丰富的特征层次结构，以准确的目标检测和语义分割

罗斯吉尔希克[1]杰夫多纳休[12]特雷弗达雷尔[12]吉登德拉马利克[1]
[1]加州大学伯克利分校和[2]1CS1
{rbg, jcionahue, ~ revor, nal ik} @eecs .伯克 1 ey。edu

摘要 *R-CNN：具有 CNN 特征的区域*

物体检测性能，在该仪器上测量。在过去的几年里，他一直在谨慎行事。内增强˙方法是一种复杂的集成系统，通常将裸体的低级图像特征与高级上下文结合起来。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 VOC2012 的最佳结果，平均精度（niAP）提高了 30%以上，达到了 53.3˙的 mAP。我们的方法结合了两个关键的签名：（1 丿 1 可以应用高容量卷积神经网络（CNN）。为了定位和分割对象，并且当标记的训练数据稀缺时，对辅助任务的监督预训练，然后进行特定领域的微调，产生显著的执行提升。由于我们将区域建议与 cnn 结合起来，我们称我们的方法为 R-CNN：具有 CA ‘A 的区域[1]特征我们还提出了实验，提供了对网络学习内容的见解，揭示了丰富的图像特征层次。整个系统的源代码是公民 ci！∴：：一：：　.

1. 介绍

特点很重要。过去十年在各种识别任务上的进展在很大程度上是基于 SIFT [26]和 HOG |˜|的使用。但是，如果我们看典型视觉识别任务的所有表现。流氓 VOC 目标检测| 2|。人们普遍认为，在 2010-2012 年期间，研究进展缓慢。通过构建集成系统和使用少量成功方法所获得的小收益。

SIFT 和 HOG 是块状方向的直方图，我们可以将其粗略地与 VI 中的复杂细胞联系起来。灵长类动物视觉通路中的第一个皮质区域。但我们也知道，识别发生在下游的几个阶段。这表明，这可能存在等级制度。用于计算特征的多阶段过程，提供更有利于识别的信息。

福岛，新诺克尼龙，11[1]J.生物学上

启发了模式识别的层次模型和移位不变模型的方法。这是对这样一个过程的早期尝试。新时代的认知。howe\ cr.缺乏超级的训练算法。*LeCun 等人*。|2 J 通过显示 lhat 的随机梯度下降，提供了缺失的算法。\ ia

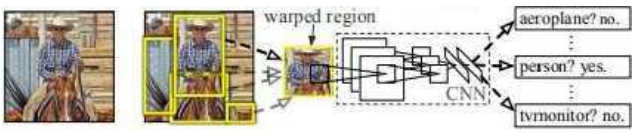


图 1：目标检测系统。我们的系统(1)湖泊的输入图像。(2>提取了大约 2000 个以上地区的提案。(3)利用大型卷积神经网络（CNN），然后(4)对每个区域进行分类。R-CNN 在 PASCAL VOC 2010 上的平均平均精度（mAP）为 53.7%。进行比较。|| 报 告 35.1˚!(使用(同一地区专业

posals.但采用了空间金字塔和视觉词汇袋的方法。流行的可拆卸部分模型为 33.4˙(。

backpropagation.可以训练卷积神经网络（CNNs）。一类扩展网络加速器的模型。

cnn 在 20 世纪 90 年代得到了大量的使用。[24]).但后来就不再流行了，尤其是在计算机领域。随着支持性扩展机器的兴起。在 2012 年。克里日耶夫斯基等人。I 21 通过在大规模视觉识别挑战<ILSVRC)|.>上显示出更高的图像分类精度，重新点燃了人们对 cnn 的兴趣。他们的成功源于训练一个大型 CNN 拍摄 120 万张带标签的图像。再加上勒昆的 CNNmax (.r。0)校正非线性线和“正正则化”)。

在 2012 年的 ILSVRC 研讨会上，人们对犯罪结果的重要性进行了激烈的辩论。中心问题可以提炼为以下：lbCNN 分类结果在多大程度上推广到帕斯卡 VOC 挑战的目标检测结果？

我们通过弥合图像分类和目标检测之间的鸿沟来明确地回答这个问题。本文旨在表明，与基于更简单的类 IIOG 功能的系统相比，CNN 可以在流氓

VOC 上导致更高的目标检测性能。[1]要实现这一结果，需要解决两个问题：将对象定位为一个深度网络和训练一个只使用少量带注释的检测数据的高容量模型。

与图像分类不同，检测需要在图像中定位（可能有很多）对象。一种方法将定位框架定义为一个回归问题。Howex er.来自 Szegedy 等人的工作。\|。和我们自己的。表明这种方法在实践中可能表现不佳（他们报告的 VOC 2007 的 mAP 为 30.5%，而我们的方法获得的结果为 58.5%）。另一种选择是建立一个滑动窗口检测器。cnn 已经以这种方式使用了至少几十年。通常是针对受约束的对象类别。比如脸|。　　|和行人，|　‘|。按顺序排列

为了保持高空间分辨率，这些 cnn 通常只有卷积层和池化层。我们还考虑了采用滑动窗口的方法。我越来越低了。在我们的网络中高的单元，有 5 个卷积层，在输入图像中有非常大的接受域（195 x 195 像素）和步幅（32 x 32 像素）。这使得在滑动窗口范式中的精确定位成为一个开放的技术挑战。

