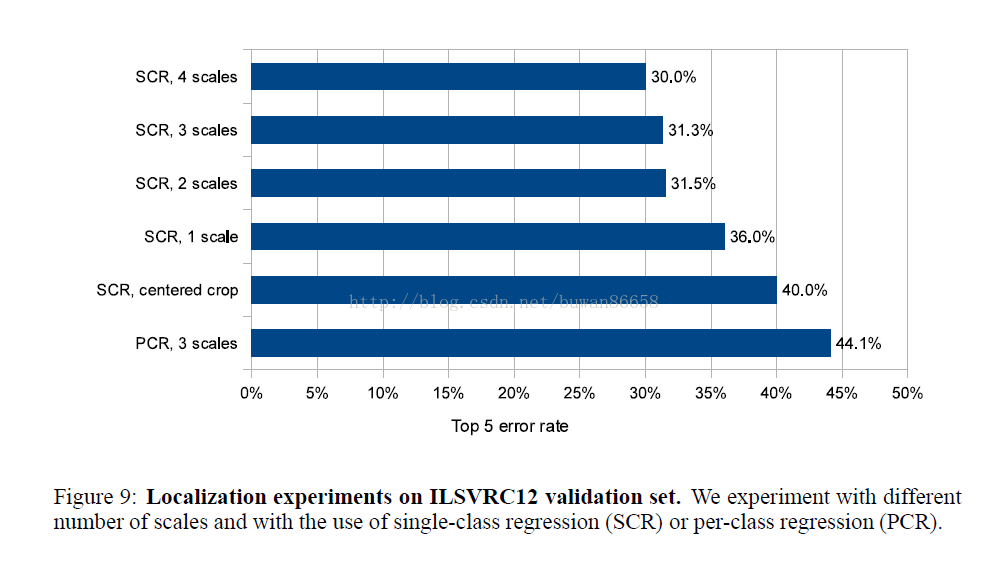
我们结合通过贪心合并策略的个体预测（参见图7）应用到回归边界框，使用下面的算法。

****



在上文中，我们使用两个边界框的中心之间的距离和框的交叉区域之和来计算匹配分数。框合并计算边界框坐标的平均值。

最后的预测是通过采用具有最大类分数的合并边界框来给出。这是通过累积地添加与预测每个边界框的输入窗口相关联的检测类别输出来计算的。对于合并成单个高可信区域框的边界框的示例，参见图6。在这个例子中，一些乌龟和鲸鱼边框出现在中间的多尺度步骤中，但在最终检测图像中消失。这些边界框不仅具有较低的分类置信度（分别至多为0.11和0.12），它们的收集并不像熊边框那样一致，以获得显着的置信度。熊的边框具有较强的置信度（平均每个规模约为0.5）和高匹配分数。因此，合并后，许多熊绑定框被融合成一个具有高置信度的边界，而错误的负样本在检测阈值以下，因为它们缺乏边界框的一致性和置信度。这个分析表明，我们的方法自然地比纯传统的非最大抑制方法更有利于增强分类模型的假阳性，通过奖励边界框的一致性。



**4.4 实验**

我们将网络应用于Imagenet 2012验证集，并使用比赛指定的定位衡量标准。其结果如图9所示，图10显示了2012年和2013年本定位比赛的结果（训练和测试数据都和近年的数据相同）。我们的方法以29.9%的错误率获得2013年比赛的冠军。

