arXiv:1311.2524v5 [cs.CV] 2014年

丰富的要素层次结构，用于准确的对象检测和语义分割

技术报告（v5）

Ross Girshick Jeff Donahue Trevor Darrell
Jitendra Malik加州大学伯克利分校
{RBG，jdonahue，特雷弗，马利克} @ eecs.berkeley.edu

摘要

在规范的PASCAL VOC数据集上测量的物体检测性能在过去几年中已经趋于稳定。表现最好的方法很复杂

通常将多个低级图像特征与高级上下文组合在一起的系统。在本文中，我们

提出了一种简单且可扩展的检测算法，相对于之前对VOC 2012的最佳结果，平均精度（mAP）提高了30%以上 - 实现了53.3%的mAP。我们的方法结合了两个关键见解：

（1）可以将大容量卷积神经网络（CNN）应用于自下而上的区域提议，以便定位和分割对象；（2）当标记的训练数据稀缺时，监督辅助任务的预训练，然后是特定领域的微调，可以显着提升性能。由于我们将区域提案与CNN结合起来，我们将方法称为R-CNN：具有CNN功能的区域。我们还将R-CNN与最近提出的基于类似CNN架构的滑动窗口检测器OverFeat进行了比较。我们发现R-CNN在200级ILSVRC2013检测数据集上大大优于OverFeat。完整系统的源代码可在以下位置获得 http://www.cs.berkeley.edu/~rbg/rcnn.

1. 介绍

特色重要。各种视觉识别任务的最后十年进展基于SIFT的使用[29]和HOG [7].但是如果我们看一下规范视觉识别任务的性能，PASCAL VOC物体检测[15]，人们普遍承认，2010 - 2012年的进展缓慢，通过建立集合系统和采用成功方法的微小变体获得了小幅增长。

SIFT和HOG是块状方向直方图，我们可以粗略地与V1中的复杂细胞相关联，这是灵长类动物视觉通路中的第一个皮层区域。但我们也知道识别发生在下游的几个阶段，这表明可能存在

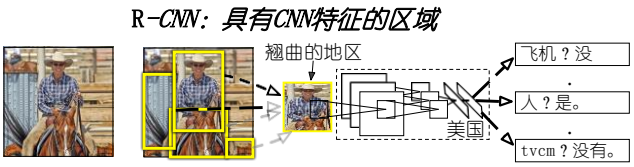

图1：对象检测系统概述。我们的系统（1）采用输入图像，（2）提取大约2000个自下而上区域提议，（3）使用大型卷积神经网络（CNN）计算每个提议的特征，然后（4）使用类对每个区域进行分类特定的线性SVM。R-CNN在PASCAL VOC 2010上实现了53.7%的平均精度（mAP）。为了进行比较，[39]使用相同的区域提案报告35.1%mAP，但使用空间金字塔和视觉词袋方法。流行的可变形零件模型的性能为33.4%。在200级ILSVRC2013检测数据集中，R-CNN的mAP为31.4%，比OverFeat有很大改进[34]，之前的最佳结果为24.3%。

用于计算功能的逻辑，多阶段过程，这些功能对于视觉识别更具信息性。

福岛的“neocognitron”[19]，一种用于模式识别的生物学启发的分层和移位不变模型，是这种过程的早期尝试。然而，新认知缺乏监督训练算法。以Rumelhart等人为基础。[33]，LeCun等。[26]表明通过反向传播的随机梯度下降对于训练卷积神经网络（CNN）是有效的，CNN是一类扩展新神经元的模型。

CNN在20世纪90年代被大量使用（例如，[27]），但随着支持向量机的兴起而失去了时尚。2012年，Krizhevsky等人。[25]通过在ImageNet大规模视觉识别挑战赛（ILSVRC）上显示更高的图像分类准确度，重新燃起了对CNN的兴趣[9，10].他们的成功来自于在120万张标记图像上训练大量CNN，以及LeCun CNN上的一些曲折（例如，max（x，0）整流非线性和“丢失”正则化）。

ImageNet结果的重要意义非常强烈

在ILSVRC 2012研讨会期间进行了辩论。中心问题可以归结为以下几点：ImageNet的CNN分类结果在多大程度上推广到PASCAL VOC挑战的目标检测结果？

我们通过弥合图像分类和对象检测之间的差距来回答这个问题。本文首次表明，与基于简单HOG类功能的系统相比，CNN可以在PASCAL VOC上实现更高的物体检测性能。为了实现这一结果，我们专注于两个问题：使用深层网络本地化对象，并仅使用少量带注释的检测数据来训练高容量模型。

与图像分类不同，检测需要在图像内定位（可能很多）对象。一种方法将定位框架化为回归问题。但是，Szegedy等人的工作。[38]与我们自己同时发现，这种策略在实践中可能表现不佳（他们报告VOC 2007的mAP为30.5%，而我们的方法实现的58.5%）。另一种方法是构建一个滑动窗口探测器。CNN已经以这种方式使用了至少二十年，通常是在受约束的对象类别上，例如面[32，40]和行人[35]。为了保持高空间分辨率，这些CNN通常仅具有两个卷积和池化层。我们还考虑采用滑动窗口方法。然而，在我们的网络中具有五个卷积层的单元在输入图像中具有非常大的感受域（195×195像素）和步幅（32×32像素），这使得在滑动窗口范例内的精确定位成为开放的技术。挑战。

相反，我们通过在“使用区域识别”范例内操作来解决CNN本地化问题[21]，已成功进行物体检测[39]和语义分割[5]。在测试时，我们的方法为输入图像生成大约2000个与类别无关的区域提议，使用CNN从每个提案中提取固定长度的特征向量，然后使用类别特定的线性SVM对每个区域进行分类。我们使用简单的技术（仿射图像变形）来计算来自每个区域建议的固定大小的CNN输入，而不管区域的形状如何。数字1 概述了我们的方法，并重点介绍了我们的一些结果。由于我们的系统将区域提案与CNN结合起来，我们称之为R-CNN方法：具有CNN功能的区域。

在本文的更新版本中，我们提供了R-CNN与最近提出的OverFeat的头对比[34通过在200级ILSVRC2013检测数据集上运行R-CNN检测系统。OverFeat使用滑动窗口CNN进行检测，到目前为止，它是ILSVRC2013检测中性能最佳的方法。我们证明R-CNN明显优于OverFeat，mAP为31.4%而24.3%。

检测中面临的第二个挑战是标记数据很少，目前可用的数量不足以培训大型CNN。该问题的传统解决方案是使用无监督的预训练，然后进行有监督的微调（例如，[35]）.本文的第二个主要贡献是显示对大型辅助数据集（ILSVRC）进行有监督的预训练，然后对小数据集（PASCAL）进行区域特异性微调，这是在数据时学习高容量CNN的有效范例。很稀缺。在我们的实验中，用于检测的微调将mAP性能提高了8个百分点。经过微调后，我们的系统在VOC 2010上实现了54%的mAP，而高度调整的基于HOG的可变形零件模型（DPM）则达到33%[17，20]。我们还将读者指向Donahue等人的同期作品。[12]，表明Krizhevsky的CNN可以作为黑盒特征提取器使用（没有微调），在几个识别任务上产生出色的表现，包括场景分类，细粒度子分类和域适应。

我们的系统也很有效率。唯一的类特定计算是相当小的矩阵向量乘积和贪婪的非最大抑制。这种计算属性来自于所有类别共享的特征，并且比以前使用的区域特征的维度低两个数量级（参见[39]）.了解我们的方法的失效模式对于改进它也是至关重要的，因此我们报告了Hoiem等人的检测分析工具的结果。[23].作为此分析的直接结果，我们证明了一种简单的边界框回归方法可以显着减少错误定位，这是主要的错误模式。在开发技术细节之前，我们注意到因为R-CNN在区域上运行，将其扩展到语义分割的任务是很自然的。通过微小的修改，我们还在PASCAL VOC分割任务中获得了有竞争力的结果，VOC 2011测试集的平均分割准确度为47.9%。

2. 使用R-CNN进行物体检测

我们的物体检测系统由三个模块组成。第一个生成与类别无关的区域提案。这些建议定义了我们的探测器可用的候选检测集。第二个模块是一个大型卷积神经网络，它从每个区域中提取固定长度的特征向量。第三个模块是一组特定于类的线性SVM。在本节中，我们将介绍每个模块的设计决策，描述其测试时间使用情况，详细说明其参数的学习方式，并在PASCAL VOC 2010-12和ILSVRC2013上显示检测结果。

2.1. 模块设计

地区提案。最近的各种论文提供了用于生成与类别无关的区域提议的方法。

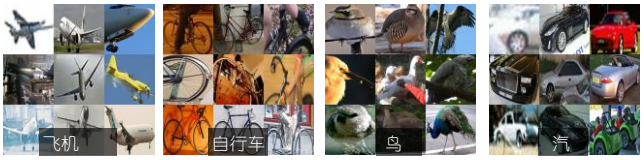

图2：VOC 2007列车的扭曲训练样本。

例子包括：对象[1]，选择性搜索[39]，与类别无关的对象提案[14]，约束参数最小割（CPMC）[5]，多尺度组合分组[3]和Cires¸an等人。[6]通过将CNN应用于规则间隔的方形作物来检测有丝分裂细胞，这是区域提案的特例。虽然R-CNN对特定区域提议方法不可知，但我们使用选择性搜索来实现与先前检测工作的受控比较（例如，[39,41]）.

特征提取。我们使用Caffe从每个区域提案中提取4096维特征向量[24] Krizhevsky等人描述的CNN的实施。[25].通过向前传播平均减去的227 227 RGB图像通过五个卷积层和两个完全连接的层来计算特征。我们推荐读者[24,25]了解更多网络架构细节。

为了计算区域提议的特征，我们必须首先将该区域中的图像数据转换为与CNN兼容的形式（其架构需要输入固定的227 227像素大小）。在我们任意形状区域的许多可能变换中，我们选择最简单的。无论候选区域的大小或宽高比如何，我们都会将其周围的紧密边界框中的所有像素扭曲到所需的大小。在变形之前，我们扩展紧密的边界框，以便在扭曲的尺寸上，在原始框周围有正好p像素的扭曲图像上下文（我们使用p = 16）。数字2 显示了翘曲训练区域的随机抽样。翘曲的替代方案在附录中讨论A.

2.2. 测试时间检测

在测试时，我们对测试图像进行选择性搜索以提取大约2000个区域提议（我们在所有实验中使用选择性搜索的“快速模式”）。我们对每个提案进行扭曲并向前传播它以通过CNN来计算功能。然后，对于每个类，我们使用针对该类训练的SVM对每个提取的特征向量进行评分。给定图像中的所有得分区域，我们应用贪婪的非最大抑制（对于每个类独立），如果它具有交叉结合（IoU）重叠且具有大于学习阈值的更高得分选定区域，则拒绝该区域。

运行时分析。两个属性使检测有效。首先，所有类别共享所有CNN参数。第二，由CNN计算的特征向量与其他常见方法相比，它们是低维的，例如带有视觉字编码包的空间金字塔。UVA检测系统中使用的功能[39例如，比我们大两个数量级（360k对4k维）。

这种共享的结果是计算区域提议和特征（在GPU上为13s /图像或在CPU上为53s /图像）所花费的时间在所有类上摊销。唯一的类特定计算是要素与SVM权重和非最大抑制之间的点积。在实践中，图像的所有点积都被分批到单个矩阵 - 矩阵产品中。特征矩阵通常为2000 4096，SVM权重矩阵为4096 N，其中N是类的数量。

该分析表明，R-CNN可以扩展到数千个对象类，而无需采用近似技术，例如散列。即使有100k类，在现代多核CPU上产生的矩阵乘法只需10秒。这种效率不仅仅是使用区域提案和共享功能的结果。由于其高维特征，UVA系统将慢两个数量级，而仅需要134GB内存来存储100k线性预测器，而我们的低维特征仅需1.5GB。

将R-CNN与Dean等人最近的工作进行对比也很有意思。使用DPM和散列进行可扩展检测[8].他们报告说，当引入10k干扰分类时，每张图像的运行时间为5分钟，VOC 2007的mAP约为16％。通过我们的方法，10k探测器可以在CPU上运行大约一分钟，并且由于没有近似值，因此mAP将保持在59％（部分3.2).