相反，正如 Gu 等人所主张的那样，我们通过使用“使用区域识别”范式来解决 CNN 定位问题。在[|]中。在测试时，我们的方法为 lhe 输入图像生成大约 2000 个类别独立的区域建议，使用 CNN 从每个提案中提取一个固定长度的特征向量。我们使用一种简单的技术（仿射图像翘曲）来计算来自每个区域建议的固定大小的 CNN 输入，而不考虑区域的形状。图 I 展示了我们的 mclhod 的概述，并强调了我们的一些结果。因为我们的系统结合了区域建议和 cnn。我们将该方法称为 R-CNN：具有 CNN 特征的区域。

在检测方面面临的第二个挑战是，标记数据稀缺，而且可用的数量是训练大型 CNN 的。这个问题的传统解决方案是不使用它的预训练，然后进行监督微调（例如，[29]）。本文的第二个主要贡献是，我们展示了在一个大型辅助数据集（ILSVRC）上的监督保留。然后在小沙拉（流氓）上进行特定领域的微调，是在数据稀缺时学习高容量 cnn 的有效范式。在我们的实验中，微调 tor 检测将 mAP 性能提高了 8 个百分点。经过微调后，我们的系统在 VOC 2（）10 上实现了 54%的 mAP，而在高调优时的 mAP 为 33%。基于 IIOG 的可变形桶模型（DPM）|　　|.

我们的系统也相当有效。唯一特定于类的计算是一个相当小的矩阵向量生产和贪婪的非极大抑制。这种计算主要遵循在所有类别中共享的特征，这些特征也是比以前使用的区域特征的两个低维数级（cf \ 21）。

HOG-likc 特性的一个优点是它们的简单性：它更容易理解它们所携带的信息（尽管 1341 表明我们的直觉可能会让我们失望）。我们能深入了解 CNN 所学到的表现方式吗？也许拥有超过 5400 万参数的紧密连接层是关键？他们不是。W[1]CNN 发现，一个惊人的比例，94%的参数可以被去除，而检测精度只有适度的下降。相反，通过探测 lhe 网络中的单元，我们可以看到 lhe 卷积层学习了一组不同的丰富的特征（图 3）。

理解我们的方法的低故障节点对改进它也至关重要。因此，我们报告了来自 Hoiem el al 的实时检测分析工具的结果。| i |.作为这一分析的直接结果，我们认为一个简单的边界盒回归方法显著减少了定位化。这是主要的错误模式。

在开发技术细节之前。我们注意到，由于 R-CNN 操作于区域，因此扩展它是语义分割的任务。与较小的修改。我们还在流氓 VOC 分割任务上获得了状态分析结果，在 VOC 2011 测试集上的轴向分割准确率为 47.9%。

2. 用 R-CNN 检测对象

我们的目标探测系统由三个模块组成。第一个方案是产生与类别独立的区域提案。这些建议定义了我们的检测器可用的候选检测集。第二个模块是一个大型的卷积神经网络，它从每个区域中提取一个固定长度的特征向量。第三个模块是一组特定于类的线性 SV Ms。在本节中，我们将介绍每个模块的设计决策，描述它们的测试时间使用情况，详细说明如何学习它们的参数，并显示关于 PASCAL VOC 2010-12 的结果。

[1]一份描述 R-CNN iirsi 的技术报告出现在 http: //arxiv 上。orq/ahs/1311.2524vi 在 Nt>\。201.<

2.1. 模块设计

区域建议。各种新的论文提供了生成计算独立的区域建议的方法。例子包括：对象 111。选择性搜索[类别独立的对象建议||。约束参数最小切割（CPMC）[ ]。多尺度的组合分组[|]。和 Ci resan 等人，他们通过应用 CNN 到规则间隔的方形作物来选择有丝分裂细胞，这是区域建议的一个特例。虽然 R-CNN 对特定区域的建议方法是不可知的，但我们使用选择性搜索来实现与之前的控制比较

Eigurc 2：来自 VOC 2007 列车的<*样本。

检测工作|。　　|).

特征抽取我们使用 CNN 的||实现从每个区域提案中提取一个 4096 维的特征向量。I.-I 特征是通过正向传播一个平均减去的 227 x 227 RGB 图像通过 5 个卷积层和 2 个全连接层来计算的。我们参考读者罗[一一 211 更多网络拱我领导的细节。

为了计算一个区域提议的特征，我们使用 tirsl 将该区域中的图像数据转换为与 CNN 兼容的形式（它的架构需要一个固定的 227 x 227 像素大小的输入）。在我们的任意形状区域的许多可能的转换中，我们选择了最简单的转换。无论候选区域的大小或长宽比，我们都在一个紧密的边界框中扭曲所有像素到 il 到所需的大小。在扭曲之前，我们扩张了紧边界盒子，以便在扭曲的大小下，在原始盒子周围恰好有 p 个扭曲的图像上下文的像素（我们使用= 16），图 2 显示了扭曲训练区域的随机采样。补充材料讨论了翘曲的替代方法。