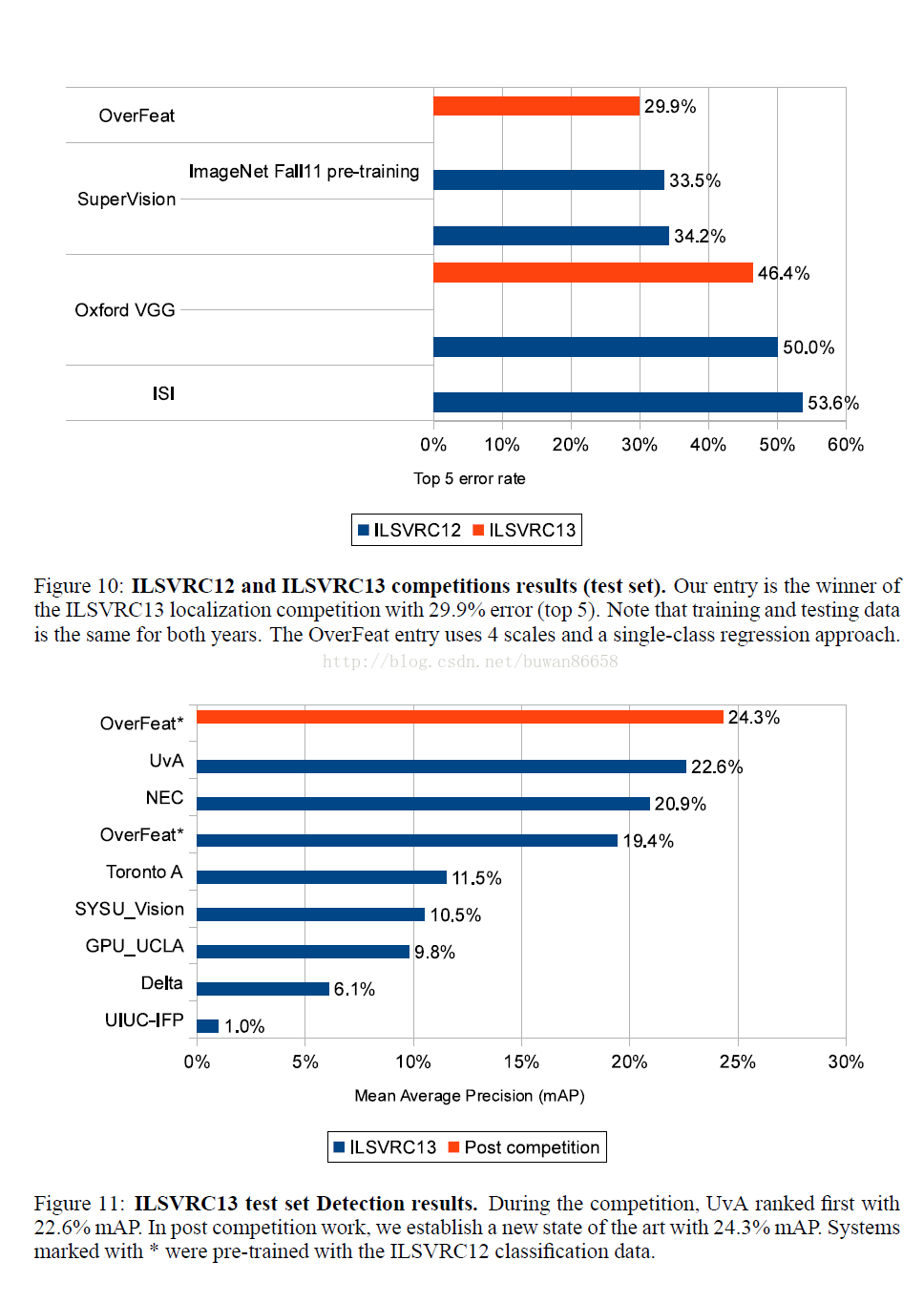
我们的多尺度和多视角方法对于获得良好的性能至关重要，如图9所示：仅使用单个居中的作物，我们的回归网络实现了40％的错误率。通过将来自所有空间位置的回归预测结合到两个尺度，我们实现了更好的误码率为31.5％。添加第三和第四比例进一步将性能提高到30.0％的误差。

对于每个类的回归网络使用不同的顶层（图9中的每类回归（PCR）），令人惊讶地不是仅使用所有类中共享的单个网络（44.1％对31.3％）。这可能是因为在训练集中用边界框注释的类别相对较少，而网络具有1000倍以上的顶层参数，导致训练不足。可以通过仅在类似类别之间共享参数来改善该方法（例如，针对所有类型的狗训练一个网络，针对车辆训练一个网络等）。