2.3. 训练

监督预训练。我们仅使用图像级注释在大型辅助数据集（ILSVRC2012分类）上对CNN进行有区别的预训练（边界框标签不可用于此数据）。使用开源Caffe CNN库进行预训练[24].简而言之，我们的CNN几乎与Krizhevsky等人的表现相符。[25]，在ILSVRC2012分类验证集上获得高出2.2个百分点的前1错误率。这种差异是由于培训过程的简化。

特定领域的微调。为了使我们的CNN适应新任务（检测）和新域（扭曲的提议窗口），我们仅使用扭曲区域提议继续CNN参数的随机梯度下降（SGD）训练。除了用随机初始化的（N + 1）路分类层替换CNN的ImageNet特定的1000路分类层（其中N是对象类的数量，加上背景为1），CNN架构保持不变。对于VOC，N = 20，对于ILSVRC2013，N = 200.我们对所有区域提案进行处理

≥ 0.5 IoU与地面实况框重叠，作为该框类的正面，其余为负面。我们以0.001的学习率（初始预训练率的1/10）启动SGD，这允许微调进行而不会破坏初始化。在每次SGD迭代中，我们统一采样32个正窗口（在所有类别上）和96个背景窗口以构建一个小批量的大小 我们将采样偏向正窗口，因为与背景相比它们非常罕见。

对象类别分类器。考虑训练二元分类器来检测汽车。很明显，紧紧包围汽车的图像区域应该是一个积极的例子。同样，很明显，与汽车无关的背景区域应该是一个反面的例子。不太清楚的是如何标记与汽车部分重叠的区域。我们使用IoU重叠阈值解决此问题，低于该阈值将区域定义为负数。通过0,0.1，...的网格搜索选择重叠阈值0.3。..，验证集合为0.5。我们发现仔细选择此阈值非常重要。将其设置为0.5，如[39]，将mAP降低了5个点。同样，将其设置为0会使mAP降低4个点。正例被简单地定义为每个类的基础真值边界框。

一旦提取了特征并应用了训练标签，我们就会优化每个类的一个线性SVM。由于训练数据太大而无法记忆，我们采用标准的硬负采矿方法[17, 37].硬负采矿快速收敛，并且在实践中，mAP仅在一次通过所有图像后停止增加。

在附录中B 我们讨论为什么在微调和SVM训练中对正面和负面例子的定义不同。我们还讨论了训练检测SVM所涉及的权衡，而不是简单地使用来自微调CNN的最终softmax层的输出。

2.4. PASCAL VOC 2010-12的结果

遵循PASCAL VOC最佳实践[15]，我们验证了VOC 2007数据集上的所有设计决策和超参数（Section3.2).对于VOC 2010-12数据集的最终结果，我们对VOC 2012列车上的CNN进行了微调，并优化了我们在VOC 2012 trainval上的检测SVM。我们仅针对两种主要算法变体（包括和不包含边界框回归）将测试结果提交给评估服务器一次。

表1 显示VOC 2010的完整结果。我们将我们的方法与四个强基线进行比较，包括SegDPM [18]，它将DPM探测器与语义分割系统的输出结合起来[4]并使用额外的检测器间上下文和图像分类器重新计算。最相关的是Uijlings等人的UVA系统。[39]，因为我们的系统使用相同的区域提议算法。为了对区域进行分类，他们的方法构建了一个四级空间金字塔并用它填充密集采样的SIFT，扩展的OpponentSIFT和RGB-SIFT描述符，每个矢量用4000字的码本量化。使用直方图交叉核SVM执行分类。与他们的多特征，非线性核SVM方法相比，我们实现了mAP的大幅改进，从mAP的35.1%到53.7%，同时也更快（Section2.2).我们的方法在VOC 2011/12测试中实现了类似的性能（53.3%mAP）。

2.5. ILSVRC2013检测结果

我们使用与PASCAL VOC相同的系统超参数在200级ILSVRC2013检测数据集上运行R-CNN。我们遵循相同的协议，仅向ILSVRC2013评估服务器提交测试结果两次，一次使用，一次没有边界框回归。

数字3 将R-CNN与ILSVRC 2013竞赛中的参赛作品以及赛后OverFeat成绩进行比较[34].R-CNN的mAP达到31.4%，远远高于OverFeat的24.3%的第二好成绩。为了了解类别上的AP分布情况，还提供了箱形图，并在表格末尾列出了每类AP的表格。8.大多数竞争性提交（OverFeat，NEC-MU，UvA-Euvision，Toronto A和UIUC-IFP）使用卷积神经网络，表明CNN如何应用于物体检测存在重大差异，导致结果大不相同。

在节中4，我们概述了ILSVRC2013检测数据集，并提供了有关在其上运行R-CNN时所做选择的详细信息。

3. 可视化，消融和错误模式

3.1. 可视化学习的功能

第一层过滤器可直接显示，易于理解[25].它们捕捉定向边缘和对手颜色。了解后续层更具挑战性。Zeiler和Fergus提出了一种视觉上具有吸引力的反卷积方法[42].我们提出了一种简单（和互补）的非参数方法，可直接显示网络学到的内容。

我们的想法是在网络中挑出一个特定的单元（特征）并使用它，就像它本身就是一个物体探测器一样。也就是说，我们在一大组保留区域提案（约1000万）中计算单位的激活，从最高到最低激活对提案进行排序，执行非最大抑制，然后显示得分最高的区域。我们的方法通过准确显示它所触发的输入，让所选单元“说出自己”。我们避免求平均值以便查看不同的视觉模式并深入了解由单元计算的不变性。

VOC 2010测试	航空	自行车	鸟	船	瓶子	总线	汽车	猫	椅子	牛	表	狗	马	姆福尔	人	厂	羊	沙发	培养	电视	地图
DPM v5 [20]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
乌瓦［39]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
区域［41]	65.0	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54.0	45.9	39.7
SegDPM [18]†	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32.0	30.5	56.4	57.2	65.9	27.0	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	**71.8**	**65.8**	**53.0**	**36.8**	**35.9**	**59.7**	**60.0**	**69.9**	**27.9**	**50.6**	**41.4**	**70.0**	**62.0**	**69.0**	**58.1**	**29.5**	**59.4**	**39.3**	**61.2**	**52.4**	**53.7**

表1：VOC 2010测试的检测平均精度（%）。R-CNN与UVA和Regionlet最直接可比，因为所有方法都使用选择性搜索区域提议。边界框回归（BB）在章节中描述C. 在出版时，SegDPM是PASCAL VOC排行榜的最佳表现者。 †DPM和SegDPM使用其他方法未使用的上下文重新绑定。

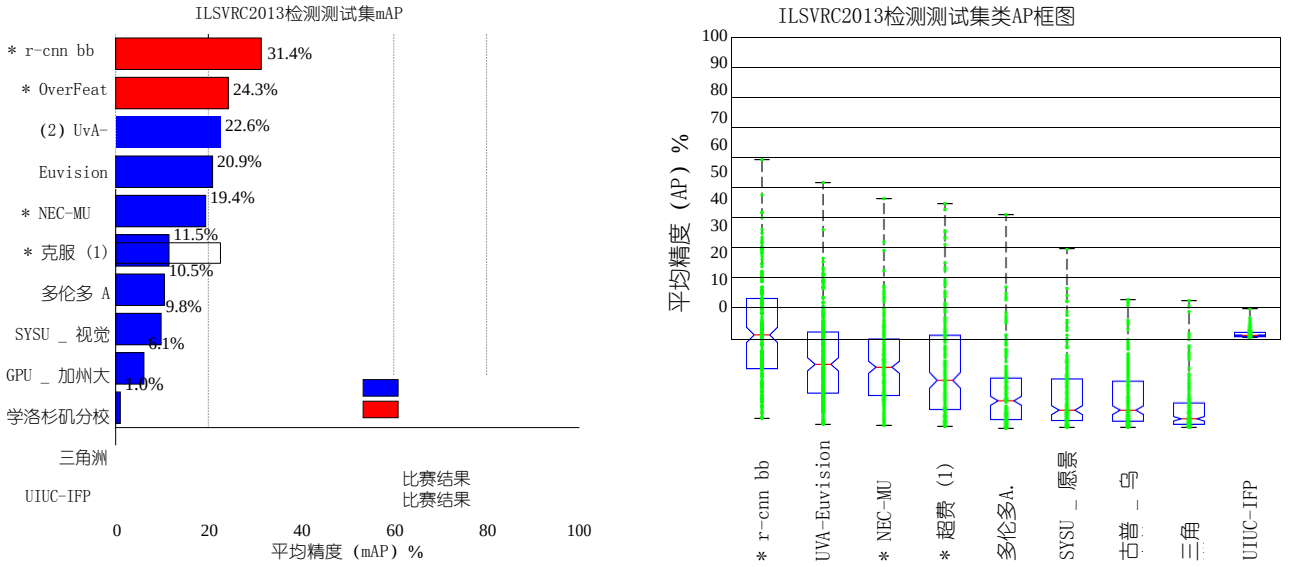

图3：（左）ILSVRC2013检测测试装置的平均平均精度。方法之前*使用外部训练数据（在所有情况下来自ILSVRC分类数据集的图像和标签）。（右）每个方法的200个平均精度值的箱形图。没有显示赛后OverFeat结果的方框图，因为每类AP尚不可用（R-CNN的每类AP在表中8 并且还包含在上传到arXiv.org的技术报告来源中;见R-CNN-ILSVRC2013-APs.txt）。红线表示中位数AP，方框底部和顶部是第25和第75百分位数。晶须延伸到每种方法的最小和最大AP。每个AP在胡须上绘制为绿点（最好以数字方式使用缩放查看）。

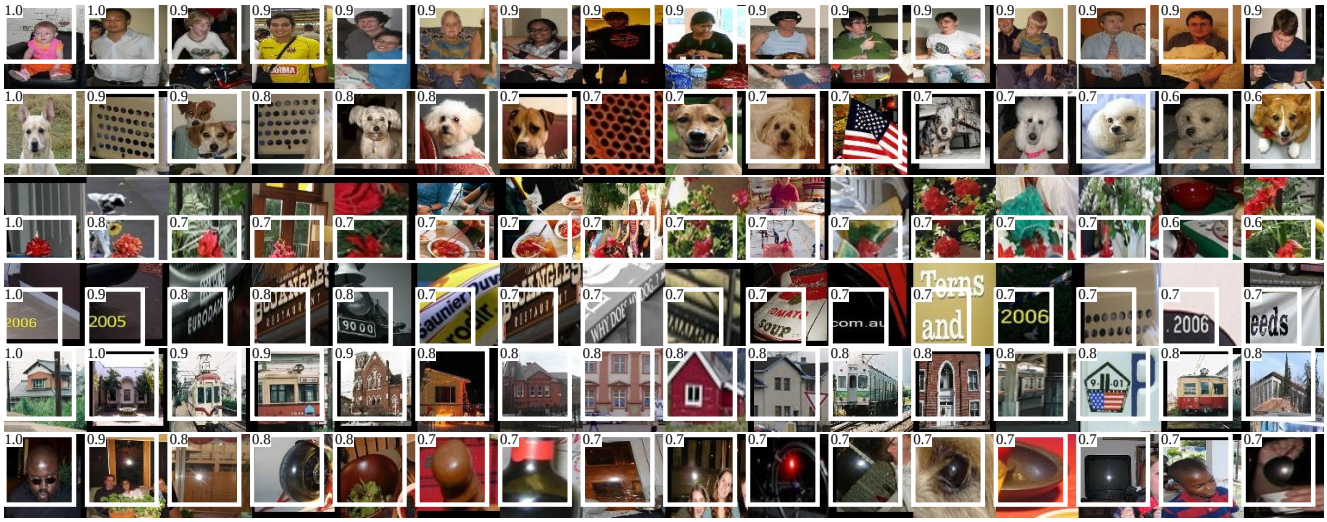

图4：六个池5单元的顶部区域。接收字段和激活值以白色绘制。一些单元与概念对齐，例如人（第1行）或文本（4）。其他单位捕获纹理和材质属性，例如点阵列（2）和镜面反射（6）。