2.2. 测试时间检测

在所有测试时间内，我们对测试图像进行选择性搜索，提取了大约 2000 个区域建议（我们在所有实验中使用选择性搜索，*fasl 模式）。我们扭曲每个提议，并通过 lhe CNN 向前传播它，以便从期望的层读取 eff 特征。然后，对于每个类，我们使用为该类训练的 SVM 对每个提取的特征向量进行评分。给定一个图像中所有有得分的区域。我们应用一个贪婪的非最大值抑制（对于每个类独立），如果一个区域有一个交叉过联合（loU）重叠，而一个更高的得分选择区域大于一个学习阈值，我们将拒绝该区域。

运行时分析。有两个属性使检测效率更高。首先，所有的 CNN 参数都在所有类别中共享。其次，与其他常用方法相比，CNN 计算出的特征向量是低维的。比如带有视觉单词编码的空间金字塔。例如，在 UVA 检测系统 I 中使用的特征比我们的特征大两个数量级(360k v.v。4k-diniensional).

这种共享的结果是，计算区域建议和特性（GPU 上 13 秒/图像或 CPU 上 53 秒/图像）所花费的时间被分摊到所有类中。唯一的类特定计算是特征和 SVM 权重之间的点积和非镍氮抑制。实际上，al! 一个图像的点产品被分割成一个单一的数学矩阵产品。特征矩阵通常为 2000 x 4096，SVM 权重矩阵为 4096 x A\，其中-V 为类数。

这一分析表明，R-CNN 可以扩展到数千个对象类，而不需要求助于近似的技术。如散希。即使有 100k 个类，在一个现代的多核 CPU 上，产生矩阵乘法也只需要 10 秒。这种效率不仅仅是使用区域建议和共享特性的结果。UVA 系统，由于其高维特性，需要 134GB 内存来存储 100k 线性预测器，而我们的低维特性只需要 1.5GB。

将 R-CNN 与 Dean el al 最近的工作进行对比也很有趣。使用 dpm 和散列||进行可扩展的数据选举。他们报告说，当引入 10k 粉碎器类时，每张图像运行 5 分钟时，VOC 2007 的 mAP 约为 16%。用我们的方法。10k 个检测器可以在一个 CPU 上运行大约一分钟。由于没有做出近似，mAP 将保持在 59%（第 3.2 节）。

2.3. 训练

监督预训练。我们在一个大型的辅助数据集（ILSVRC 2012）上对 CNN 进行了有区别的预训练。没有边界框标签）。使用开源的 Caffe CNN 库进行预训练[2|]。简而言之，我们的 CNN 几乎与克里热夫斯基等人的表现相当。| .i |.在 lhe ILSVRC 2012 验证集上，误差率高出 2.2 个百分点。这种差异是由于法律培训过程的简化。

领域特定的线调优，lb 使我们的 CNN 适应新的任务（检测）和新的域（扭曲的 VOC 窗口），我们继续使用 lhe CNN 参数的随机梯度下降（SGD）训练。除了用随机初始化的 21 路分类层（lhe2（）VOC 类加背景）替换分类层 lhe（）00 路之外，CNN 算法没有改变。我们将所有使用>（）.5 loU 与地面真值框重叠的区域提议视为该框类的正值，其余的视为负值。我们在 0.001（初始训练前率的 1/1）开始 SGD，这允许在不阻碍初始化的情况下进行微调。在每个 SGD 充气中。我们均匀地采样了 32 个正窗口（超过所有类）和 96 个背景窗口，以构建一个大小为 128 的小批量。我们将抽样偏向于正窗口，因为与背景相比，它们极其罕见。

对象类别分类器。考虑训练一个二值分类器来检测汽车。很明显，一个轻微包围汽车的图像区域应该是一个积极的例子。同样的。很明显，一个与汽车无关的背景区域应该是一个负面的例子。不太清楚的是如何标记一个与汽车部分重叠的区域。我们用一个循环重叠部分来解决这个问题。在以

为了对区域进行分类，他们的方法建立了一个四级空间金字塔，并使用密集采样的 SIFT 填充它。扩展对手 I FT.和 RGBSIFT 描述符，每个描述对象都用码本进行量化。使用直方图相交核 SVM 进行分类。与他们的基本特征相比。利用非线性核 SVM 方法，我们在 mAP 中实现了一个较大的小像素点。

VO(吗？201（）测试航空自行车，鸟船，公共汽车，牛（狗马），植物羊沙发火车

DPM 侦它	49	53	13	15	35.	53	49	27	17	2S	14	17	46.	51.2	47.7	10.	34.	20	43	38	33
乌瓦伊吉	56	42	15	12	21.	49	36	46	12	32	30	36	43.	52.9	32.9	153	41.	31	47	44	35
kegionlcts||	65	48	25	24	24.	56	54	51	17	28	3(	35	40.	55.7	43.5	143	43.	32	54	45	39
SegDPM | f	61	53	25	25	35.	51	50	50	19	33	26	40	48.	54.4	47.1	14.	38.	35	52	43	40
R-CNN	67	64	46	32	30.	56	57	65	27	47	40	66	57.	65.9	53.6	26.	56.	38	52	50	50
R-CNN BB	71	65	53	36	35.	59	60	69	27	50	41	70	62.	69.0	58.1	29.	59.	39	61	52	53

