指导教师：杨涛　　提交时间：16.3.21

CVPR2015 Paper Translation

No: 01

姓名：吕长命

学号：2013302556

班号：10011304

卷积功能屏蔽的联合对象和物体分割

作者：Jifeng Dai Kaiming He Jian Sun

Microsoft Research

{jifdai,kahe,jiansun}@microsoft.com

译者：吕长命

摘要：

语法分割已经看到了相当大的进步，因为卷积神经网络的强大学习特征。最近语法分割的前沿研究是通过从模糊图像区域提取卷积神经网络特征，探索图形信息。这种方法提出在图像方面人工的局限，并且可能影响提取的特征质量。而且在未处理过的图像领域操作，需要在一幅图像上计算上千个神经网络，这个非常耗时。

在这篇文章中，我们提出通过模糊卷积方法探索图形信息。该提案在卷积特点图上被当作面具（遮罩层）/*选一*/。分片的卷积神经网络特点被从这些图中直接遮盖掉，并且用于识别分类。我们进一步提出一个联合方法在相同的框架中处理对象和事物。最先进的计算结果用令人信服的计算速度在 PASCAL VOC 基准和 PASCAL-CONTEXT 上演示。

1、介绍

语法分割[14, 19, 24, 2]目的是把每一个图形像素标记为语法类别。最近在卷积神经网络上的重大突破已经大大提高了基于 R-CNN 方法的语法分割的技术水平。语法分割的 R_CNN 方法提取 CNN 特征的两种类型，一种是，从建议包围盒提取的区域特征；另一种是从被分割遮盖的原生图形中提取分割特点。这些特点的一系列相关事物被用于训练分类。这种方法在这个长期挑战性的项目中已经演示了一个可靠的结果。然而基于原生图像的 R_CNN 方法有两个小问题。第一在图片内容上的遮罩可能导致“人工边界”。这

些边界在提前训练的网络中的样本上并未出现。这个问题可能降低所提取的分割特点的质量。第二类似于 R_CNN 方法对物体的识别，这些方法需要把网络应用到成千上万的有（没有）遮罩的原生图片区域。这是非常耗时的甚至在高端 GPU 上也是如此。

第二个问题同样存在于基于对象识别的 R_CNN 中。幸运的是，它在很大程度上可以通过最近一种称为 SPP_Net 的技术解决。这种技术只在整幅图像上计算一次卷积特点图并且应用一种空间金字塔池的技术形成对于分类的裁剪功能。通过这些裁剪特征的检测结果表明有竞争力的检测精度并且速度可以提高 50 倍以上。因此，在这篇文章中我们提出一个问题：对于语法分割，我们可以只用卷积特点图吗？

这个工作的第一部分对这个问题做了肯定的回答。我们设计了一种卷积特点屏蔽法，直接从特点图而不是原生图像中直接提取分割特征。用区域建议法给出的分割，我们预测他们到最后卷积特征映射域。投影段对屏蔽卷积功能表现为二进制函数。被屏蔽的特征然后被送到识别的完全连接层。

因为卷积特征是从未屏蔽的图像上计算得到的所以其质量不受影响。因而这种方法在卷积特征映射上是有效的，且仅需计算一次。前面提到的两个语义分割问题就这样解决了。图 1 比较了基于原始图像的流水线和我们基于特点映射的流水线。这篇文章的第二部分为了链接对象和事物进一步一般化我们的方法。和对象不同的是，事物（例如：天空，草地，水）通常被作为图像中的环境。事物大部分表现为颜色或纹理，并且有着不易确定的外形。因此用一个简单的矩形框或简单分割来描述事物是不合适的。基于我们的屏蔽卷积特征，我们提出了一个训练过程——把一个事物作为一个多分割特征的紧密组合体。这使我们能够解决在同一框架内的对象和事物。

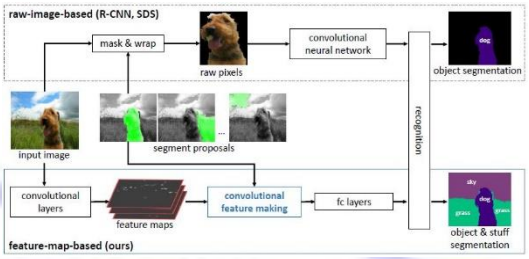

图 1

基于以上方法，我们在 PASCAL_VOC2012 基准上展示了高技术的成果。我们的方法可以在少于 1 秒钟的时间内处理一幅图像，这比基于 SDS 的 R_CNN 方法快乐 150 多倍。而且我们的方法也是第一个基于深度学习的方法，曾被应用到新标记 PASCAL_CONTEXT 标杆，在对象和事物分割两方面我们的结果都大大由于以前的最优技术。