VOC 2007测试	航空	自行车	鸟	船	瓶子	总线	汽车	猫	椅子	牛	表	狗	马	姆福尔	人	厂	羊	沙发	培养	电视	地图
R-CNN游泳池$_5$	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc$_6$	59.3	61.8	43.1	34.0	25.1	53.1	60.6	52.8	21.7	47.8	42.7	47.8	52.5	58.5	44.6	25.6	48.3	34.0	53.1	58.0	46.2
R-CNN fc$_7$	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT池$_5$	58.2	63.3	37.9	27.6	26.1	54.1	66.9	51.4	26.7	55.5	43.4	43.1	57.7	59.0	45.8	28.1	50.8	40.6	53.1	56.4	47.3
R-CNN FT fc$_6$	63.5	66.0	47.9	37.7	29.9	62.5	70.2	60.2	32.0	57.9	47.0	53.5	60.1	64.2	52.2	31.3	55.0	50.0	57.7	63.0	53.1
R-CNN FT fc$_7$	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN FT fc$_7$ BB	**68.1**	**72.8**	**56.8**	**43.0**	**36.8**	**66.3**	**74.2**	**67.6**	**34.4**	**63.5**	**54.5**	**61.2**	**69.1**	**68.6**	**58.7**	**33.4**	**62.9**	**51.1**	**62.5**	**64.8**	**58.5**
DPM v5 [20]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
DPM ST [28]	23.8	58.2	10.5	8.5	27.1	50.4	52.0	7.3	19.2	22.8	18.1	8.0	55.9	44.8	32.4	13.3	15.9	22.8	46.2	44.9	29.1
DPM HSC [31]	32.2	58.3	11.5	16.3	30.6	49.9	54.8	23.5	21.5	27.7	34.0	13.7	58.1	51.6	39.9	12.4	23.5	34.4	47.4	45.2	34.3

表2：VOC 2007测试的检测平均精度（%）。第1-3行显示没有微调的R-CNN性能。第4-6行显示CNN在ILSVRC 2012上进行预训练的结果，然后在VOC 2007 trainval上进行微调（FT）。第7行包括一个简单的边界框回归（BB）阶段，可以减少本地化错误（SectionC).第8-10行将DPM方法作为强基线。第一个仅使用HOG，而接下来的两个使用不同的特征学习方法来增强或替换HOG。

VOC 2007测试	航空	自行车	鸟	船	瓶子	总线	汽车	猫	椅子	牛	表	狗	马	姆福尔	人	厂	羊	沙发	培养	电视	地图
R-INNN T-Net	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-INNN T-Net BB	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5
R-CNN O-Net	71.6	73.5	58.1	42.2	39.4	70.7	76.0	74.5	38.7	71.0	56.9	74.5	67.9	69.6	59.3	**35.7**	62.1	64.0	66.5	**71.2**	62.2
R-CNN O-Net BB	**73.4**	**77.0**	**63.4**	**45.4**	**44.6**	**75.1**	**78.1**	**79.8**	**40.5**	**73.7**	**62.2**	**79.4**	**78.1**	**73.1**	**64.2**	35.6	**66.8**	**67.2**	**70.4**	71.1	**66.0**

表3：两种不同CNN架构的VOC 2007测试的检测平均精度（%）。前两行是Table的结果2 使用Krizhevsky等人的架构（T-Net）。第三和第四行使用最近提出的Simonyan和Zisserman（O-Net）的16层架构[43].

我们从层池$_5$可视化单元，这是网络的第五个和最后一个卷积层的最大池输出。池$_5$特征图是6 6 256 = 9216维。忽略边界效应，每个池$_5$单元在原始227x227像素输入中具有195 195像素的感受域。中央游泳池$_5$单元具有近乎全局的视野，而靠近边缘的一个具有较小的剪切支撑。

图中的每一行4 显示我们在VOC 2007 trainval上微调的CNN的池$_5$单元的前16次激活。256个功能独特单元中的六个可视化（附录D 包括更多）。这些单位是提示显示网络学习的代表性样本。在第二行中，我们看到一个在狗脸和点阵列上发射的单位。对应于第三行的单元是红色斑点检测器。还有人脸探测器和更抽象的图案，如文字和带窗户的三角形结构。该网络似乎学习了一种表示，该表示将少量的类调整特征与形状，纹理，颜色和材料属性的分布式表示相结合。随后的完全连接层fc$_6$ 能够模拟这些丰富特征的大量组合。

3.2. 消融研究

性能逐层，无需微调。为了了解哪些层对检测性能至关重要，我们分析了每个CNN最后三层的VOC 2007数据集的结果。层池$_5$在章节中简要描述3.1. 最后两层总结如下。

层fc$_6$ 完全连接到池$_5$。为了计算特征，它将4096 9216权重矩阵乘以池$_5$特征图（重新整形为9216维向量）和
然后添加一个偏差矢量。该中间矢量是分量半波整流的（x←max（0，x））。

层fc$_7$ 是网络的最后一层。这是实施通过将由fc$_6$ 计算的特征乘以4096 4096权重矩阵，并类似地添加偏差矢量并应用半波整流来进行分析。

我们首先查看CNN的结果，而不对PASCAL进行微调，即所有CNN参数仅在ILSVRC 2012上预先训练。逐层分析性能（表2 第1-3行显示fc$_7$ 的特征比fc$_6$的特征更加普遍。这意味着可以在不降低mAP的情况下移除29%或约1680万个CNN参数。更令人惊讶的是，即使仅使用6%的CNN参数计算池$_5$特征，同时去除fc$_7$ 和fc$_6$ 也会产生非常好的结果。CNN的大部分代表性力量来自其卷积层，而不是来自更大的密集连接层。该发现表明，通过仅使用CNN的卷积层，在HOG的意义上计算任意大小的图像的密集特征图的潜在效用。这种表示将使得能够在池$_5$特征之上使用滑动窗口检测器（包括DPM）进行实验。

逐层执行，具有微调功能。我们现在看看我们的CNN的结果，经过微调它的pa-

关于VOC 2007 trainval的电表。改善是惊人的（表2 第4-6行：微调将mAP提高8.0个百分点至54.2%。对于fc_6和fc_7而言，微调的提升要比对于池$_5$大得多，这表明从ImageNet学到的池$_5$特征是通用的，并且大部分改进都是从学习中获得的。特定于域的非线性分类器。

与最近的特征学习方法的比较。在PAS-CAL VOC检测中尝试了相对较少的特征学习方法。我们看一下基于可变形零件模型的两种最新方法。作为参考，我们还包括基于标准HOG的DPM的结果[20].

第一个DPM特征学习方法，DPM ST [28]，使用“草图标记”概率的直方图增强HOG特征。直观地，草图标记是通过图像块中心的轮廓的紧密分布。通过随机森林在每个像素处计算草图标记概率，该随机森林被训练以将35个35像素斑块分类为150个草图标记或背景之一。

第二种方法，DPM HSC [31]，用稀疏代码（HSC）的直方图替换HOG。为了计算HSC，使用100 7 7像素（灰度）原子的学习字典在每个像素处求解稀疏码激活。由此产生的激活以三种方式（全波和两波）进行校正，空间合并，单位f_2归一化，以及然后进行功率变换（$x \leftarrow \text{sign}(x) \cdot |x|^{\alpha}$）。

所有R-CNN变体都强于三个DPM基线（表2 第8-10行，包括使用特征学习的两个。与仅使用HOG功能的DPM的最新版本相比，我们的mAP高出20多个百分点：54.2%对比33.7% - 相对改善率为61%。HOG和草图标记的组合比单独的HOG产生2.5 mAP点，而HSC比HOG提高4 mAP点（内部与其私有DPM基线进行比较 - 都使用DPM的非公开实现，其表现不如开源版本[20]).这些方法分别实现了29.1%和34.3%的mAP。

3.3. 网络架构

本文中的大多数结果都使用了Krizhevsky等人的网络架构。[25].但是，我们发现架构的选择对R-CNN检测性能有很大影响。在表中3 我们使用Simonyan和Zisserman最近提出的16层深度网络显示了VOC 2007测试的结果[43].该网络是最近ILSVRC 2014分类挑战中表现最佳的网络之一。网络具有均匀的结构，由13层33个卷积核组成，其中散布着5个最大池层，并且顶部有3个完全连接的层。我们称这个网络为“O-Net”对于OxfordNet和基线为TorontoNet的“T-Net”。

为了在R-CNN中使用O-Net，我们从Caffe Model Zoo下载了VGG ILSVRC 16层模型的公开预先训练的网络权重。[1] 然后，我们使用与T-Net相同的协议对网络进行微调。唯一的区别是为了适应GPU内存，需要使用更小的微型计算机（24个示例）。结果见表3 表明带有O-Net的R-CNN大幅优于R-CNN和T-Net，使mAP从58.5%增加到66.0%。然而，在计算时间方面存在相当大的缺点，O-Net的前向传输大约比T-Net长7倍。

3.4. 检测错误分析

我们应用了Hoiem等人的优秀检测分析工具。[23]为了揭示我们方法的错误模式，了解微调如何更改它们，以及查看我们的错误类型与DPM的比较。分析工具的完整摘要超出了本文的范围，我们鼓励读者参考[23]了解一些更精细的细节（例如“规范化的AP”）。由于分析最好在相关图的背景下被吸收，因此我们在图的标题内进行讨论5 和图6.

3.5. 边界框回归

基于错误分析，我们实现了一种简单的方法来减少本地化错误。受到DPM中使用的边界框回归的启发[17]，我们训练一个线性回归模型来预测一个新的检测窗口，给出选择性搜索区域提议的池$_5$功能。详细内容见附录C. 结果见表1，表2和图5 表明这种简单的方法修复了大量错误定位的检测，将mAP提高了3到4个点。

3.6. 定性结果

ILSVRC2013的定性检测结果如图所示8 和图9 在论文的最后。从val$_2$组中随机采样每个图像，并且显示来自所有检测器的所有检测，精度大于0.5。请注意，这些都不是策划的，并给出了实际的探测器的实际印象。更多定性结果如图所示10 和图11，但这些已被策划。我们选择了每张图片，因为它包含有趣，令人惊讶或有趣的结果。此外，还显示了精度大于0.5的所有检测。

4. ILSVRC2013检测数据集

在节中2 我们在ILSVRC2013检测数据集上提供了结果。该数据集不如同质

[1] https://github.com/BVLC/caffe/wiki/Model-Zoo

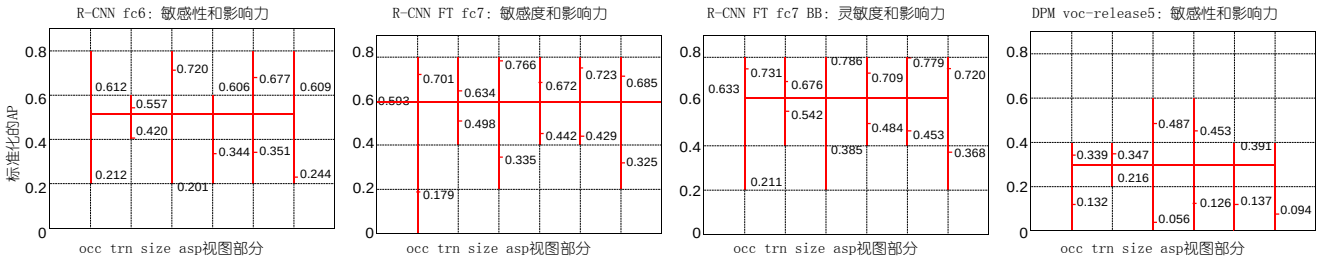

图6：对象特征的敏感度。每个图显示归一化AP的平均值（超过类）（见[23]）用于六个不同对象特征（遮挡，截断，边界框区域，纵横比，视点，部分可见性）内的最高和最低性能子集。我们显示了我们的方法（R-CNN）的图表，有和没有微调（FT）和边界框回归（BB）以及DPM voc-release5。总的来说，微调不会降低灵敏度（最大值和最小值之间的差异），但几乎可以显着改善几乎所有特性的最高和最低性能子集。这表明微调不仅仅是简单地改进宽高比和边界框区域中性能最低的子集，因为人们可能会根据我们如何扭曲网络输入进行猜测。相反，微调可以提高所有特征的鲁棒性，包括遮挡，截断，视点和零件可见性。