R-CNN 是最直接比较 UVA 和区域，因为所有 n 心 hod 都使用选择性搜索区域建议。边界箱回归(BB 在第 3.4 节中描述。公共石灰。SegDPM 在 VOC 排行榜上表现出色。DPM 和 SegDPM 使用的上下文评分我方法没有使用。

下区域被定义为阴性。重叠的阈值。（）.3.通过{（）上的网格搜索选择。0.1.....0.5} 在一个

验证集我们发现，仔细地选择这个阈值是很重要的。设置为 0.5。在[”。将 mAP 降低 5 分。同样，卖给（）也使 mAP 降低了 1 个百分点。这些例子被简单地定义为每个类的地面边界框。

一旦特征被提取并应用训练标签。我们优化了每个类的一个线性 SVM。由于训练雷达太大，无法进入记忆，我们采用了标准的硬负挖掘方法||.我倒了

有效挖掘收敛迅速，在实际中，只通过所有图像就会增加。

在补充材料中，我们讨论了为什么积极和消极的例子在色调调整和 SVM 训练中的定义不同。我们还讨论了为什么有必要训练检测分类器，而不是简单地使用来自 lhe 微调 CNN 的最后一层（fcs）的输出。

2.4. 帕斯卡挥发性有机化合物 2010-12 的研究结果

遵循最佳 VOC 最佳实践| 2|。我们验证了 lheVOC2007dalasel（ϟ）上的所有设计决策和超参数。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 VOC 2012 列车上的检测支持向量机。我们只向两种主要算法的评估服务器提交了一次测试结果（有和没有边界箱回归）。

表 1 显示了 VOC 2010 的完整结果。我们将我们的方法与四种强基线进行了比较，包括 SegDPM [|.它结合了 DPM 探测器与

语义分割系统的输出|4]，并使用额外的插入选择器上下文和图像分类器重新评分。与系统的比较。|32|.因为我们的系统使用相同的区域建议算法。mAP 从 35.1%增加到 53.7%。同时也要快得多（第 2.2 节）。我们的方法在 VOC 2011/12 上达到了类似的性能（53.3%的 mAP）。

3. 可视化、消融术和误差模式

3.1. 正在可视化已学习到的特性

第一层填料可以直接可视化，易于理解[二|]。它们会捕捉到定向的边缘和对手的颜色。理解后续的图层更具挑战性。Zeiler 和 Fergus 在|中提出了一种可视化的地图集解卷积方法]。腿提出了一种简单的（互补的）非辅助方法，直接显示网络学到的东西。

这个想法是在网络中挑出一个特定的单位（壮举我），并把它当作是它自己的目标探测器一样使用。这就是说。我们在大量的持有区域提案（约 1（）00 万）上计算单元的活动，从最高的最低激活中排序提案，执行非最大抑制，然后显示得分最高的区域。我们的方法通过显示它来选择“为自己说话”的输入。我们避免平均，以看到不同的视觉单元，并获得洞察由 lhe 单位计算的不变性。

我们从层池-中可视化单元，这是 lhe 网络的第五层和最后一个卷积层的最大池输出。池特征图是 6x（）x256=921（）维。忽略边界效应，每个池单元在原始的 227 x 227 像素的输入中都有一个 195 x 195 像素的接受域。一个中央池单元有一个几乎全局的视野，而一个靠近边缘的池单元有一个较小的剪辑支撑。

图 3 中的每一行都显示了我们在 CNN 上对 VOC 2007 Irainval 进行了微调的一个池单元的前 16 个激活情况。256 个功能独特的单元中有 6 个是可视化的（补充材料包括更多）。这些

編|匯良虐亩區网繼圖谖画牛倒!：：|

)006

舞 3^岡 iM11j 就岡芝 iT ■•*飢辭 律圓叫面思同蛔亏画側也國回封国&

藏受帮助器的激活值是 dKtun。有些单位与概念相一致，如“人”（第 1 行）或 1 ex 1 (4)。其他单位的结构和恶意适当的谎言，如 1 阵列(2)和螺旋后跟离子(6)。

VOC 2007 测试	航	自	鸟	船	bol	公	小	ca	椅	牛	li	狗	hi	mbi	人	植	羊	沙	有	(	in
R-CNN 池，	5L	60	36	27	23.	52	60	49	是	47	44	40	56	5S.	42.4	23	46.	36	51	55	44
R-CNN 傅提	59	61	43	34	25.	53	60	52	21	47	42	47	52	58.	44.6	25	48.	34	53	58	46
R-CNN	57	57	3S	31	23.	51	58	51	2(	50	40	46	51	55.	43.3	23	48.	35	51	57	44
R-CNN 1 T 池	58	63	37	27	26.	54	66	51	26	55	43	43	57	59.	45.S	2S	50.	40	53	56	47
R-CNN IT Ic.i	63	66	47	37	29.	62	70	60	32	57	47	53	60	64.	52.2	3I	55.	50	57	63	53
R-CNN 1	64	69	50	41	32.	62	71	60	32	58	46	56	60	66.	54.2	3L	52.	48	57	64	54
R-CNN 1'一、沙	68	72	56	43	36.	66	74	67	34	63	54	61	69	68.	58.7	33	62.	51	62	64	5S
DPM v5 |i	33	60	10	16	27.	54	5S	23	20	24	26	12	5S	48.	43.2	12	21.	36	46	43	33
DPM 是	23	5S	10	S.	27.	50	52	7.	19	22	IS	S.	55	44.	32.4	13	15.	22	46	44	29
DPM HSC |2 |	32	5S	H.	16	30.	49	54	23	21	27	34	13	58	51.	39.9	12	23.	34	47	45	34