2、卷积特征屏蔽法

2.1 卷积特征屏蔽层

卷积神经网络作为一个一般特征提取器已经逐渐在计算机视觉领域显示出它的强大之处。在 Krizhevsky——et——al 的工作中，他们建议在全连接层的特征可以被用来作为整幅图像的特征，例如：图像检索。在【6，25】这些全面的功能是作为通过迁移学习通用的功能在其他数据集全图像分类任务。在破目标检测的 R_CNN 的文章中。CNN 特征也被当作整体特征使用，但要从原生图像修剪后的子图像中提取。在基于 CNN 的语法分割的文章中，R_CNN 方法一般化为遮罩原生图像区域。对所有的方法来说，整个网络被训练为完整特征提取器，无论是在整个图像或子图像中。在 SPP_Net 最近的工作中，它表明，该卷积特征图可以用作局部特征。在一个整幅图像的卷积特征图中，局部矩形区域语义信息（通过激活的优势）和空间信息（通过位置）编码。这些局部区域的特征可以直接汇集起来做识别。

SPP 在【11】实际上扮演两个角色：1）用一个矩形区域屏蔽特征图，之外的激活部分被移除。2）产生从该任意大小的区域的定长特征。所以，如果用一个矩形屏蔽可以是有效的，我们要是用一个良好的不规则分割图形屏蔽特征图会如何呢？

卷积特征屏蔽层就这样产生了。我们首先获得原始图像上的候选段（如超象素）。很多局部建议法是基于超像素的。每一个建议框通过分组几个超像素给出。我们称这样的一个组为分割提议。因此我们可以同时获得候选段和他们的提议框（在文中称为区域）而不用额外的努力。这些段都在原始图像的二进制屏蔽。

接下来我们预计这些二进制口罩的域最后卷积特征图。因为在卷积特征图中每个激活区域是由一个在图像区域的感受视野贡献的。我们首先预测每一个激活区到图像区域作为它的接受视野的中心（在【11】中的细节之后）。在图像上的二进制掩模的每个像素被分配到其最接近的感受野中心。之后

这些像素被预测回到卷积特征图域中，基于中心和它的激活位置。在特征图中，每个位置将要从二进制屏蔽中收集多个被预测的像素。然后对这些二进制值求取平均值和阈值（0.5）化操作。这给了我们一个在特征图中的屏蔽层（图 2）。这个面具被用到卷积特征图中。实际上我们只需要在特征图的每一个通道上乘以这个二进制屏蔽。在我们的方法中称最终特征为分割特征。

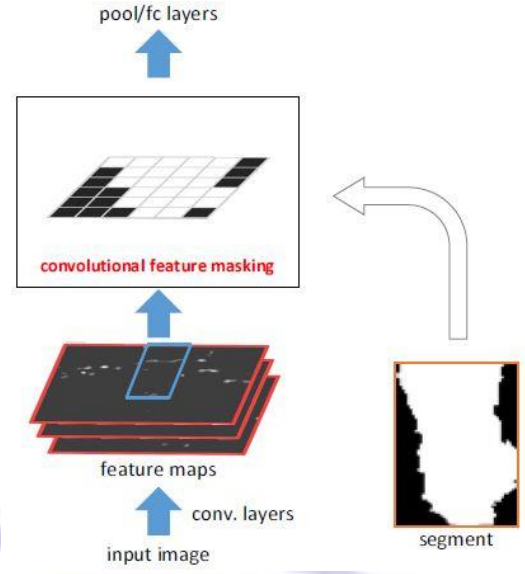

图 2

2.2 网络设计

在【10】中已经展示仅有分割特征是不够的。这些分割特征应该和被用一种像 R_CNN 方法一般化的原始特征(来自边界框）一块用。基于我们的 CFM 层，有两种方法可以这样做。