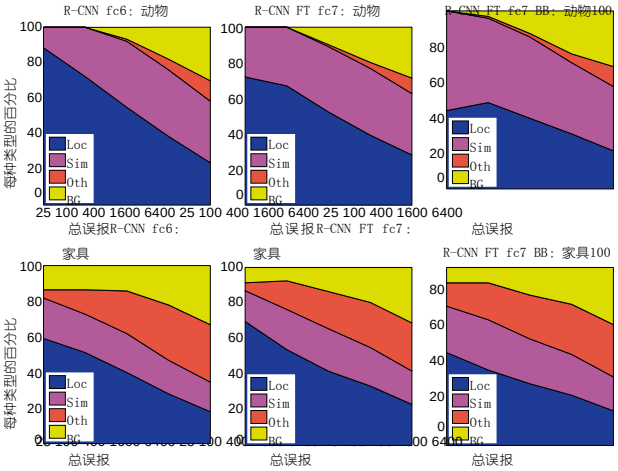

图5：排名靠前的假阳性（FP）类型的分布。每个图显示FP类型的演变分布，因为按照得分降低的顺序考虑更多的FP。每个FP分为4种类型中的1种：Loc-poor定位（IoU重叠检测，正确等级在0.1和0.5之间，或重复）；模拟与类似的混淆；与不同对象类别的混淆；BG-a FP在背景上发射。与DPM相比（见[23]），我们的错误显然更多是由于本地化不良而不是与背景或其他对象类混淆，这表明CNN特征比HOG更具辨别力。松散定位可能是由于我们使用自下而上区域提议以及从预训练CNN进行全图像分类所学到的位置不变性。第三列显示了我们的简单边界框回归方法如何修复许多本地化错误。

PASCAL VOC，需要选择如何使用它。由于这些决定非常重要，我们将在本节中介绍它们。

4.1. 数据集概述

ILSVRC2013检测数据集分为三组：train（395,918），val（20,121）和test（40,152），其中每组中的图像数量在括号中。该val和测试分割是从相同的图像分布中提取的。这些图像是类似场景的，并且与PASCAL VOC图像的复杂性（物体数量，杂波量，姿势可变性等）相似。val和测试拆分被详尽地注释，这意味着在每个图像中，来自所有200个类的所有实例都用边界框标记。该

相比之下，火车组来自ILSVRC2013的分类 -

*图*像分布。这些图像具有更多变化的复杂性，并且偏向于单个中心对象的图像。与val和测试不同，火车图像（由于它们的数量很大）没有被详尽地注释。在任何给定的火车图像中，来自200个类的实例可以标记或不标记。除了这些图像集之外，每个类还有一组额外的负像。负面图像

手动检查以验证它们不包含

他们相关课程的任何实例。在这项工作中没有使用负片图像集。有关如何收集和注释ILSVRC的更多信息，请参见[11, 36].

这些分裂的性质为训练R-CNN提供了许多选择。火车图像不能用于硬负挖掘，因为注释并非详尽无遗。负面例子应该从哪里来？此外，火车图像具有与val和测试不同的统计数据。是否应该使用火车图像，如果使用，在何种程度上？虽然我们还没有彻底评估大量的选择，但我们根据之前的经验提出了看似最明显的路径。

我们的总体策略是严重依赖val集并使用一些列车图像作为正例的辅助来源。为了使用val进行训练和验证，我们将其分成大致相同大小的“val_1”和“val_2”集。由于某些类在val中的示例非常少（最小的只有31而一半只有少于110），因此生成近似类平衡的分区非常重要。为此，生成了大量候选分割并且具有最小相对最小分割

选择了班级不平衡。[2] 每个候选分割是通过使用类别计数作为特征聚类val图像，然后是可以改善分割平衡的随机本地搜索来生成的。这里使用的特定分裂具有约11％的最大相对不平衡和4％的中值相对不平衡。val_1/ val_2 拆分和用于生成它们的代码将公开提供，以允许其他研究人员比较他们在本报告中使用的val拆分方法。

4.2. 地区提案

我们遵循用于在PASCAL上检测的相同区域提议方法。选择性搜索[39在val_1，val_2和测试（但不在火车上的图像上）的每个图像上以“快速模式”运行。需要进行一个小修改来处理选择性搜索不是尺度不变的事实，因此产生的区域数量取决于图像分辨率。ILSVRC图像大小范围从非常小到几个几百万像素，因此我们在运行选择性搜索之前将每个图像调整为固定宽度（500像素）。在val上，选择性搜索导致每个图像平均有2403个区域提议，所有地面实况边界框的回忆率为91.6％（0.5 IoU阈值）。此次召回明显低于PASCAL，其约为98％，表明该地区提案阶段有很大的改进空间。

4.3. 培训数据

对于训练数据，我们形成了一组图像和方框，其中包括来自val_1 的所有选择性搜索和地面实况框以及来自火车的每类最多N个地面实况框（如果一个类别少于N个地面 - 火车中的真值箱，然后我们把所有这些都带走了）。我们将这个图像和框的数据集称为val_1+$train_n$ 。在消融研究中，我们在val_2 上显示N0,500,1000的mAP（Section4.5). ∈ { }

R-CNN中的三个程序需要培训数据：（1）CNN微调，（2）检测器SVM训练，（3）边界框回归训练。使用完全相同的方法在val + train上运行CNN微调以进行50k SGD迭代 N 用于PASCAL的设置。使用Caffe对单个NVIDIA Tesla K20进行微调需要13个小时。对于SVM训练，来自val_1+$train_n$ 的所有地面实况框用作其各自类别的正例。对来自val_1的5000个图像的随机选择子集进行硬阴性采矿。最初的实验表明，从所有val_1挖掘负数而不是5000个图像子集（大约一半），只得到了mAP降低0.5个百分点，同时将SVM培训时间缩短一半。没有任何负面例子

[2]相对不平衡测量为a | − |b / (a + b) 其中a和b是分裂的每一半中的类别数。

因为注释并非详尽无遗而进行训练。未使用额外的验证负片图像集。边界框回归量训练为val_1。

4.4. 验证和评估

在将结果提交给评估服务器之前，我们使用上述训练数据验证了数据使用选择以及微调和边界框回归对val_2 集的影响。所有系统超参数（例如，SVM C超参数，区域扭曲中使用的填充，NMS阈值，边界框回归超参数）被固定在用于PAS-CAL的相同值。毫无疑问，这些超参数选择中的一些对于ILSVRC来说略微不理想，但是这项工作的目的是在没有大量数据集调整的情况下在ILSVRC上产生初步的R-CNN结果。在选择val_2上的最佳选择后，我们将两个结果文件提交给ILSVRC2013评估服务器。第一次提交没有边界框回归，第二次提交是边界框回归。对于这些提交，我们扩展了SVM和boundingbox回归训练集，分别使用val + $train_{1k}$ 和val。我们使用了在val_1+列车$_{1k}$ 上微调的CNN，以避免重新运行微调和特征计算。

4.5. 消融研究

表4 显示了不同数量的训练数据，微调和边界框回归效果的消融研究。第一个观察结果是val上的mAP_2与测试中的mAP非常接近。这让我们相信，mAP on val_2 是测试集性能的良好指标。第一个结果，20.9％，是R-CNN使用在ILSVRC2012分类数据集上预训练的CNN（没有微调）实现的，并且可以访问val中的少量训练数据$_1$ （回想一半） val中的类$_1$ 有15到55个例子。将训练集扩展到val_1+ $train_n$ 可将性能提高到24.1％，N = 500和N = 1000之间基本没有差别。使用以下示例对CNN进行微调

然而，只有val_1 适度改善了26.5％

由于少数积极的训练样例，可能存在严重的过度拟合。将微调设置扩展为val_1+列车$_{1k}$，从列车组中每班增加1000个正例，有助于显着提升mAP至29.7％。边界框回归将结果改善至31.0％，这是在PASCAL中观察到的较小的相对增益。

4.6. 与OverFeat的关系

R-CNN和OverFeat之间存在一种有趣的关系：OverFeat可以（大致）看作是R-CNN的一个特例。如果要替换选择性搜索区域

测试集	val_2	val_2	val_2	val_2	val_2	val_2	测试	测试
SVM训练集	val_1	VAL_1+列车$_{.5k}$	VAL_1+列车$_{1k}$	VAL_1+列车$_{1k}$	VAL_1+列车$_{1k}$	VAL_1+列车$_{1k}$	VAL +火车$_{1k}$	VAL +火车$_{1k}$
CNN微调套装	n/a	n/a	n/a	val_1	VAL_1+列车$_{1k}$	VAL_1+列车$_{1k}$	VAL_1+列车$_{1k}$	VAL_1+列车$_{1k}$
bbox reg set	n/a	n/a	n/a	n/a	n/a	val_1	n/a	瓦尔
CNN功能层	fc_6	fc_6	fc_6	fc_7	fc_7	fc_7	fc_7	fc_7
地图	20.9	24.1	24.1	26.5	29.7	**31.0**	30.2	**31.4**
中位数AP	17.7	21.0	21.4	24.8	29.2	**29.6**	29.0	**30.3**

表4：ILSVRC2013对数据使用选择，微调和边界框回归的消融研究。

具有规则方形区域的多尺度金字塔的建议，并将每类边界框回归量更改为单个边界框回归量，然后系统将非常相似（模拟它们如何训练的一些潜在显着差异：CNN检测微调，使用SVM等）。值得注意的是，OverFeat比R-CNN具有明显的速度优势：它大约快9倍，基于[每个图像引用的2秒数]34].这个速度来自OverFeat的滑动窗口（即区域提议）在图像级别没有扭曲的事实，因此可以在重叠窗口之间轻松共享计算。通过在任意大小的输入上以卷积方式运行整个网络来实现共享。应该以各种方式加速R-CNN，并将继续作为未来的工作。

5. 语义分割

区域分类是语义分割的标准技术，允许我们轻松地将R-CNN应用于PASCAL VOC分段挑战。为了便于与当前领先的语义分割系统（称为O_2P进行直接比较，用于“二阶汇集”）[4]，我们在他们的开源框架内工作。O_2P使用CPMC为每个图像生成150个区域建议，然后使用支持向量回归（SVR）预测每个类别的每个区域的质量。他们的方法的高性能是由于CPMC区域的质量和多种特征类型的强大二阶汇集（SIFT和LBP的丰富变体）。我们还注意到Farabet等人。[16最近在使用CNN作为多尺度每像素分类器的几个密集场景标记数据集（不包括PAS-CAL）上展示了良好的结果。

我们关注[2，4]并扩展PASCAL分段训练集，以包括Hariharan等人提供的额外注释。[22].设计决策和超参数在VOC 2011验证集上进行了交叉验证。最终测试结果仅评估一次。

CNN功能用于细分。我们评估了在CPMC区域上计算特征的三种策略，所有这些策略都是通过扭曲重新定位的矩形窗口开始的。

gion到227×227.第一个策略（完整）忽略了gion的形状和直接在扭曲窗口上计算CNN特征，就像我们检测时一样。但是，这些功能会忽略该区域的非矩形形状。两个区域可能具有非常相似的边界框，同时具有非常小的重叠。因此，第二个策略（fg）仅在区域的前景掩码上计算CNN特征。我们用平均输入替换背景，以便在平均减法后背景区域为零。第三种策略（完整+ fg）简单地连接完整和fg功能;我们的实验验证了它们的互补性

	完整的R-CNN		*fg* R-NNN		*全+ fg R-CNN*	
O_2P [4]	Dct	fc_7	Dct	fc_7	Dct	fc_7
46.4	43.0	42.5	43.7	42.1	**47.9**	45.8