第 1-3 行显示了没有进行微调的 R-CNN 性能。第 4-6 行显示了 CNN 在 ILSVRC 2012 之前的结果，然后行边缘(FT>在 VOC 2007 的结果。第 7 行包括一个简单的边界框回归(BB >阶段，减少定位错误(第 3.4 节>。第 8-1 行（）将 DPM 方法作为一个强基线。1irsl 只使用 HOG。而接下来的两种方法则使用迪非勒伦特征学习方法来增加或取代 HOG。

单元被选择来显示网络学习的内容的代表性样本。在第二行中。我们看到一个单位，轮胎在狗的脸和点阵列。与第三行对应的单位是一个红色的斑点选择器。还有人脸探测器和更抽象的模式，如文本和带有窗口的三角形结构。该网络似乎学习了一种表示，它结合了少量的类调优特征以及形状、纹理、颜色的不同表示。和材料的属性。随后的全连接层 fet，具有建模这些丰富特征的大量组成集的能力。

3.2. 消融研究

没有逐行调优的性能。为了了解哪些层对检测性能至关重要，我们分析了 CNN 在 VOC 2007 数据集上的最后三层的结果。在第 3.1 节中简要描述了图层池。最后两层总结如下。

层 fc_{6i} 是完全连接到池-。结核病的计算特征。它将一个 1（）96x921（）的权重矩阵乘以池特征映射（重塑为一个 9216 维向量），然后添加一个偏差向量。这个中间向量是逐分量半波整流的 r)).

层 fc：是网络的最后一层。它是通过将 fc(>计算的特征乘以一个 -1096x-409（）权重矩阵，并类似地添加一个偏差向量并应用半波校流来实现的。

*我们从来自 CNN 的结果开始，“没有关于流氓的^^，即。*所有的 CNN 参数都仅在 ILSVRC 2012 上进行了预训练。分层分析性能（表 2 行 1-3)表明，fc 的特性比 fc 的特性更差。这意味着有 29%。或者大约 1680 万，CNN 的参数可以通过可降解的 niAP 被删除。更令人惊讶的是，同时删除两个 fc？和 fc(我产生了相当好的结果，即使池-，特性的计算只使用了 6%的 CNN 的参数。CNN 的大部分代表性权力都来自于它的卷积层，而不是来自于

更大的紧密连接的层。这一发现表明，计算机计算机在计算密集特征图方面具有实用性。在 IIOG 的意义上。通过只使用 CNN 的卷积层的任意大小的图像。这种抑制作用将使滑动窗口探测器的实验成为可能。包括 DPM。在泳池的上。

逐行优化性能。我们现在看看 CNN 在 VOC 2007 训练上微调其参数后的结果。改进是引人注目的（“Rible2 行 4-6）：精细化增加 mAP8。（）个百分点至 54.2%。fc（；微调的提升比池更大，这表明池。从|学习到的特征是通用的，而且大部分改进是通过学习域的非线性分类器中获得的。

并与最近的特征学习方法进行了比较。针对流氓 VOC 检测的特征学习方法相对较少。我们来看看两种最近建立在可变形部件模型上的方法。作为参考，我们还包括了基于 hog 的标准 DPM 的结果[|]。

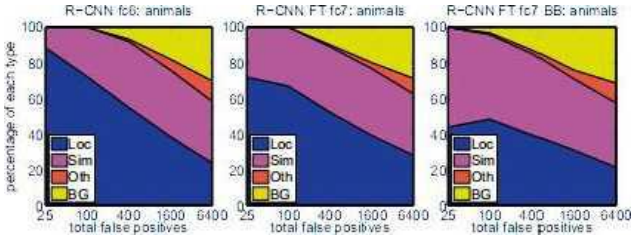
第一种 DPM 是一种有缺陷的学习方法。DPM ST [_].用“草图标记*”概率的直方图来增强 HOG 特征。直观地说，草图标记是通过图像补丁中心的轮廓。草图标记概率由一个随机森林计算，该森林训练将 35 x 35 像素块划分为 15 个（）草图标记或背景中的一个。

第二种方法。DPM HSC |用稀疏码（I ISC）代替 I IOG。强制使用 I-ISC。使用 100 个 7 x 7 像素（灰度）原子的学习字典在每个像素处解稀疏码肌动子。振动激活以三种方式重新关联（全波和两个半波）。空间合并，单元/■>标准化，然后动力伊朗形成(.r-sign（.r）|.r|").