设计 A：在最后的卷积层：就像在图 3（左部）展示的一样，在最后的卷积层之后我们一般化了两种特征源。一种是由 SPP 层产生的区域特征，就像在【11】中一样。另一个是由接下来的方法产生的分割特征。CFM 层被用到整幅图像的卷积特征图中。这给了我们一个任意大小（就边界框而言）的段特征。然后我们用另一个 SPP 层在这个特征上产生一个固定长度的输出。两类汇集特征被送到两个分开的 FC 层中。最后的 FC 层的特征被级联来训练分类器，就像【10】中的分类器。在本设计中，我们在 FC 层的训练和测试中有两种途径。

设计 B：空间金字塔汇集层。我们首先采用 SPP 层来汇集特征。我们用一个{6 × 6, 3 × 3, 2 × 2, 1 × 1}的 4 层金字塔像在【11】中的一样。6X6 层实际上是一个很小的 6X6 的特征图，它仍有足够的空间信息。我们把 CFM 层用在这个极小的特征图中产生分割特征。这个特征则与其它三个层拼接后送到 FC 层，想在图 3 右部展示的一样。

在这个设计中，我们保留一种 fc 层的通路来减少计算开销和过拟合的风险。

2.3 训练和推理

基于这两个设计和 CFM 层，顺着在【8，11，10】中的普通练习，训练和推理阶段可以很容易进行。在这两个阶段中，我们用区域提案算法产生大约 2000 个区域提案和相关的段。输入图像调整为多尺度（较短的边 s《{480, 576, 688, 864, 1200}），并且
卷积特征图从全图像中提取然后固定（不会进一步调整）。

训练。我们首先用 SPP 方法为物体检测微调一个网络。然后我们用设计 A 或 B 中的体系架构更换微调的网络，然后进一步为分割微调网络。在第二个微调步骤中，段提案覆盖地面实况前景的【0.5-1】是效果较好的，【0.1-0.3】效果是相当差的。该覆盖由基于两个段区域（而不是他们的边界框）的 IoU 得分评定。在微调后，我们训练一个在网络输出上的现行 SVM 分类器，为每一个类别。在 SVM 训练中，仅有地面实况段被用作良好样例。

推理。每一个区域提案被分配一个合适的尺度如在【11】中的一样。每一个区域的特征和他的相关段被提取出来如在设计 A 或 B。SVM 分类器被用作给每一个区域打分。

给出所有的评分区域提案，我们通过在 SDS 通过的方案获取像素等级分类标签。这个通过的方案顺序选择最高分的区域提案，执行区域细化，抑制重叠的建议，和把像素标签粘贴到标签的结果上。区域细化提高了 PASCAL VOC 2012 大约 1%的精度为 SDS 和我们的方法。

2.4 在对象分割上的结果

我们在 PASCAL VOC 2012 语法分割基准上评估我们的方法，那有 20 个对象类别。我们遵从“comp6”评估协议，那也被用在【4，8，11】。PASCAL VOC 2012 的系列训练和来自【9】附加的分割注释被用来训练和在【4，8，10】中评估。研究了两个场景：语法分割和同时检测和分割。