表5：VOC 2011验证的分段平均准确度（%）。第1列呈现O2P;2-7使用我们在ILSVRC 2012上预培训的CNN。

VOC 2011的结果。表5 显示了我们在VOC 2011验证集上与O_2P相比的结果摘要。（请见附录E 对于完整的每类别结果。）在每个特征计算策略中，层fc_6 总是优于fc_7 ，以下讨论涉及fc_6 特征。fg策略稍微优于完整，表明屏蔽区域形状提供了更强的信号，与我们的直觉相匹配。然而，full + fg实现了47.9%的平均准确度，我们的最佳结果是4.2%的优势（也略微优于O_2P），表明即使给出fg功能，完整功能提供的上下文也是非常有用的信息. .值得注意的是，在我们的完整+ fg功能上训练20个SVR在单个核心上花费一个小时，而在O_2P功能上训练需要10个小时。

在表中6 我们在VOC 2011测试集上展示结果，将我们表现最好的方法fc_6 （全+ fg）与两个强基线进行比较。我们的方法在21个类别中的11个中实现了最高的分割准确度，并且最高的总分割准确度为47.9%，在各类别之间平均（但在任何合理的误差范围内可能与O_2P结果相关）。通过微调可以实现更好的性能。

VOC 2011测试	bg	航空自行车鸟船瓶巴士车猫椅牛表狗马mbike人植物羊沙发火车电视																				意思
R&P [2]	83.4	46.8	18.9	36.6	31.2	42.7	57.3	47.4	44.1	8.1	39.4	**36.1**	36.3	49.5	48.3	50.7	26.3	47.2	22.1	42.0	43.2	40.8
O2P [4]	**85.4**	**69.7**	22.3	45.2	**44.4**	46.9	66.7	57.8	56.2	**13.5**	**46.1**	32.3	41.2	**59.1**	55.3	51.0	**36.2**	50.4	**27.8**	46.9	**44.6**	47.6
我们的（全+ fg R-CNN fc_6）	84.2	66.9	**23.7**	**58.3**	37.4	**55.4**	**73.3**	**58.7**	**56.5**	9.7	45.5	29.5	**49.3**	40.1	**57.8**	**53.9**	33.8	**60.7**	22.7	**47.1**	41.3	**47.9**

表6：VOC 2011测试的分段准确度（%）。我们比较两个强大的基线：[区域和部分]（R&P）方法[2]和[二阶汇集（O2P）方法[4].在没有任何微调的情况下，我们的CNN实现了最高的分割性能，优于R&P并且大致匹配O2P。

6. 结论

近年来，物体检测性能停滞不前。表现最佳的系统是复杂的集合，将多个低级图像特征与来自物体探测器和场景分类器的高级上下文相结合。本文介绍了一种简单且可扩展的物体检测算法，与PASCAL VOC 2012上的最佳结果相比，可提供30%的相对改进。

我们通过两个见解实现了这一表现。首先是应用大容量卷积神经网络 -
适用于自下而上的区域提案，以便对象进行本地化和细分。第二个是在标记的训练数据稀缺时训练大型CNN的范例。我们表明，对具有丰富数据（图像分类）的辅助任务进行预训练是非常有效的，然后为数据稀缺（检测）的目标任务微调网络。我们推测，“监督的预训练/领域特定的微调”范例对于各种数据稀缺的视力问题将非常有效。

最后，我们指出通过结合使用计算机视觉和深度学习（自下而上区域提议和卷积神经网络）的经典工具来实现这些结果是很重要的。这两者不是反对科学探究的对象，而是自然而不可避免的合作伙伴。

致谢。这项研究得到了DARPA Mind's Eye和MSEE计划的部分支持，NSF颁发了IIS-0905647，IIS-1134072和IIS-1212798，
MURI N000014-10-1-0933，并得到丰田的支持。本研究中使用的GPU由NVIDIA公司慷慨捐赠。

附录

A. 对象提案转换

在这项工作中使用的卷积神经网络需要227 227像素的固定大小输入。对于检测× 我们将对象提议视为任意图像矩形。我们评估了两种将对象提议转换为有效CNN输入的方法。

第一种方法（“带有上下文的最严格的正方形”）将每个对象提案包含在最紧凑的方块内

图7：不同的对象提议转换。（A）相对于转换的CNN输入的实际规模的原始对象提案；（B）具有背景的最严格的正方形；（C）没有背景的最严格的方格；（D）翘曲。在每列和示例提议中，顶行对应于上下文填充的p = 0像素，而底行具有p = 16像素的上下文填充。

然后将该方块中包含的图像（各向同性地）缩放到CNN输入大小。数字7 列（B）显示了这种转变。该方法的变体（“没有上下文的最严格的正方形”）排除了围绕原始对象提议的图像内容。数字7 柱
（C）显示了这种转变。第二种方法（“warp”）各向异性地将每个对象提议缩放到CNN输入大小。数字7 列（D）显示了扭曲变换。

对于每个转换，我们还考虑在原始对象提议周围包含其他图像上下文。上下文填充量（p）被定义为变换的输入坐标系中原始对象提议周围的边界大小。数字7显示每个示例的顶行中的p = 0像素，并且底行中的p = 16像素。在所有方法中，如果源矩形超出图像，则将缺失的数据替换为图像均值（然后在将图像输入CNN之前将其减去）。一组试验表明，使用上下文填充（p = 16像素）的变形优于备选方案（3-5 mAP点）。显然，更多的替代方案是可能的，包括使用复制而不是平均填充。对这些替代方案的彻底评估留作未来的工作。

B. 正面与负面的例子和softmax

两种设计选择值得进一步讨论。第一个是：为什么正面和负面的例子被定义为微调CNN而不是训练对象检测SVM？为了简要回顾这些定义，对于微调，我们将每个对象提议映射到具有最大IoU重叠（如果有的话）的地面实例，并且如果IoU至少为0.5，则将其标记为匹配的地面实例类的正数. .所有其他提案都标有“背景”（即所有类别的反面例子）。相比之下，对于训练SVM，我们仅将地面实例框作为其各自类别的正面示例，并且标签提议与IoU重叠小于0.3，该类的所有实例都为该类的负数。落入灰色区域的建议（超过0.3 IoU重叠，但不是基本事实）将被忽略。

从历史上看，我们达到了这些定义 -
因为我们开始通过训练SVM来计算特征
通过ImageNet预先培训的CNN，进行微调

不是当时的考虑因素。在该设置中，我们发现我们用于训练SVM的特定标签定义在我们评估的选项集中是最佳的（其中包括我们现在用于微调的设置）。当我们开始使用微调时，我们最初使用与我们用于SVM训练相同的正面和负面示例定义。然而，我们发现结果比使用我们目前的正面和负面定义所获得的结果要糟糕得多。

我们的假设是，如何定义正面和负面的差异并不是根本重要的，而是由于微调数据有限这一事实。我们当前的方案引入了许多“抖动”的例子（这些提议在0.5和1之间重叠，但不是基本事实），它将正例的数量扩大了大约30倍。我们猜想在微调整个网络时需要这个大集合以避免过度拟合。但是，我们还注意到使用这些抖动的示例可能不是最理想的，因为网络没有针对精确定位进行微调。

这导致了第二个问题：为什么在微调之后，根本就训练SVM？简单地应用微调网络的最后一层（一种21向软最大回归分类器）作为对象检测器将更加清晰。我们尝试了这一点，发现VOC 2007的性能从mAP的54.2%下降到50.9%。这种性能下降可能源于几个因素的组合，包括微调中使用的正例的定义不强调精确定位，而softmax分类器是在随机抽样的负例上训练而不是在使用的“硬阴性”子集上用于SVM培训。

这个结果表明它可以获得接近
在没有训练SVM的情况下具有相同的性能水平

微调。我们推测，通过一些额外的调整来微调剩余的性能差距可能会被关闭。如果这是真的，这将简化并加速R-CNN训练而不会损失检测性能。

C. 边界框回归

我们使用简单的边界框回归阶段来提高本地化性能。在使用特定于类的检测SVM对每个选择性搜索提议进行评分后，我们使用特定于类的边界框回归器来预测用于检测的新边界框。这与可变形零件模型中使用的边界框回归在精神上类似[17].两种方法之间的主要区别在于，这里我们从CNN计算的特征回归，而不是从推断的DPM零件位置计算的几何特征回归。

我们的训练算法的输入是一组N训练 -

对$\{(P^i, G^i)\}_{i=1,\ldots,N}$，*其中*$P^i = (P^i_x, P^i_y, P^i_w, P^i_h)$指定建议$P^i$的边界框的像素坐标以及$P^i$的宽度和高度（以像素为单位）。因此，除非需要，否则我们删除上标i。每个地面实况边界框G以相同的方式指定：$G = (G_x, G_y, G_w, G_h)$。我们的目标是学习将拟议的方框P映射到地面实况框G的转换。

我们根据四个函数$d_x(P)$，$d_y(P)$，$d_w(P)$和$d_h(P)$对转换进行参数化。前两个指定P的边界框中心的尺度不变的平移，而后两个指定P的边界框的宽度和高度的对数空间平移。在学习这些功能之后，我们可以通过应用变换将输入提议P变换为预测的地面实况框$\hat{G}$

$$\hat{G}_x = P_w d_x(P) + P_x \quad (1)$$

$$\hat{G}_y = P_h d_y(P) + P_y \quad (2)$$

$$\hat{G}_w = P_w \exp(d_w(P)) \quad (3)$$

$$\hat{G}_h = P_h \exp(d_h(P))。 \quad (4)$$

每个函数$d_*(P)$（其中*是x，y，h，w之一）被建模为提议P的池$_5$特征的线性函数，由$\varphi_5(P)$表示。（隐含地假设$\varphi_5(P)$对图像数据的依赖性。）因此我们得到$d_*(P) = w_*^T \varphi_5(P)$，其中$w_*$是可学习的模型参数向量。我们通过优化正则化最小二乘目标（岭回归）来学习w_*：

$$w_* = \underset{\hat{w}}{\operatorname{argmin}} \sum_i^N (t_*^i - \hat{w}_*^T \varphi_5(P^i))^2 + \lambda \|\hat{w}_*\|^2。 \quad (5)$$

训练对（P，G）的回归目标t_*定义为

$$t_x = (G_x - P_x)/P_w \quad (6)$$

$$t_y = (G_y - P_y)/P_h \quad (7)$$

$$t_w = \log(G_w/P_w) \quad (8)$$

$$t_h = \log(G_h/P_h)。 \quad (9)$$

作为标准正则化最小二乘问题，这可以以封闭形式有效地解决。

我们在实现边界框回归时发现了两个微妙的问题。第一个是正则化很重要：我们根据验证集设置λ= 1000。第二个问题是在选择使用哪些训练对（P，G）时必须小心。直觉上，如果P远离所有地面实况框，则将P转换为地面实况框G的任务没有意义。使用像P这样的例子会导致无望的学习问题。因此，我们只从提案P中学习它是否在至少一个地面实况框附近。当且仅当重叠大于阈值（我们使用a设置为0.6）时，我们通过将P分配给具有最大IoU重叠的地面实况框G（如果它重叠多于一个）来实现“接近度”。验证集）。所有未分配的提案都将被丢弃。我们为每个对象类执行一次，以便学习一组特定于类的边界框回归量。

在测试时，我们对每个提案进行评分并仅预测其新的检测窗口一次。原则上，我们可以迭代此过程（即，重新对新预测的边界框进行评分，然后从中预测新的边界框，依此类推）。但是，我们发现迭代不会改善结果。

D. 其他功能可视化

数字12 显示20个池$_5$单元的其他可视化。对于每个单元，我们在所有VOC 2007测试中显示了24个区域建议，这些建议最大限度地激活了整个大约1000万个区域中的单元。

我们通过其中的（y，x，通道）位置标记每个单元6 6 256维池$_5$特征图。在每个通道内，CNN计算与输入区域完全相同的功能，（y，x）位置仅改变感受野。

E. 每个类别的细分结果

在表中7 除了O$_2$P方法之外，我们还显示了我们的六种分割方法中每种类型的VOC 2011 val分类准确度[4]. 这些结果显示哪些方法在20个PASCAL类中的每一个中都是最强的，加上背景类。