所有的 R-CNN 变体都强烈优于三个 DPM 基线（表 2 行 8-10）。包括那两个使用有缺陷学习的人。与 DPM 的最新版本进行了比较。它只使用了 HOG 功能，我们的 mAP 要高出超过 20 个百分点： 54.2%的 r.v.。33.7%-相对为 61%。I IOG 和草图令牌的组合比 HOG 单独产生 2.5 个 mAP 点，而 HSC 比 HOG 提高了 4 个 mAP 点（与它们的原始 DPM 基线相比——两者都使用了开源||的非公开实现）。这些方法实现的 niap 值分别为 29.1%和 34.3%。

3.3. 检测误差分析

我们应用了 Iloieni 等人的优秀检测分析工具。| ‘)]为了揭示我们的方法的错误模式，了解微调是如何改变它们。看看我们的错误类型与 DPM 的比较。该分析工具的完整总结超出了本文的范围，我们鼓励读者咨询|。”|来了解一些更详细的细节（如“标准化 AP”）。由于分析最好集中在相关图中，我们在图 4 和图 5 的标题中进行讨论。



3.4. 边界框回归

在误差分析的基础上，实现了一种简单的定位误差减少方法。灵感来自于 DPM|-中使用的边界盒回归]。我们训练一个线性回归模型来预测一个新的网格窗口，给定的池特征 i‘或一个选择性搜索区域的建议。详情载于补充资料。表 I 中的结果。表 2。和图 4 显示，这种简单的方法修复了大量的错误定位检测，将 mAP 提高了 3 到 4 个点。

4. 语义分割

区域分类是语义分段的标准分类。允许我们轻松地将 R-CNN 应用到

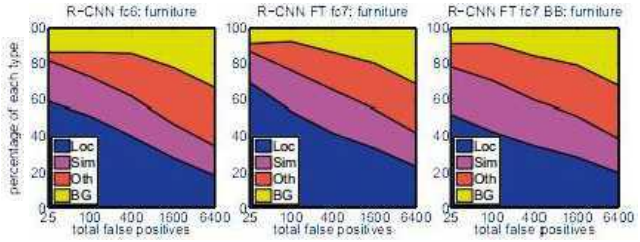


图 4：排名最高的假阳性（FP）类型的分布情况。每个图都显示了随着 FPs 的增加，FP 类型的进化变化。每个 FP 被分为 4 种类型中的 1 种：loc 差定位(在 0.1 到 0.5 之间的正确类。或重复：>：(>：与不同对象类别的融合：BG-一个在背景上触发的 FP。与 DPM 相比（见|2 |）。显然，我们更多的错误是由于糟糕的定位，而不是与背景或其他对象类的混淆，这表明 CNN 的情况更多。我们使用了放大区域建议和 CNN 全图像预训练后的位置不变性。第三列显示了我们的简单边界盒回归尼尔霍德如何减少许多定位错误。

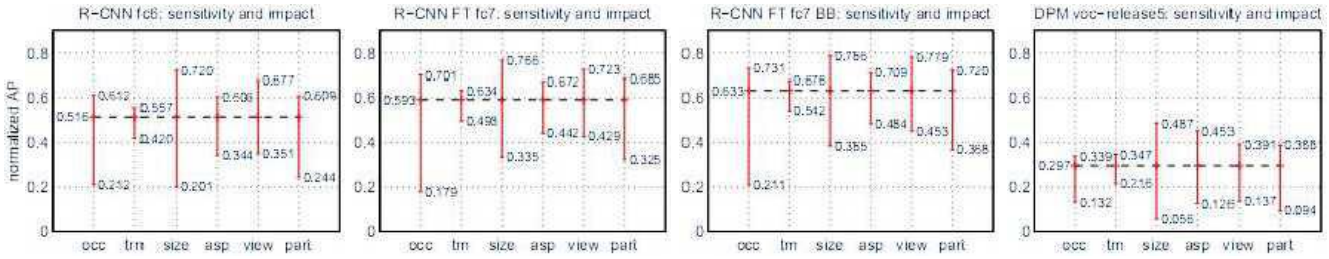


图 5：对对象特征的敏感性。每个图显示了 6 个不同物体中最高和最低穿孔子集的标准化 AP（见|-，|）。边界框区域，方面为零。视点，相对可见性）。我们展示了我们的 inelhod（R-CNN）有和没有 iine 更新（FT）和边界盒回归（BB）以及 DPMvoc 释放 5 的图。全面的线月并没有降低敏感度（最大和最小之间的差异），但显著改善了几乎所有方法的最高和最低表现子集。这表明，线调优做了更多的 Ilian 简单地改进了长宽比和边界盒面积的最低执行子集，正如我们可能根据我们如何扭曲网络输入所推测的那样。相反，线调优提高了所有优点的鲁棒性，包括遮挡、耳轴。视点和部分可见性。

帕斯卡 VOC 分段的挑战。为了便于与当前领先的语义分段系统（称为 0？P 表示“二阶池”）进行直接比较，||。我们在他们的开源框架内工作。O_2P 使用 CPMC 为每幅图像生成 15 个（）区域建议，然后为每个类预测每个区域的质量。使用支持向量回归（SVR）。他们的方法的高性能是由于 CPMC 区域的低质量和多种特征类型（SIFT 和 LBP 的丰富变体）的强大的二阶池化。我们还注意到，Farabet el al. | |最近在使用 CNN 作为多像素分类器的密集场景标记数据集（不包括流氓）上显示了良好的结果。