**5.检测**

检测训练类似于分类训练，但是以空间方式。可以同时训练图像的多个位置。由于模型是卷积的，所有权重都在所有位置之间共享。与定位任务的主要区别是当没有对象存在时，必须预测背景类。传统上，负样本最初是随机的训练。然后将最有罪的负面错误添加到自举传递中的训练集中。独立引导通过渲染训练的负样本集和训练时间之间的复杂和潜在风险不匹配。另外，需要调整自举通行的大小，以确保训练不会在一小组上过度使用。为了规避所有这些问题，我们通过选择一些有趣的负样本，例如随机的或最令人讨厌的，来进行负样本训练。这种方法在计算上更昂贵，但是使程序更简单。并且由于特征提取最初是用分类任务训练的，所以检测微调不是很长。

在图11中，我们报告了ILSVRC 2013比赛的结果，我们的检测系统排名第三，平均精度为19.4％（mAP）。之后，以24.3％的mAP建立了新的检测状态。 请注意，前3名方法与其他队伍之间存在巨大差距（第4种方法产生11.5％的mAP）。此外，我们的方法与使用初始分段步骤将候选窗口从大约20万到2000减少的前2个其他系统有很大不同。这种技术加速了推理，并大大减少了潜在的假阳性数量。[29，1]表明，当使用密集滑动窗口时，检测精度下降，而不是选择性搜索，从而丢弃不太可能的对象位置，从而减少误报。结合我们的方法，我们可以观察到类似的改进，如传统密集方法和基于分割的方法之间所见。还应该指出的是，我们没有对NEC和UvA的检测验证集进行微调。验证和测试集分布与培训集中的差异显着不足，单独将测试结果提高约1个点。图11中两个OverFeat结果之间的改善是由于更长的训练时间和上下文的使用，即每个刻度也使用较低的分辨率刻度作为输入。