F. 跨数据集冗余分析

在辅助数据集上进行训练时，一个问题是它与测试集之间可能存在冗余。尽管对象检测和整个图像分类的任务有很大不同，使得这种交叉设置冗余更加令人担忧，我们仍然进行了彻底的调查，量化了ILSVRC 2012培训中PAS-CAL测试图像的包含程度。和验证集。我们的研究结果可能对有兴趣使用ILSVRC 2012作为PASCAL图像分类任务的训练数据的研究人员有用。

我们对重复（和近似重复）图像执行了两次检查。第一个测试基于flickr图像ID的完全匹配，这些匹配包含在VOC 2007测试注释中（这些ID对于后续的PASCAL测试集有意保密）。所有PASCAL图像和大约一半的ILSVRC都是从flickr.com收集的。这项检查在4952（0.63%）中出现了31场比赛。

第二次检查使用GIST [30描述符匹配，显示在[13]在大型（> 100万）图像集中，在近似重复图像检测方面具有出色的性能。关注[13]，我们计算了所有ILSVRC 2012 trainval和PASCAL 2007测试图像的扭曲32 32像素版本的GIST描述符。

GIST描述符的欧几里德距离最近邻匹配揭示了38个近似重复的图像（包括通过flickr ID匹配找到的所有31个）。匹配在JPEG压缩级别和分辨率方面略有不同，并且在较小程度上裁剪。这些发现表明重叠很小，不到1%。对于VOC 2012，由于flickr ID不可用，我们仅使用GIST匹配方法。根据GIST比赛，1.5%的VOC 2012测试图像在ILSVRC 2012 trainval中。VOC 2012的比率稍高可能是因为两个数据集的收集时间比VOC 2007和ILSVRC 2012更接近。

G. 记录更改日志

本文档跟踪R-CNN的进展情况。为了帮助读者了解它如何随着时间的推移而发生变化，这里有一个描述修订版的简短更新日志。

v1初始版本。

v2 CVPR 2014相机准备版。包括（1）从更高的学习率（0.001而不是0.0001）开始微调，（2）在准备CNN输入时使用上下文填充，以及（3）边界框回归以修复本地化所带来的检测性能的显着改进错误。

v3 ILSVRC2013检测数据集的结果以及与OverFeat的比较被整合到几个部分（主要是部分2 和部分4).

VOC 2011 val	bg	航空自行车鸟船瓶巴士车猫椅牛表狗马mbike人植物羊沙发火车电视																				意思
O_2P [4]	**84.0**	**69.0**	21.7	47.7	42.2	42.4	**64.7**	**65.8**	57.4	**12.9**	37.4	20.5	43.7	35.7	52.7	51.0	**35.8**	**51.0**	28.4	59.8	49.7	46.4
*完整的*R-CNN fc_6	81.3	56.2	23.9	42.9	40.7	38.8	59.2	56.5	53.2	11.4	34.6	16.7	48.1	37.0	51.4	46.0	31.5	44.0	24.3	53.7	51.1	43.0
*完整的*R-CNN fc_7	81.0	52.8	**25.1**	43.8	40.5	42.7	55.4	57.7	51.3	8.7	32.5	11.5	48.1	37.0	50.5	46.4	30.2	42.1	21.2	57.7	**56.0**	42.5
fg R-NNN fc_6	81.4	54.1	21.1	40.6	38.7	**53.6**	59.9	57.2	52.5	9.1	36.5	**23.6**	46.4	38.1	53.2	51.3	32.2	38.7	**29.0**	53.0	47.5	43.7
fg R-NNN fc_7	80.9	50.1	20.0	40.2	34.1	40.9	59.7	59.8	52.7	7.3	32.1	14.3	48.8	42.9	54.0	48.6	28.9	42.6	24.9	52.2	48.8	42.1
全+ fg R-CNN fc_6	83.1	60.4	23.2	48.4	**47.3**	52.6	61.6	60.6	**59.1**	10.8	**45.8**	20.9	**57.7**	43.3	**57.4**	**52.9**	34.7	48.7	28.1	60.0	48.6	**47.9**
满+ fg R-CNN fc_7	82.3	56.7	20.6	**49.9**	44.2	43.6	59.3	61.3	57.8	7.7	38.4	15.1	53.4	**43.7**	50.8	52.0	34.1	47.8	24.7	**60.1**	55.2	45.7

表7：VOC 2011验证集上的每类别分段准确度（%）。

v4 softmax与SVM的结果见附录B 包含错误，已修复。我们感谢Sergio Guadarrama帮助确定了这个问题。

v5使用Simonyan和Zisserman的新16层网络架构添加了结果[43到科3.3 和表3.

参考

[1] B. Alexe，T。Deselaers和V. Ferrari。测量图像窗口的对象性。TPAMI，2012。2

[2] P. Arbela'ez，B。Hariharan，C。Gu，S。Gupta，L。Bourdev，and J. Malik。使用区域和部分的语义分割。在 *cvpr*，2012。10，11

[3] P. Arbela'ez，J。Pont-Tuset，J。Barron，F。Marques和J. Malik。多尺度组合分组。在CVPR，2014年。3

[4] J. Carreira，R。Caseiro，J。Batista 和 C. Sminchisescu。具有二阶池的语义分割。在ECCV，2012年。4，10，11，13，14

[5] J. Carreira和C. Sminchisescu。CPMC：使用约束参数最小割的自动对象分割。 *tpami，2012*。2，3

[6] D. Cires¸an，A。Giusti，L。Gambardella 和 J. Schmidhuber。用深度神经网络检测乳腺癌组织学图像中的有丝分裂。在MICCAI，2013年。3

[7] N. Dalal和B. Triggs。用于人体检测的定向梯度的直方图。在CVPR，2005年。1

[8] T. Dean，MA Ruzon，M。Segal，J。Shlens，S。Vijayanarasimhan和J. Yagnik。在一台机器上快速，准确地检测100，000个对象类。在CVPR，2013年。3

[9] J. Deng，A。Berg，S。Satheesh，H。Su，A。Khosla 和 L. FeiFei。ImageNet 大型视觉识别竞赛2012（ILSVRC2012）。http://www.image-net.org/挑战/ LSVRC / 2012 /。1

[10] J. Deng，W。Dong，R。Socher，L.-J。Li，K。Li和L. Fei Fei。ImageNet：一个大规模的分层图像数据库。在CVPR，2009年。1

[11] J. Deng，O。Russakovsky，J。Krause，M。Bernstein，AC Berg和L. Fei-Fei。可扩展的多标签注释。在CHI，2014年。8

[12] J. Donahue，Y。Jia，O。Vinyals，J。Hoffman，N。Zhang，
E. Tzeng和T. Darrell。DeCAF：用于通用视觉识别的深度卷积激活功能。在ICML，2014年。2

[13] M. Douze，H。Je'gou，H。Sandhawalia，L。Amsaleg，and C. 施密德。用于web级图像搜索的gist描述符的评估。在Proc。ACM国际图像和视频检索会议，2009年。13

[14] I. Endres和D. Hoiem。类别独立对象提案。在ECCV，2010年。3

[15] M. Everingham，L。Van Gool，CKI Williams，J。Winn，and A. Zisserman。PASCAL视觉对象类（VOC）挑战。IJCV，2010。1，4

[16] C. Farabet，C。Couprie，L。Najman和Y. LeCun。学习场景标注的分层功能。TPAMI，2013年。10

[17] P. Felzenszwalb，R。Girshick，D。McAllester和D. Ramanan。使用有区别训练的基于部件的模型进行物体检测。TPAMI，2010年。2，4，7，12

[18] S. Fidler，R。Mottaghi，A。Yuille和R. Urtasun。自上而下检测的自下而上分割。在CVPR，2013年。4，5

[19] K. Fukushima。Neocognitron：一种自组织神经网络模型，用于模式识别机制，不受位置偏移的影响。生物控制论，36（4）：193-202，1980。1

[20] R. Girshick，P。Felzenszwalb和D. McAllester。经过专业训练的可变形零件模型，第5版。http://www.cs.berkeley.edu/~rbg/latent-v5/。2，5，6，7

[21] C. Gu，JJ Lim，P。Arbela'ez和J. Malik。使用地区识别。在CVPR，2009年。2

[22] B. Hariharan，P。Arbela'ez，L。Bourdev，S。Maji 和J. Malik。逆检测器的语义轮廓。在ICCV，2011年。10

[23] D. Hoiem，Y。Chodpathumwan和Q. Dai。诊断物体探测器中的错误。在ECCV中。2012。2，7，8

[24] Y. Jia。Caffe：一种用于快速特征嵌入的开源卷积架构。http://caffe. berkeleyvision. org/org/，2013. 3

[25] A. Krizhevsky，I。Sutskever和G. Hinton。使用深度卷积神经网络的ImageNet分类。在NIPS，2012年。1，3，4，7

[26] Y. LeCun，B。Boser，J。Denker，D。Henderson，R。Howard，W. Hubbard和L. Jackel。反向传播适用于手写邮政编码识别。神经比较，1989。1

[27] Y. LeCun，L。Bottou，Y。Bengio和P. Haffner。基于梯度的学习应用于文档识别。PROC。IEEE，1998年。1

[28] JJ Lim，CL Zitnick和P. Doll'r。草图标记：轮廓和物体检测的学习中级表示。在CVPR，2013年。6，7

类	美联社	类	美联社	类	美联社	类	美联社	类	美联社
手风琴	50.8	蜈	30.4	发胶	13.8	铅笔盒	11.4	雪犁	69.2
飞机	50.0	链锯	14.1	汉堡包	34.2	卷笔刀	9.0	皂液器	16.8
蚂蚁	31.8	椅子	19.5	锤子	9.9	香水	32.8	足球	43.7
羚羊	53.8	钟	24.6	仓鼠	46.0	人	41.7	沙发	16.3
苹果	30.9	鸡尾酒调酒器	46.2	口琴	12.6	钢琴	20.5	抹刀	6.8
狈狯	54.0	咖啡机	21.5	竖琴	50.4	菠萝	22.6	松鼠	31.3
朝鲜蓟	45.0	计算机键盘	39.6	宽边帽	40.5	乒乓球	21.0	海星	45.1
斧头	11.8	电脑鼠标	21.2	白菜头	17.4	投手	19.2	听筒	18.3
婴儿床	42.0	螺旋形的	24.2	头盔	33.4	比萨	43.7	火炉	8.1
背包	2.8	奶油	29.9	河马	38.0	塑料袋	6.4	过滤器	9.9
面包圈	37.5	槌球	30.0	单杠	7.0	板架	15.2	草莓	26.8
平衡木	32.6	拐杖	23.7	马	41.7	石榴	32.0	担架	13.2
香蕉	21.9	黄瓜	22.8	热狗	28.7	棒冰	21.2	墨镜	18.8
创可贴	17.4	杯子或杯子	34.0	iPod的	59.2	豪猪	37.2	游泳裤	9.1
班卓琴	55.3	尿布	10.1	等足	19.5	电钻	7.9	猪	45.3
棒球	41.8	数码时钟	18.5	海蜇	23.7	椒盐卷饼	24.8	注射器	5.7
篮球	65.3	洗碗机	19.9	考拉	44.3	打印机	21.3	表	21.7
沐浴帽	37.2	狗	76.8	杓	3.0	冰球	14.1	磁带播放器	21.4
烧杯	11.3	家猫	44.1	瓢虫	58.4	出气筒	29.4	网球	59.1
熊	62.7	蜻蜓	27.8	灯	9.1	钱包	8.0	蜱	42.6
蜜蜂	52.9	鼓	19.9	笔记本电脑	35.4	兔子	71.0	领带	24.6
灯笼椒	38.8	哑铃	14.1	柠檬	33.3	球拍	16.2	虎	61.8
长凳	12.7	电扇	35.0	狮子	51.3	射线	41.1	烤面包机	29.2
自行车	41.1	象	56.4	口红	23.1	小熊猫	61.1	红绿灯	24.7
粘合剂	6.2	香粉	22.1	蜥蜴	38.9	冰箱	14.0	培养	60.8
鸟	70.9	图	44.5	龙虾	32.4	遥控	41.6	长号	13.8
书架	19.3	文件柜	20.6	运动衣	31.0	橡皮擦	2.5	喇叭	14.4
领结	38.8	花盆	20.2	马拉卡	30.1	橄榄球	34.5	龟	59.1
弓	9.0	长笛	4.9	麦克风	4.0	统治者	11.5	电视或显示器	41.7
碗	26.7	狐狸	59.3	微波	40.1	盐或胡椒粉	24.6	独轮车	27.2
乳罩	31.2	圆号	24.2	牛奶可以	33.3	萨克管	40.8	真空	19.5
卷饼	25.7	青蛙	64.1	迷你裙	14.9	蝎	57.3	小提琴	13.7
总线	57.5	平底锅	21.5	猴	49.6	螺丝刀	10.6	排球	59.7
蝴蝶	88.5	大熊猫	42.5	摩托车	42.2	密封	20.9	华夫饼干	24.0
骆驼	37.6	金鱼	28.6	蘑菇	31.8	羊	48.9	垫圈	39.8
开罐器	28.9	高尔夫球	51.3	钉	4.5	滑雪	9.0	水壶	8.1
汽车	44.5	高尔夫车	47.9	颈托	31.6	臭鼬	57.9	船舶	40.9
大车	48.0	鳄梨	32.3	双簧管	27.5	蜗牛	36.2	鲸	48.6
黄牛	32.3	吉他	33.1	橙子	38.8	蛇	33.8	酒瓶	31.2
大提琴	28.9	吹风机	13.0	獭	22.2	雪地车	58.8	斑马	49.6