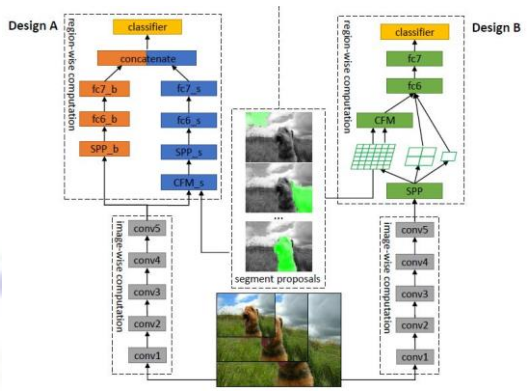

图 3

场景 I：语法分割

在语法分割的试验中，类标签被分配给图像中的所有像素，并且精确度有区域 IoU 得分评估。

我们首先研究用“ZF SPPnet”模型作为我们的特征提取器。这个模型是基于 Zeiler 和 Fergus 的快速模型但是带有 SPP 层。它有五个卷积层和三个 fc 层。这个模型适合【11】的代码一起发布的。我们注意到，在 R-CNN[8]和 SDS[10]的结果用“AlexNet”[13]来代替。要了解预先训练模式的影响，我们以 PASCAL VOC 2012 系列数字的形式报告它们的物体探测 mAP：SPP-Net (ZF) 是 51.3%，R-CNN (AlexNet) 是 51.0%，和 SDS (AlexNet) 是 51.9%。这意味着这两个预训练的模型是可作为一般特征提取器的。所以 CFM 以下的收益并不只是由于预先训练的模型。展示出 CFM 层的作用，我们展现出一个没有 CFM 的基准线——在我们的设计 B 中，我们移除了 CFM 层但仍用相同的整个流水线。我们称这种基线为我们方法的“无 CFM”的版本。实际上，这个基准线降低到原始 SPP-net 使用，除了正/负样本的定义是分割。表 1 无 CFM 的结果和 CFM 的两种设计比较。我们发现，CFM 具有明显的在非 CFM 基线优点。这是预料之中的，因为无 CFM 基线也没有任何基于细分特点。此外，我们发现，设计 A 和 B 只是表现相当，而 A 需要计算 FC 层的两种途径。因此，本文的其余部分，我们采用 B 型设计的 ZF SPPnet。

在表 2 中，我们使用不同的区域建议算法评估我们的方法。我们采取两种方案算法：选择性搜索（SS），和多尺度组合分组（(MCG)）。遵从【10】中的协议，“快”模式用于 SS，“准确”模式用于 MCG。表 2 显示，我们的方法实现在 MCG 建议更高的精度。这表明我们的特征掩蔽方法能够通过更准确的分割建议生成信息。

no-CFM	CFM (A)	CFM (B)
43.4	51.0	50.9

表 1

	ZF SPPnet	VGG net
SS	50.9	56.3
MCG	53.0	60.9

表 2

	ZF SPPnet	VGG net
5-scale	53.0	60.9
1-scale	52.9	60.5

表 3

	conv time	fc time	total time
SDS (AlexNet) [10]	17.8s	0.14s	17.9s
CFM, (ZF, 5 scales)	0.29s	0.09s	0.38s
CFM, (ZF, 1 scale)	0.04s	0.09s	0.12s
CFM, (VGG, 5 scales)	1.74s	0.36s	2.10s
CFM, (VGG, 1 scale)	0.21s	0.36s	0.57s

表 4

在表 2 中，我们还评估预训练网络的影响。我们比较 ZF——SPPnet 与公开 VGG——16 模型【20】在图像分类中的最新进展表明，非常深的网络[20]可以显著提高分类准确率。该 VGG-16 模型有 13 卷积和 3 层 FC。因为该模型具没有 SPP 层，我们考虑它最后的聚集层（7×7），是一个{7×7}单级金字塔特殊 SPP 层。在这种情况下，我们的设计 B 不适用，因为没有粗糙层。因此我们用设计 A 代替。表 2 表明，在使用 VGGnet 我们的结果时大幅改善。这表明，我们的方法，从比较有代表性的特征通过更深层次的模型学习而获益。

在表 3 我们评估图像尺寸的影响。除了使用 5 尺度，我们只是从单级图像提取特征，它的短边为 S=576。表 3 表明，我们的单尺度变体具有可忽略的降解。但单规模变体具有更快的计算速度如表 4。

接下来，我们与国家的最先进成果的比较在 PASCAL——VOC2012 测试表 5 中设置。在这里，SDS【10】在这方面以前最先进的方法，并且 O2P[4]是一个领先的非基于 CNN 的方法。我们方法用 ZF——SPPnet 和 MCG 取得了 55.4 分。这比在【10】中报道的 SDS 结果提高了 3.8%，它用的是 AlexNet 和 MCG。这表明，我们的 CFM 方法在没有掩盖原始像素图像的条件下可以产生有效的特征。用 VGG——net，我们的方法对试验组的得分为 61.8。

除了高精确性，我们的方法比 SDS 快得多。用 SDS 和我们的方法在特征提取步骤中运行时间如表 4 所示。这两种方法都是基于 Caffe 库并在 Nvidia 的 GTX GPU 泰坦上运行。时间是取从 PASCAL——VOC 随机的 100 张图像的平均值。用尺度 5，我们用 ZF——SPPnet 的方法比 SDS 快了 47 倍以上；用尺度 1，我们用 ZF——SPPnet 的方法比 SDS 快 150 倍以上而且更加精确。速度改善是因为我们的方法只需要计算一次特征图。表 4 表明我们的方法在用 VGGnet 时仍是可行的。