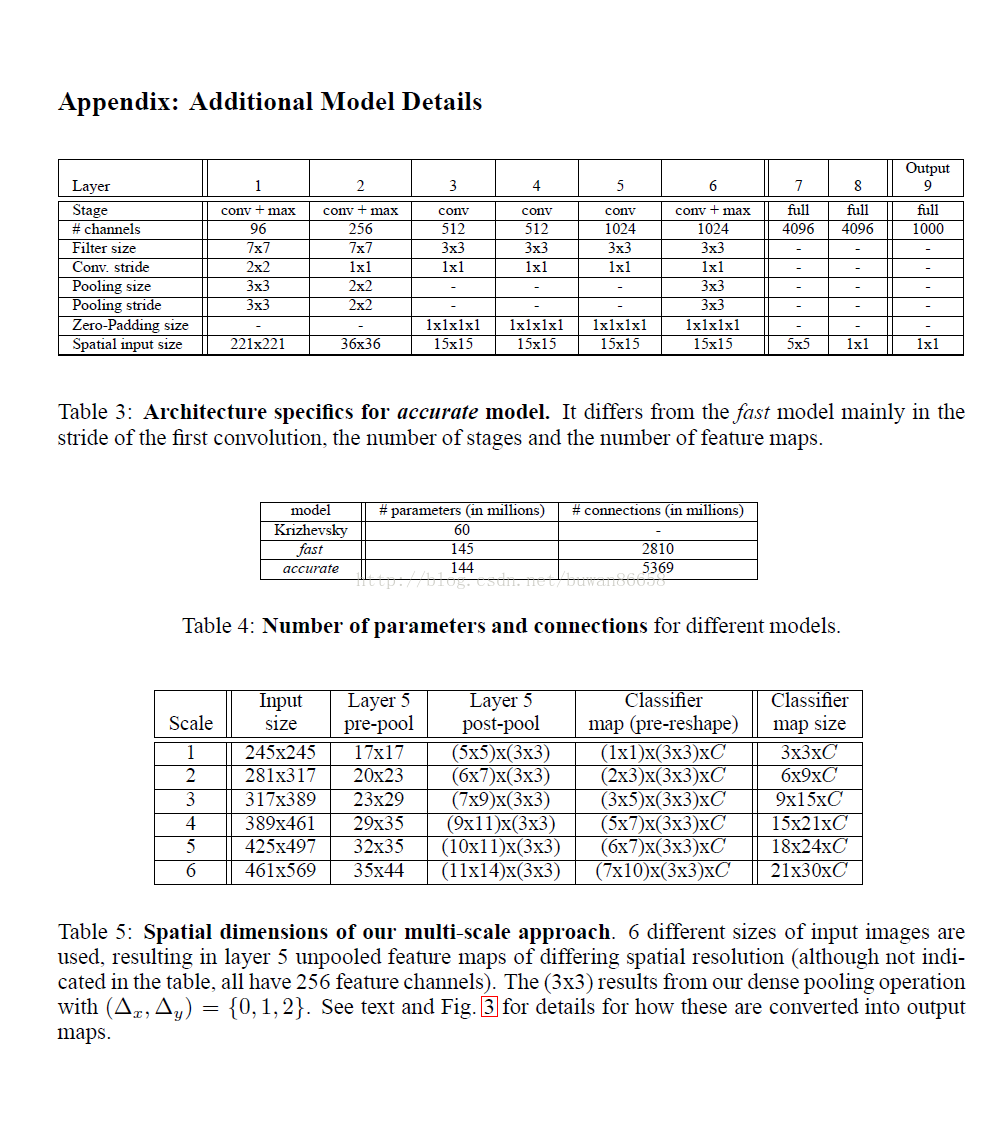


**6.讨论**

我们提出了一种可以用于分类，定位和检测的多尺度滑动窗口方法。我们将其应用于ILSVRC 2013数据集，目前在分类中排名第四，定位第一，检测第一。本文的第二个重要贡献是解释了ConvNets如何有效地用于检测和定位任务。这些从未在[15]中得到解决，因此我们是第一个解释如何在ImageNet 2012的背景下进行的。我们提出的方案涉及到对分类设计的网络的大量修改， 但清楚地表明，ConvNets有能力应对这些更具挑战性的任务。我们的定位方法赢得了2013年的ILSVRC比赛，并且明显优于2012和2013年的所有方法。检测模式在比赛中名列前茅，排名第一。我们提出了一个完整的方法，可以在共享一个共同的特征提取基础时执行不同的任务，直接从像素学到。

**我们的方法在一下几个方面任需要继续改善。（i）对于定位，我们目前并不支持整个网络;这样做可能会提高性能。（ii）我们使用l2损失，而不是直接优化测量性能的交叉联合（IOU）标准。交换损失应该是可能的，因为IOU仍然是可以区分的，只要有一些重叠。（iii）边界框的替代参数化可能有助于将输出关联起来，这将有助于网络训练。**