表8：ILSVRC2013检测测试集的每类平均精度（%）。

[29] D. Lowe。从尺度不变的关键点获得独特的图像特征。IJCV，2004。1

[30] A. Oliva和A. Torralba。建模场景的形状：

空间包络的整体表示。IJCV，2001。13

[31] X. Ren和D. Ramanan。稀疏代码的直方图

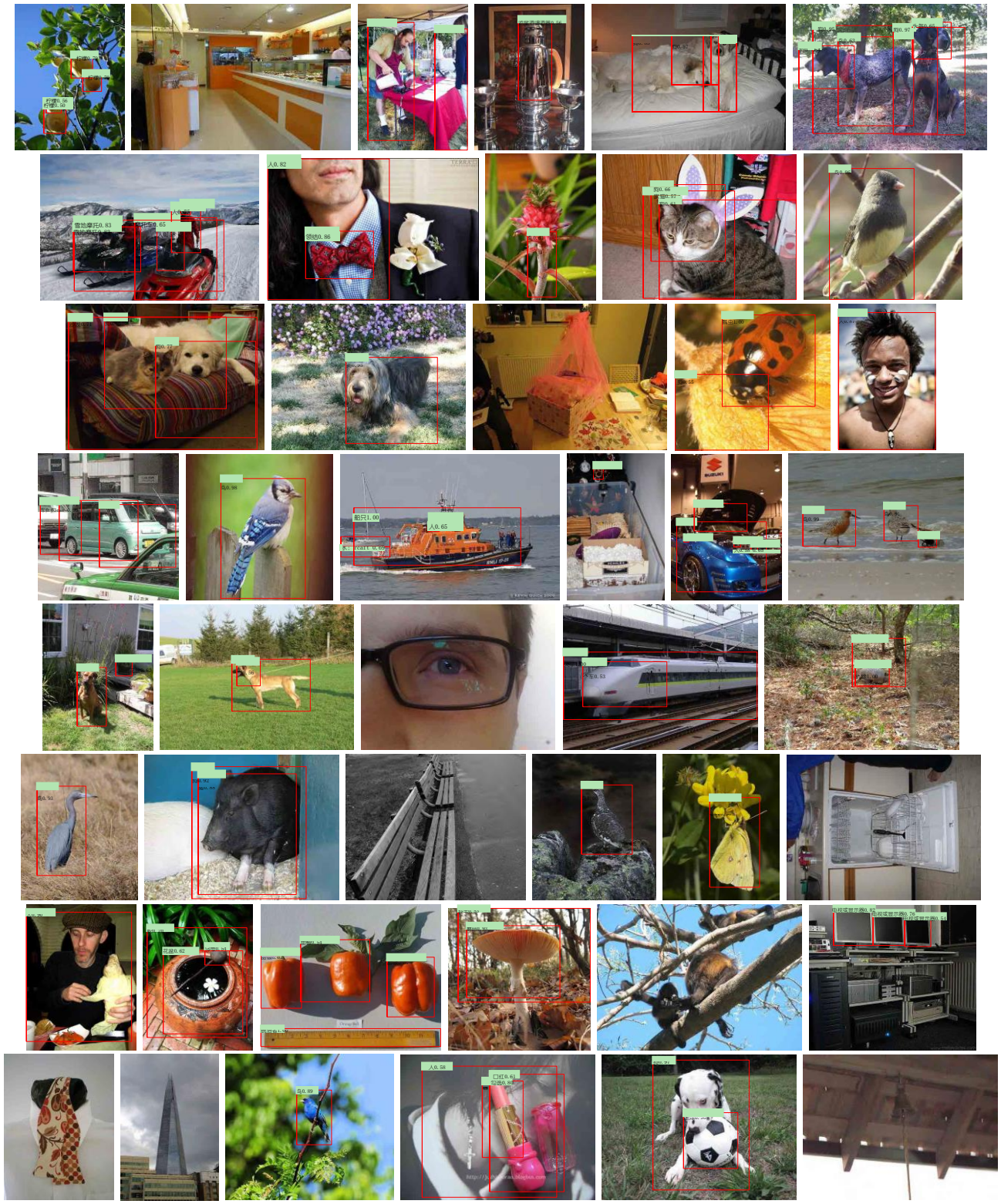

图8：val2上的设置检测示例，该配置在val2上实现了31.0%mAP。每个图像都是随机采样的（这些图像不是策划的）。显示精度大于0.5的所有检测。每个检测都用检测器的精确回忆曲线标记预测的类别和检测的精度值。建议使用缩放数字查看。