与我们的工作同步，一种全卷积网络（FCN）方法被提出用于语法分割。它有一个分数（在测试集上 62.2）与我们的方法相当，且速度快因为它也只需在整幅图像上进行一次卷积。但是 FCN 不能产生一个实例明智的结果，这是 10 中的另一个评价指标。我们的方法也可以适用于这种情况下，如下评估。

场景 II：同时检测和分割

在同时检测和分割的评估协议中，所有的对象实例和它们的分割遮罩都被标记。和语法分割相对比，这个场景要求除了标记逐像素的语法分类还要进一步识别不同的对象实例。精度是由在【10】中定义的 APr 得分确定的。接下来我们公布在 VOC2012 确认集的 APr 平均值，作为地面真标签测试集不可用。如在表 6 中展示的，当用 ZF——SPPnet 和 MCG 时我们的方法的 APr 平均值为 53.2。这要比在【10】中公布的 SDS 结果要好。用 VGGnet 我们的 APr 平均值为 60.7，那是这方面的国际最先进结果。注意到，FCN 方法在评价公知 APr 平均值是不适用的。因为它不能产生对象实例。

	mean	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
O_2P [4]	47.8	64.0	**27.3**	54.1	39.2	48.7	56.6	57.7	52.5	14.2	54.8	29.6	42.2	58.0	54.8	50.2	36.6	58.6	31.6	48.4	38.6
SDS (AlexNet + MCG) [10]	51.6	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5
CFM (ZF + SS)	53.5	63.3	21.5	59.1	40.3	52.4	68.6	55.4	66.6	25.4	60.5	48.5	60.0	53.6	58.6	59.8	40.5	**68.6**	31.7	49.3	53.6
CFM (ZF + MCG)	55.4	65.2	23.5	59.0	40.4	61.1	68.9	57.9	70.8	23.9	59.4	44.7	66.2	57.5	62.1	57.6	44.1	64.5	42.5	52.9	**55.7**
CFM (VGG + MCG)	**61.8**	**75.7**	26.7	**69.5**	**48.8**	**65.6**	**81.0**	**69.2**	**73.3**	**30.0**	**68.7**	**51.5**	**69.1**	**68.1**	**71.7**	**67.5**	**50.4**	66.5	**44.4**	**58.9**	53.5

表 5

method	mean AP^r
SDS (AlexNet + MCG) [10]	49.7
CFM (ZF + SS)	51.0
CFM (ZF + MCG)	53.2
CFM (VGG + MCG)	**60.7**

表 6

3. 链接对象和事物分割

在自然图像中的语法分类可以被大致分为对象和事物。对象有确定的外形且每一个实例都是可数的，而事物有确定的颜色或纹理且表现为任意形状，如，草地，天空和水。因此不同于对象，事物区域是不恰当的被表示为矩形区域或边界框。当我们的方法可以产生段特征，每一个段仍然和一个边界框相关联因为他的产生方法。当区域或段提议被提出后，事物可以被一个单一段完全覆盖是很稀少的。即使事物被单个矩形框区域覆盖，几乎可以确定在这个区域的很多像素并不属于这个事物。因此事物分割有着不同于对象分割的问题。

接下来我们将展示我们的框架的一个推广，以减少在事物中包含的问题。我们可以通过一个单一的解决方案，同时处理对象和东西。尤其，卷积特征图仅需要被计算一次。因此如果算法被要求进一步处理事物会有一个较小的额外的开销。

我们的推广是训练期间修改样品的底层概率分布。不同于同等对待样本，我们的训练将要有偏差的对待提议并尽可能紧凑（如下讨论的）的覆盖事物。一个段追求过程被提出来发现紧凑提案。

3.1 由段组合代表的事物

我们把事物作为一个多段建议的组合体。我们期望一个段提议可以尽可能多的覆盖一个事物的局部。与此同时，我们希望这些段提议能够紧凑一些，段越少越好。

我们首先为事物分割定义一个段建议的候选集。我们定义一个纯分数作为一个段建议和段的边界框内事物部分的 IoU 比值。其中在单个图像所有的段建议，那些具有高纯度的分数（》0.6）和事物组成潜在的组合候选集。要从此候选集产生一个紧凑的组合，我们采用类似匹配追踪程序[23，17]。我们依次从候选集中挑选段无需更换。每一步最大的段建议被选中。被选中的段然后抑制在候选集（他们以后也不会被选到）中高度重叠的建议。在本文中，抑制重叠阈值设置为 IOU= 0.2。重复该过程，直到剩余的段都具有比阈值更小的区域，这是分段区中的（图像集的）初始候选集的平均。我们称这个程序段的追求。