我们跟着我走。-)，并扩展帕斯卡分段训练集，包括伊拉里|提供的额外注释。设计决策和超参数在 VOC 2011 验证集上进行了交叉验证。最终的测试结果仅被评估了一次。

CNN 的细分功能。我们评估了三种计算 CPMC 区域特征的策略，所有这些策略都首先将该区域周围的矩形窗口扭曲到 227 x 227。第一种策略（fitll）忽略了区域的形状，并在弯曲窗口上计算 CNN 的特征，就像我们在检测时所做的一样。然而，这些特征忽略了该区域的非矩形形状。有两个区域可能有非常相似的边界框，但几乎没有重叠。因此，第二种策略（fe）针叶树脉冲 CNN 只在一个区域的前景掩模上出现特征。我们用均值输入替换背景，使均值减法后背景区域为零。第三种策略(fitll+)简单地连接了///// 和特征；我们的实验验证了它们的针叶树的增大。

VOC 2011 年生产产品的研究结果。表 3 总结了我们对 VOC 2011 验证集与 0 的比较结果 $_2P$.（每个类别的完整结果见补充材料。）在每个特征计算策略中，图层 fc（；

	完整的 R-CNN		fg R-CNN		*fiiH+fii R-CNN*	
0*1		Icr	fc_6	Icr	fC6	fCT
46.4	43.0	42.5	43.7	42.1	47.9	45.8

表 3：分割平均精度（°/（）on \（）C 2011 验证。2-7 使用我们的 CNN 在 ILSVRC2012 上的 CNN。

总是优于 fc 吗？下面的讨论是 fc（i 特性。该策略的表现略优于/h//。表明 lhatl 掩蔽区域形状提供了更强的信号，符合我们的直觉。然而，适合的+fg 的平均精度为 47.9%。我们的最佳结果是 4.2%(表现也略优于 O_2P)。即使给定 fe 特性，指示由完整特性提供的高度上下文也具有很高的信息性。值得注意的是，在我们的全功能上训练 lhe2（）SVr 在一个核心上需要一个小时，而在 0 上训练需要 10+小时 $_2P$ 特性。

在表 4 中，我们给出了 VOC 2011 最低测试集的结果。比较我们的 bcsl-执行方法。*{full+fy)*.阿甘西有两个很强的基线。我们的 nielhod 在 21 个类别中的 1i 达到最高的分割精度，最高，为 47.9%，跨类别（但在任何合理的误差范围下可能与 0jP 结果相关）。但通过进行微调，还可以实现更好的性能。

5.结论

近年来，反对者在民主选举中的表现一直停滞不前。性能最好的系统是将多个低级图像特征与来自对象选择器和场景分类器的高级上下文相结合的复杂集成。本文提出了一种简单且可扩展的目标检测算法，该算法比之前对 PASCAL VOC 2012 的最佳结果相对提高了 30%。

我们通过两个见解实现了这一表现。首先是将大容量卷积神经网络应用于自底向上的区域建议，以定位和分割对象。第二个是火车的范例

VOC 2011 测试	h	航	自	鸟	船	瓶	公	小	猫	椅	牛	表	狗	马	mbi	人	植	羊	索	有	N	意
R&P 匚]	S	4	1	?	3	42	5	4	4	8.	3	36	3	49	48.	50.	26	47	2	4	43	40
。如 14|	8	6	2	4	4	46	6	5	5	13	4		4	59	55.	51.	36	50	2	4	44	47
名词（lull+f i;	S	6	2	5	S	55	7	5	5	9.	4	29	4	40	57.	53.	33	60	2	4	41	47

■比较了两个强大的基线：“地区和部分”*(R&P>>的|。|和||的二阶池化（0？P）方法。你任何任何话。我们的 CNN 实现了分段性能。通过优化 R&P 和粗略地优化 0jP。

当标记的训练数据稀缺时，识别大型 cnn。我们表明，对具有丰富数据的辅助任务（图像分类）进行预训练，然后对数据稀缺<检测的目标任务进行微调是非常有效的。我们推测，“监督公关训练/特定微调”范式将对各种数据稀缺的视觉问题非常有效。

最后，我们指出，我们通过结合计算机视觉和深度学习的经典工具（自下向上上区域建议和卷积神经网络）来实现这些结果是很重要的。而不是反对相反的科学探究路线。这两者都是自然的、不可避免的合作伙伴。

确认。这项研究部分得到了美国 DARPA 心灵之眼和 MS EE 项目的支持，美国国家科学基金会奖 HS-0905647。IIS-1134072. 和 HS-1212798。穆 N000014-10-1-0933。以及来自丰田的支持。本研究中使用的 gpu 是由 NVIDIA 公司慷慨捐赠的。