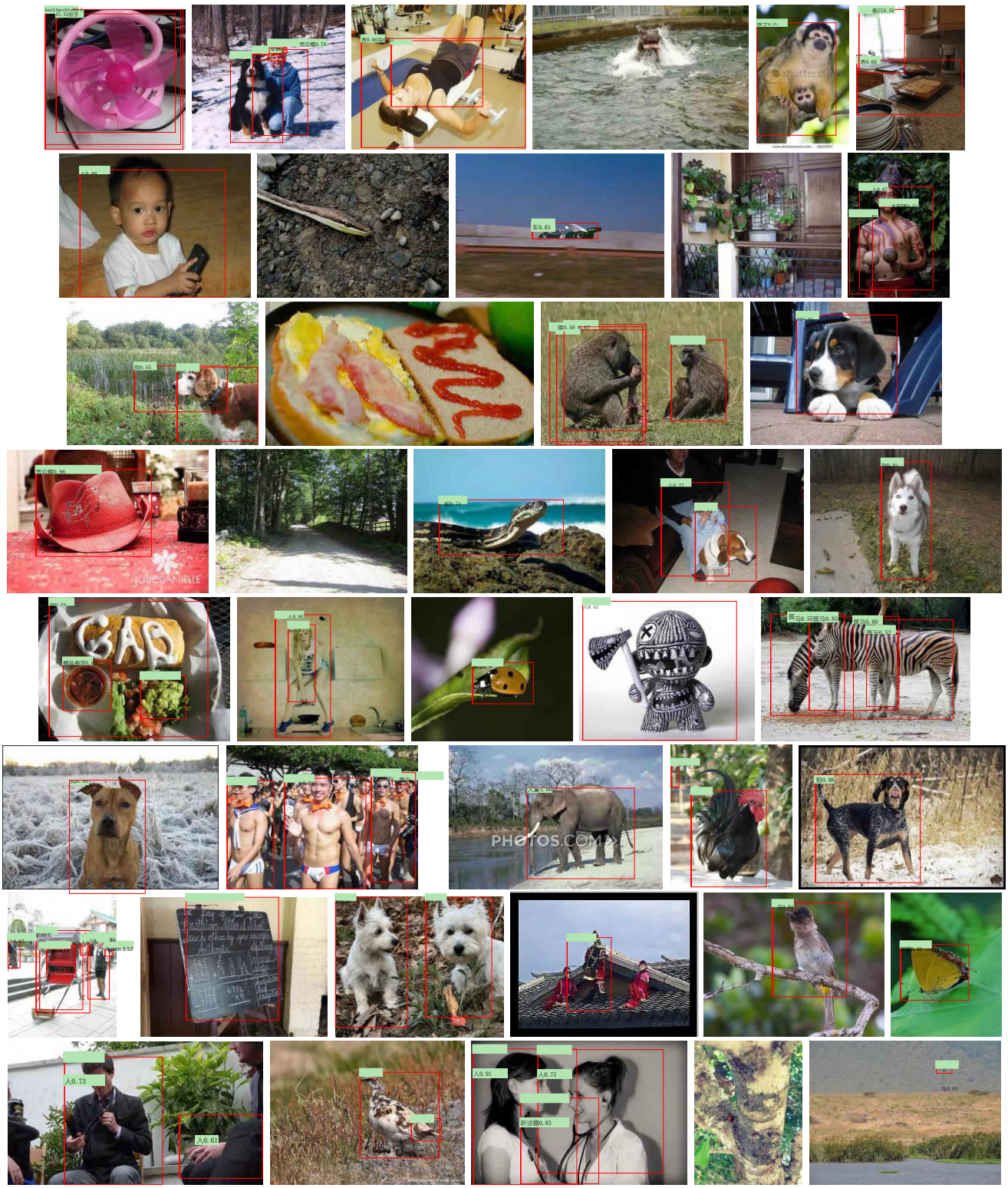

图9：更随机选择的示例。见图8 标题为细节。建议使用缩放数字查看。

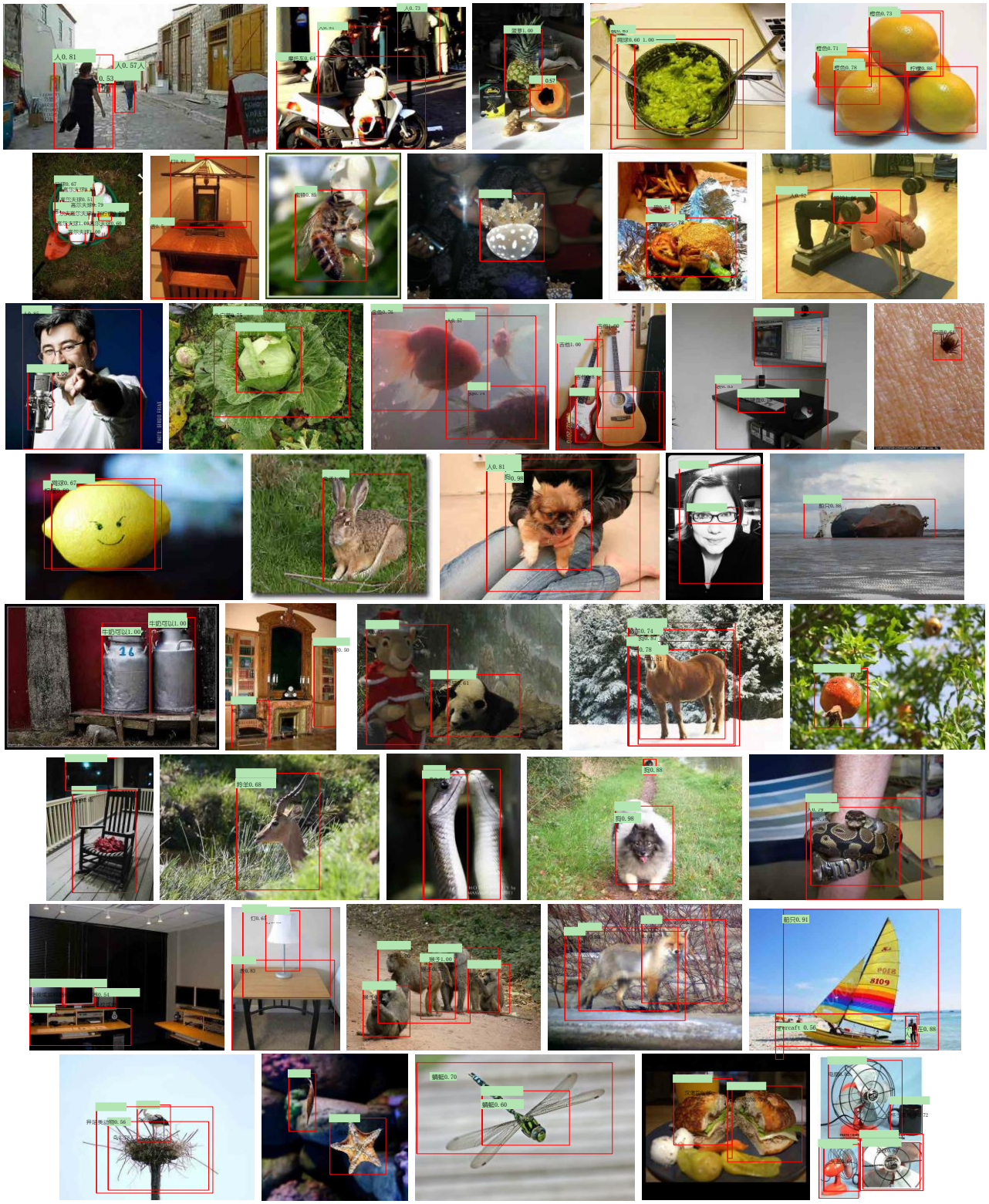

图10：策划示例。选择每张图片是因为我们发现它令人印象深刻，令人惊讶，有趣或有趣。建议使用缩放数字查看。

物体检测。在CVPR，2013年。6,7

[32] HA Rowley，S。Baluja和T. Kanade。基于神经网络的人脸检测。TPAMI，1998年。2

[33] DE Rumelhart，GE Hinton和RJ Williams。通过错误传播学习内部表示。Parallel Distributed Processing，1：318-362,1986。1

[34] P. Sermanet，D。Eigen，X。Zhang，M。Mathieu，R。Fergus和Y. LeCun。OverFeat：使用卷积网络进行集成识别，定位和检测。在ICLR，2014年。1,2,4,10

[35] P. Sermanet，K。Kavukcuoglu，S。Chintala和Y. LeCun。具有无监督多阶段特征学习的行人检测。在CVPR，2013年。2

[36] H. Su，J。Deng和L. Fei-Fei。用于视觉对象检测的众包注释。在AAAI技术报告，第四届人类计算研讨会，2012年。8

[37] K. Sung和T. Poggio。基于视图的基于视图的人脸检测学习。技术报告AI备忘录第1521号，马萨诸塞理工学院，1994年。4

[38] C. Szegedy，A。Toshev和D. Erhan。用于物体检测的深度神经网络。在NIPS，2013年。2

[39] J. Uijlings，K。van de Sande，T。Gevers和A. Smeulders。选择性搜索对象识别。IJCV，2013。1,2,3,4,5,9

[40] R. Vaillant，C。Monrocq和Y. LeCun。用于图像中对象定位的原始方法。关于视觉的 IEE 专业，*图像与信号处理，1994*。2

[41] X. Wang，M。Yang，S。Zhu和Y. Lin。用于通用对象检测的Regionlet。在ICCV，2013年。3,5

[42] M. Zeiler，G。Taylor和R. Fergus。适用于中高级特征学习的自适应解卷积网络。在CVPR，2011年。4

[43] K. Simonyan和A. Zisserman。用于大规模图像识别的非常深的卷积网络。arXiv preprint，arXiv：1409.1556,2014。6,7,14

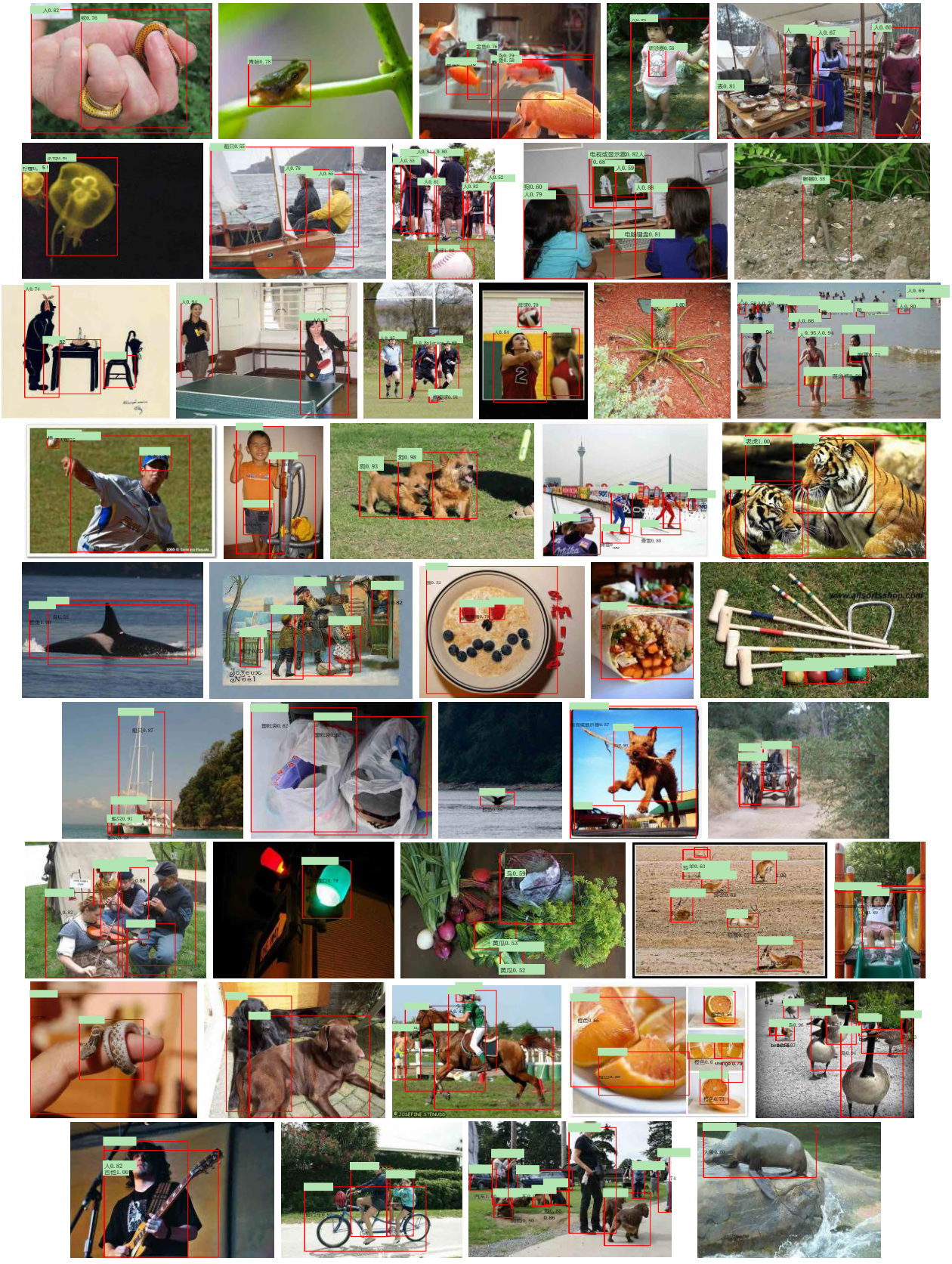

图11：更多精选示例。见图10 标题为细节。建议使用缩放数字查看。

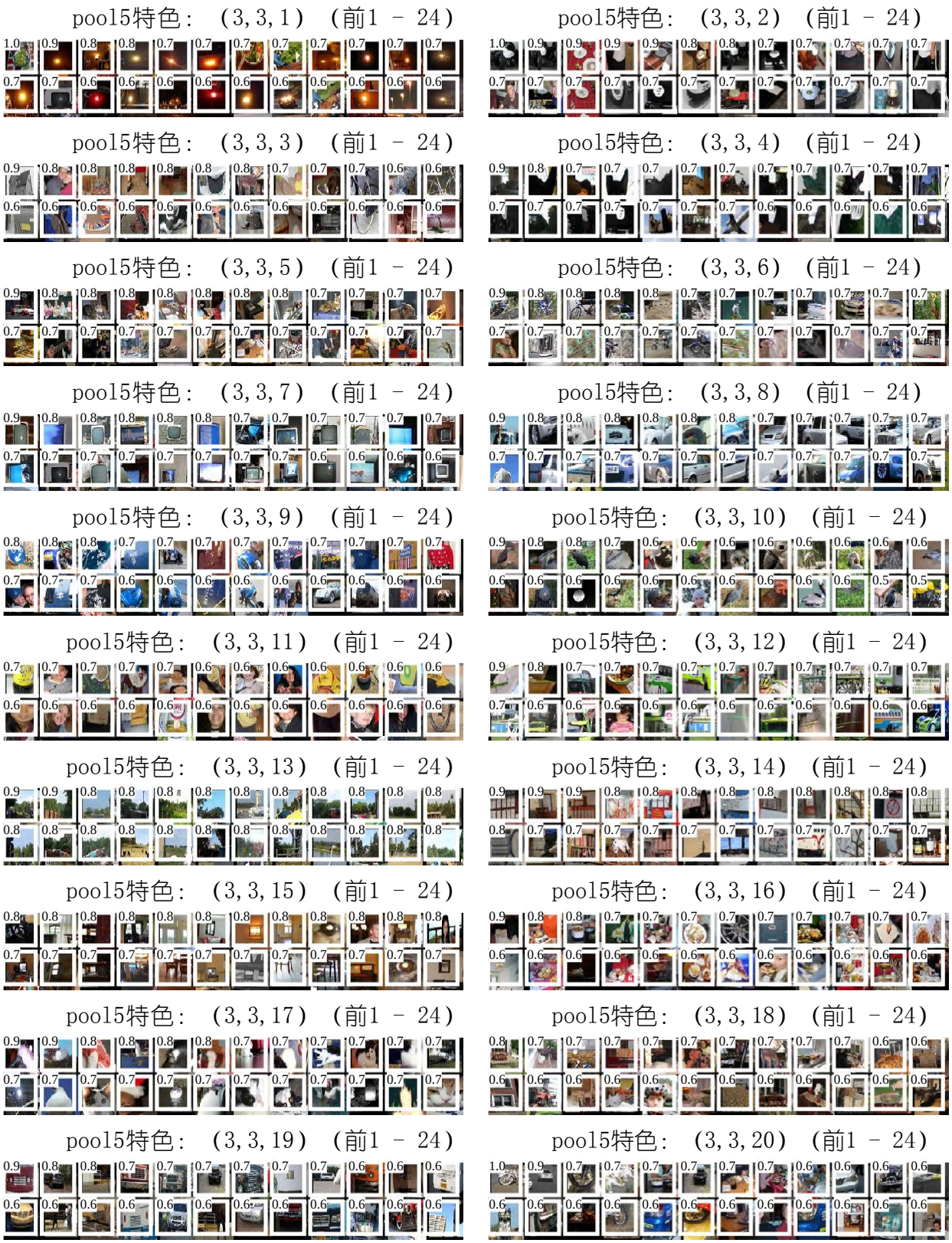

图12：我们展示了VOC 2007测试中大约1000万个区域中的24个区域提案，它们最强烈地激活了20个单元中的每一个。每个蒙太奇都由6 6 256维池$_5$特征图中的单位（y，x，通道）位置标记。每个图像区域都以白色的单位感受野的覆盖图绘制。激活值（我们通过除以通道中所有单位的最大激活值进行标准化）显示在感知区域的左上角。最佳数字缩放查看。