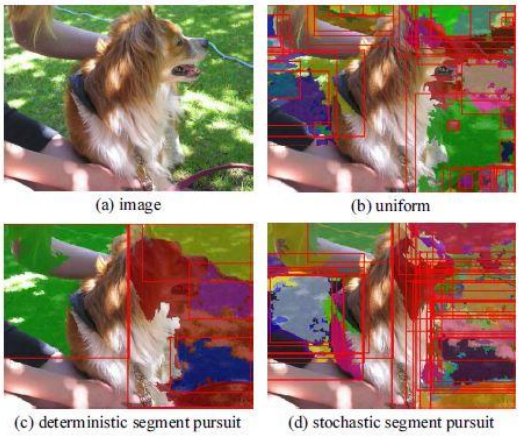

图 4

图 4（b）展示了一个例子如果段的提议是随机从候选集合采样。我们看到那有很多小的段。定义这些小段是不利的，识别力差的段要么是积极的要么是消极的样本（例如：由 IoU 评估）-如果他们是积极的，他们只是事物的一小部分；如果他们是消极的，他们共享像事物的大部分一样的纹理和颜色。因此我们更愿意在训练中忽视这些样本，所以分类器不会偏见对待这些小样本的任一方。图 4（c）展示了由段追求选出的段建议。我们看到他们可以仅通过几个很大的段覆盖事物（这里是草地）。我们期待一个更多依赖这样一个提议组合的解决方法。

然而，上述过程是确定性的，并且可

以给出每一幅图像的小样本集。例如，在图 4（c）它仅提供了 5 个段建议。在微调程序中，我们需要用来训练的大量随机样本。因此我们在上述段追求程序中注入随机性。在每一步中，我们从候选集中随机抽取段建议，而不是用最大的。采摘概率正比于一个区段的区域大小（因此一个大的仍是更有可能的）。这可以以随机的方式给我们另一个紧凑的组合。图 4（d）展示了一个在几次试验中产生的段建议的样本。

由这种方式给出的所有段建议被认为事物分类的积极样本。那些纯净分数低于 0.3 的段建议是消极样本。这些样本可以被用作微调和 SVM 训练如下详述。

在微调阶段，在每个时期的每个图像产生一个随机紧凑组合。所有在这个组合的所有图像段建议组成这个时期的样本。这些样品是随机置换并送入 SGD 解算器。虽然现在看来样本相互独立于 SGD 解算器，它们实际上是由段追求的共同规则采样。其内在的概率分布会影响 SGD 求解。这个程序每个时期都要重复。对 SGD 结算期来说，我们在 200K 的小批量训练程序后停止训练。对 SVM 训练，我们只用了确定性的段追求给出的单一组合。

用这种方式，我们可以对待事物和对象在仅用作对象的相同框架内。仅有的不同是事物样本是由一种段追求提出的方式提供的，而不是纯粹的随机性。为了平衡分类的不同，对象的一部分，事物，和在每一个小批量的背景样本被设置为约为 30%,30%,和 40%。测试阶段在仅有对象的事件中是相同的。当测试阶段未改变时，被学到的分类器对那些紧凑的建议是不同的。

图 5

3.2 联合对象和事物分割的结果

我们在新标记的 PASCALCONTEXT 数据集上为联合对象和事物分割进行了实验。在这个丰富的数据集上，每一个像素是用一个语法分类标记。它是一个有不同图像的有挑战性的数据集，多样的语义范畴，和对象/事物像素的平衡率。由【18】的协议得出，语义分割是在最常见的 59 类和一个背景类别进行（表 7）。分割精确度是由超过 60 种的 IoU 平均分评估的。由【18】得出，一个 33 种比较简单的类别上（由【18】