**附录：**



**引用：**

[1] J. Carreira, F. Li, and C. Sminchisescu. Object recognition by sequential figure-ground ranking. International  
journal of computer vision, 98(3):243–262, 2012.  
[2] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation,  
release 1. http://sminchisescu.ins.uni-bonn.de/code/cpmc/.  
[3] D. C. Ciresan, J.Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification.  
In CVPR, 2012.  
[4] M. Delakis and C. Garcia. Text detection with convolutional neural networks. In International Conference  
on Computer Vision Theory and Applications (VISAPP 2008), 2008.  
[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical  
Image Database. In CVPR09, 2009.  
[6] I. Endres and D. Hoiem. Category independent object proposals. In Computer Vision–ECCV 2010, pages  
575–588. Springer, 2010.  
[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. IEEE  
Transactions on Pattern Analysis and Machine Intelligence, 2013. in press.  
[8] C. Garcia and M. Delakis. Convolutional face finder: A neural architecture for fast and robust face  
detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004.  
[9] A. Giusti, D. C. Ciresan, J.Masci, L.M. Gambardella, and J. Schmidhuber. Fast image scanning with deep  
max-pooling convolutional neural networks. In International Conference on Image Processing (ICIP),  
2013.  
[10] R. Hadsell, P. Sermanet, M. Scoffier, A. Erkan, K. Kavackuoglu, U. Muller, and Y. LeCun. Learning  
long-range vision for autonomous off-road driving. Journal of Field Robotics, 26(2):120–144, February  
2009.  
[11] G. Hinton, N. Srivastave, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks  
by preventing co-adaptation of feature detectors. arXiv:1207.0580, 2012.  
[12] G. E. Hinton, A. Krizhevsky, and S. D.Wang. Transforming auto-encoders. In Artificial Neural Networks  
and Machine Learning–ICANN 2011, pages 44–51. Springer Berlin Heidelberg, 2011.  
[13] V. Jain, J. F. Murray, F. Roth, S. Turaga, V. Zhigulin, K. Briggman, M. Helmstaedter, W. Denk, and H. S.  
Seung. Supervised learning of image restoration with convolutional networks. In ICCV’07.  
[14] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for  
object recognition? In Proc. International Conference on Computer Vision (ICCV’09). IEEE, 2009.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks.  
In NIPS, 2012.  
[16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten  
digit recognition with a back-propagation network. In D. Touretzky, editor, Advances in Neural  
Information Processing Systems (NIPS 1989), volume 2, Denver, CO, 1990. Morgan Kaufman.  
[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition.  
Proceedings of the IEEE, 86(11):2278–2324, November 1998.  
[18] Y. LeCun, F.-J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance  
to pose and lighting. In Proceedings of CVPR’04. IEEE Press, 2004.  
[19] S.Manen, M. Guillaumin, and L. Van Gool. Prime object proposals with randomized prims algorithm. In  
International Conference on Computer Vision (ICCV), 2013.  
[20] O. Matan, J. Bromley, C. Burges, J. Denker, L. Jackel, Y. LeCun, E. Pednault, W. Satterfield, C. Stenard,  
and T. Thompson. Reading handwritten digits: A zip code recognition system. IEEE Computer, 25(7):59–  
63, July 1992.  
[21] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. Barbano. Toward automatic phenotyping of  
developing embryos from videos. IEEE Transactions on Image Processing, 14(9):1360–1371, September  
2005. Special issue on Molecular and Cellular Bioimaging.  
[22] S. Nowlan and J. Platt. A convolutional neural network hand tracker. pages 901–908, San Mateo, CA,  
1995. Morgan Kaufmann.  
[23] M. Osadchy, Y. LeCun, and M. Miller. Synergistic face detection and pose estimation with energy-based  
models. Journal of Machine Learning Research, 8:1197–1215, May 2007.  
[24] P. Sermanet, S. Chintala, and Y. LeCun. Convolutional neural networks applied to house numbers digit  
classification. In International Conference on Pattern Recognition (ICPR 2012), 2012.  
[25] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multistage  
feature learning. In Proc. International Conference on Computer Vision and Pattern Recognition  
(CVPR’13). IEEE, June 2013.  
[26] P. Sermanet and Y. LeCun. Traffic sign recognition with multi-scale convolutional networks. In Proceedings  
of International Joint Conference on Neural Networks (IJCNN’11), 2011.  
[27] G. Taylor, R. Fergus, G. Williams, I. Spiro, and C. Bregler. Pose-sensitive embedding by nonlinear nca  
regression. In NIPS, 2011.  
[28] G. Taylor, I. Spiro, C. Bregler, and R. Fergus. Learning invarance through imitation. In CVPR, 2011.  
[29] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object  
recognition. International Journal of Computer Vision, 104(2):154–171, 2013.  
[30] R. Vaillant, C. Monrocq, and Y. LeCun. Original approach for the localisation of objects in images. IEE  
Proc on Vision, Image, and Signal Processing, 141(4):245–250, August 1994.