参考文献

11) B. Alexe。 1.德斯克伦..和诉 Penan。测量(对象为图像窗口。伊帕米。2012.
(2| B Arbelacz.B. 乌安鲁安。C. 顾。S. Gupla.L.Bourdes 和 J. Malik。软骨分段使用区域和部分，hi C\ PR。2012.
|3] R Arbclaez.J.桥-1‘uset。J.巴伦。I'.马奎斯和马利克。多尺度双分组。在 CVPR。2014.
|4| J. Caneira。R. Casein>.J. Balisla.和史密斯。具有二阶池化的离子。在 ECCV。2012.
|5| J.丘埃拉和|。CPMC：使用连续参数的自动对象段。IPAML 2012.
|6] D.Cncsan。A. 朱斯蒂 L. Gambardellu.和迪迪人。利用深度神经网络在乳腺癌组织学图像中的有丝分裂检测，hi MiCCAL 2013。
|7| N. Dakil 和 B. Tnggs。用于人体检测的直方图或相关的梯度。在 CVPR。2005.
|S] T.院长 M. A. Ruzon。\L 西格尔。J. Shlens.S.这是我的名字。和叶拉圣，对 1 个（） （）的准确检测。（） （） （）对象类是单台机器。在 CVPR。201?.
|9) J.邓。A. 冰山 S.萨希什。IL Su。 A. Khski。和飞飞。图像网大规模视觉识别识 别 2012iILS\‘RC2（ ）12i 。’： ••: : 〃:、 w.image-i.et .org/challenges/LSVRC/2012/.
I!（）) J.邓。盾 R. Sovher.L.-J.列支敦士登李 ki。和 l.裴飞。一个大型的分层图像数据库。在 CVPR。2009.
111) 1 端和 D. lloicm。标有独立的奥尔>目标提案。在 ECCV。2010.
112] M. Evenn^hani.L.VanGSubarm>1。C.K. 1 威廉姆斯。J. Winn 和 A.齐森奈伊。B\SCAL 可视化对象类（VOC）挑战。Z/CV。20!().
113] C. 帕拉贝。C. 优惠券。L. Najinan.和勒村。学习场景标记的分层特征。*!PAML 2013.*
|I4| P. Felzenszwalb.R. Gnshick.D. 麦卡赫斯特公司，和 D. 雷南南。带有训练有素的监狱。LPAML 2010.
115| S. Fidler。R. Mottaghi.A. 耶。和乌尔塔森。机器人(自上而下的程序程序。在 CVPR。2013.
|16| K.福岛。新认知：一种模式认知机制的研究，不受位置变化的影响。生物控制论。36(4):193-202. 198().
(17] R. Gushick.I< relzenszwalb.和 1)。McAHester.DiscriniinativeJy tidined deibntuiblc pari niodcls. 版本 5- ： ； r.tp ： //■. ：■。:.■: .cs .Berkeley.edu/’rbg/latent—v5/.
|I8| C.顾。J. J. Lun。R 弧。和马利克。识别使用区域。在 CVPR。2009.
|19| B. Hariharan.R Arbclaez.L. Bourdev.S. \kiji.和马利克。来自反向探测器的语义桨。在 ICCV。2011.
(2()] D. Hoicni.Y. Chodpathuiiiuan.和戴问题。对象检测器中的诊断错误。在 ECCV。2012.
|2 11 Y.贾。最后一个特性嵌入的开源卷积计算，http: Z/ca： .如果 v 是 i o n. org/\
2013.
|22| A.库兹耶夫斯基。1 Suiskeser.和 g.辛顿。利用深度卷积神经网络进行分类。在 NIPS。2012.
(23] Y. LeCun。B. 波沙牌手表 J.丹克。D. 亨德森 R.霍华德。\V.槽和 L. Jackcl。传播应用于手写压缩>识别。*神经补偿*。. 1989.
|24| Y. LeCun。L. Bottou。Y.本·吉奥。和 P. Ha 衬垫。基于梯度的文档识别方法应用了>的文档识别方法。1998.
|25) J. J.伦。C.L. Ziimck.和 R 美元素描标记：一个学习的轮廓的 n 级表示和嗅觉检测。在 CVPR。2013.
|26| D.劳。显示图像为 fn>比例-图像关键点。IJCV。2004.
|27| X. Ren 和 D.雷宁金。用于 objcul 检测的代码的直方图。在 CVPR 中。2013.
|2S| II. A.行。S. Baluja.和 1：Kanadc。基于神经邻居的人脸检测。伊帕米。I99S.
|29| R Serinancl.K. Kaxukcuogki.S. Chiniala.和 LeCun。我们的学习是简单的。在 CVPR。2011
|3（）| K. Sung 和 T. Poggio。基于例子的学习，基于简单的人脸检测。技术报告。备忘录 No. 1521 年。马萨诸塞州的技术要求。1994.
|3 I] C.大小写。A. Toshev，和 D. 伊桑深度神经网络负责目标检测。在 MPS。2013.
|.’2| J. Uijlings.K.\一个德萨姆勒。T：速度。和 a。选择搜索对象识别。IJCV。2013.
|33] R. Vdillani.C. 蒙罗。和勒村。对图像中的物体进行定位的原始方法。视觉演示演示，模拟^。和信号处理。1994.
|34| C. Vondnck.A. 霍斯基。T. Malisiewicz.和钛>。HOGglcs: xisualizingobject deteuiion I’eatuivs.K‘CV。2013.
|35| X.王。M.杨。S. Zliu.和林。用于通用的嗅觉检测。InlCCV。2013.
I?61 米。泽勒。G.支付法律价值。和 r.费格斯。用于中、高级特征学习的自适应反卷积网络。在 CVPR。2011.