分类）的一个子集的平均分，也被报道

			ZF+SS			VGG+SS	VGG+MCG
	Super Parsing	O_2P	no-CFM	CFM w/o SP	CFM	CFM	CFM
mean	-	18.1	20.7	24.0	26.6	31.5	**34.4**
mean on †	15.2	29.2	32.4	37.2	40.4	46.1	**49.5**
aeroplane†	19.5	36.4	20.5	37.6	42.9	**48.9**	47.5
bicycle†	11.3	23.5	32.2	39.1	40.3	41.2	**48.0**
bird†	4.1	24.6	27.9	40.5	46.6	52.9	**59.0**
boat†	0.0	22.3	16.6	29.9	34.0	33.6	**37.7**
bottle†	1.2	15.0	40.0	39.0	39.8	41.5	**51.6**
bus†	14.0	43.2	50.0	52.4	53.5	61.0	**65.2**
car†	15.0	33.5	41.0	44.0	47.1	53.7	**57.2**
cat†	20.1	36.7	45.1	54.1	56.1	60.0	**67.4**
chair†	2.9	6.8	13.6	15.7	17.7	22.9	**24.6**
cow†	0.1	16.2	28.5	34.8	39.8	52.4	**58.9**
table†	6.4	7.0	12.1	12.3	13.9	11.5	**16.7**
dog†	11.5	26.9	39.5	48.3	51.4	57.6	**63.7**
horse†	2.0	26.4	33.0	40.2	43.1	**50.5**	**58.0**
motorbike†	14.3	32.8	40.4	45.1	47.9	54.8	**55.0**
person†	30.1	44.5	47.4	51.0	54.5	59.9	**65.0**
pottedplant†	1.1	15.9	31.4	31.5	34.9	34.1	**41.1**
sheep†	4.2	23.7	29.3	45.5	56.3	59.6	**60.7**
sofa†	3.6	16.1	15.2	19.1	22.0	22.1	**31.8**
train†	10.4	26.7	33.6	39.1	43.0	49.0	**56.1**
tv/monitor†	9.0	24.3	40.3	41.0	40.7	**50.4**	50.3
sky†	65.6	75.6	64.9	70.3	76.8	**80.6**	76.8
grass†	45.3	56.0	51.9	56.9	60.7	**66.1**	66.1
ground†	24.0	27.6	22.0	20.5	22.8	38.3	**39.4**
road†	15.8	31.2	25.3	30.6	34.0	36.2	**37.8**
building†	19.8	24.3	25.0	28.2	32.4	37.9	**39.5**
tree†	37.8	44.3	44.2	50.9	53.4	**59.8**	58.0
water†	34.5	54.8	51.4	54.1	59.7	65.3	**69.1**
mountain†	8.8	19.2	14.9	20.8	18.4	**36.7**	35.6
wall†	30.8	40.5	28.7	36.4	40.4	42.5	**43.8**
floor†	14.4	25.7	23.0	28.1	31.7	35.9	**38.9**
track†	17.5	29.5	35.7	27.3	31.9	**40.1**	38.2
keyboard†	0.1	18.2	26.1	30.2	25.1	36.7	**39.8**
ceiling†	6.4	12.7	19.3	13.4	20.3	**27.9**	23.8
bag	-	1.2	0.5	2.1	2.8	2.1	**9.0**
bed	-	0.7	0.0	1.2	**3.0**	1.1	2.9
bedclothes	-	0.0	4.0	9.9	11.9	13.7	**16.6**
bench	-	0.1	0.0	0.0	0.1	0.0	**0.2**
book	-	5.0	15.0	8.9	14.8	**20.5**	20.1
cabinet	-	4.4	4.2	5.0	8.7	11.4	**18.4**
cloth	-	1.8	0.2	2.8	**3.4**	2.7	2.7
computer	-	0.0	0.0	3.8	5.0	4.5	**8.6**
cup	-	1.4	7.4	8.1	12.4	21.2	**25.5**
curtain	-	11.6	12.5	9.2	15.5	21.4	**25.1**
door	-	2.3	3.9	5.0	6.2	**12.7**	4.9
fence	-	6.6	8.4	4.9	7.3	20.7	**23.3**
flower	-	6.8	4.1	9.0	8.4	10.7	**28.0**
food	-	10.7	22.7	23.3	29.8	35.2	**38.1**
mouse	-	0.9	1.4	4.4	9.9	**12.4**	12.3
plate	-	5.6	7.5	7.0	10.7	19.3	**21.8**
platform	-	7.5	14.7	16.2	18.4	18.7	**26.7**
rock	-	6.7	8.1	13.5	15.0	24.2	**26.4**
shelves	-	3.7	1.8	1.5	3.6	3.1	**10.4**
sidewalk	-	0.5	0.0	1.7	2.5	**7.9**	6.8
sign	-	7.0	1.6	4.1	8.7	**18.7**	17.2
snow	-	16.4	19.0	23.9	28.9	28.3	**40.5**
truck	-	0.2	0.3	0.4	4.5	3.4	**11.0**
window	-	14.6	10.7	11.9	12.4	**21.7**	19.2
wood	-	0.8	0.5	2.4	2.8	1.9	**5.8**
light	-	8.5	11.5	6.2	10.9	15.3	**20.5**

表 7

在这 60 中分割任务重。训练和评估分别表现在训练和分数集上。我们和两个领先方法相比-SuperParsing【21】和 O2P【4】，他们的结果报道在【18】。为了公平比较，该区域细化【10】未用在所有方法中。通过时间表和在 O2P 的一样。在这次比较中，我们忽视

R_CNN和SDS因为他们还没有为事物而发展。

表 7 展示了 IoU 分数的平均值。这里“no-CFM”是我们的基准线（no CFM，没有段追求）；“CFM w/o SP”是我们的 CFM 方法但没有段追求；且“CFM”是我们的 CFM 方法带有段追求。当段追求未使用时，积极的段样本不同于候选集（其中的段的纯净分数》0.6）的样本.

SuperParsing 在较简单的 33 个类别

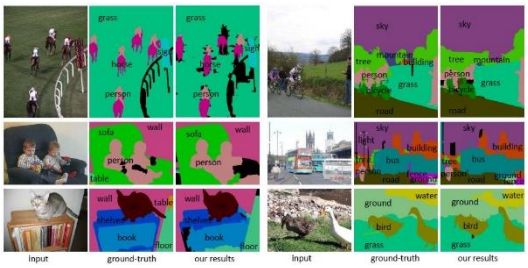

图 5

上得到了一个 15.2 的平均分，且所有的分类分数在【18】中是无用的。O2P 方法结果在 33 个较简单类别上平均分 29.2 在整体分数为 18.1，被报道在【18】。这两种方法是基于 CNN 特征。对于基于 CNN 的结果，no-CFM 基准线（20.7，用 ZF 和 SS）已经优于 O2P(18.1)。这主要是因为通过深度网络学到的一般特征。我们的没有段追求的CFM方法把整体分数提高到24.0。这表明卷积特征屏蔽法的影响。用我们的段追求，CFM 方法进一步提高整体分数到 26.6。这证明了通过段追求所产生的样本的影响。当通过 VGG 网络替换 ZF SPPnet 时，和通过 MCG 替换 SS 建议时，我们的方法产生了一个超过 34.4 的分数。因此我们的方法从深度模型和更精确的段提议中获益。我们的一些结果展示在图 5 中。

值得说明的是，在这个数据集中尽管只评估平均 IoU 分数，我们的方法也能够产生一个用于对象的良好结果。

4. 结论

我们已经展示卷积特征屏蔽，它探索形状信息在网络中的一个稍晚的阶段。我们已经进一步展示卷积特征屏蔽对链接对象和事物分割是适用的。我们计划进一步研究通过卷积特征屏蔽提高对象检测。探索连接对象和物体分割的内容信息也会是感兴趣的。

参考文献

[1] P. Arbel´aez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. CVPR, 2014.

[2] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In CVPR, 2011.

[3] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In CVPR, 2011.

[4] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In ECCV. 2012.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition.

arXiv:1310.1531, 2013.
[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.
[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524, 2013.
[9] B. Hariharan, P. Arbel´aez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In ICCV, 2011.
[10] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In ECCV. 2014.
[11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. arXiv preprint arXiv:1406.4729, 2014.
[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.
[14] M. P. Kumar, P. Ton, and A. Zisserman. Obj cut. In CVPR, 2005.
[19] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation. In ECCV, 2006.
[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[21] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In ECCV. 2010.
[22] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. IJCV, 2013.
[23] Y. N.Wu, Z. Si, H. Gong, and S.-C. Zhu. Learning active basis model for object detection and recognition. IJCV, 2010.
[24] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. In CVPR, 2010.
[25] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901, 2013.