DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs

Abstract

在这项工作中，我们通过深度学习解决了语义图像分割的任务，并做出三个主要贡献，通过实验证明具有实质性的实用价值。首先，我们使用上采样滤波器或“萎缩卷积”来强调卷积，作为密集预测任务中的强大工具。 Atrous 卷积允许我们明确地控制在深度卷积神经网络中计算特征响应的分辨率。它还允许我们有效地扩大滤波器的视野，以在不增加参数数量或计算量的情况下结合更大的上下文。其次，我们提出了一个不稳定的空间金字塔池（ASPP）来在多个尺度上稳健地分割对象。ASPP 使用多个采样率和有效视场的过滤器探测进入的卷积特征层，从而捕获多个尺度的对象和图像上下文。第三，我们通过结合 DCNN 和概率图形模型的方法来改进对象边界的定位。 DCNN 中通常部署的最大池和下采样的组合实现了不变性，但是对定位精度有影响。我们通过将最终 DCNN 层的响应与完全连接的条件随机场（CRF）相结合来克服这个问题，CRF 在定性和定量方面都有所提高，以提高定位性能。我们提出的“DeepLab”系统在 PASCALVOC-2012 语义图像分割任务中设置了新的先进技术，在测试集中达到了 79.7％mIOU，并将结果推进到其他三个数据集：PASCAL-Context，PASCAL-Person-部分和城市景观。我们所有的代码都在网上公开发布。

Introduction

深度卷积神经网络（DCNNs）[1]已经将计算机视觉系统的性能推向了各种高级问题的飙升，包括图像分类[2]，[3]，[4]，[5]， [6]和目标检测[7]，[8]，[9]，[10]，[11]，[12]，其中以端到端方式训练的 DCNN 比依赖于系统的系统提供了明显更好的结果 手工制作的功能。 这种成功的关键在于 DCNN 对局部图像变换的内置不变性，这使得他们可以学习越来越抽象的数据表示[13]。 这种不变性对于分类任务显然是可取的，但是可能妨碍密集预测任务，例如语义分割，其中空间信息的抽象是不期望的。

特别地，我们考虑将 DCNN 应用于语义图像分割中的三个挑战：（1）降低的特

征分辨率，（2）多尺度的对象的存在，以及（3）由于 DCNN 不变性而降低的定位精度。接下来，我们将讨论这些挑战以及我们在我们提出的 DeepLab 系统中克服它们的方法。

第一个挑战是由最初为图像分类设计的连续 DCNN 层执行的最大池和下采样（'跨步'）的重复组合引起的[2]，[4]，[5]。当 DCNN 以完全卷积的方式使用时，这导致特征图的空间分辨率显着降低[14]。为了克服这个障碍并有效地生成更密集的特征映射，我们从 DCNN 的最后几个最大池化层中移除下采样运算符，而是对后续卷积层中的滤波器进行上采样，从而得到以更高采样率计算的特征映射。过滤器上采样相当于在非零滤波器分接头之间插入孔（法语中为“trous”）。这种技术在信号处理方面有着悠久的历史，最初是为了在一种被称为“算法”的方案中对未抽样小波变换进行有效计算而开发的。一个愚蠢的“[15]。我们使用术语 atrous 卷积作为与上采样滤波器卷积的简写。在[3]，[6]，[16]中，在 DCNN 的背景下已经使用过这种想法的各种方法。在实践中，我们通过组合 atrous 卷积来恢复全分辨率特征图，该卷积更加密集地计算特征图，然后是对原始图像尺寸的特征响应的简单双线性插值。该方案提供了一种简单而强大的替代方法，可在密集预测任务中使用去卷积层[13]，[14]。与具有较大滤波器的常规卷积相比，紊乱卷积允许我们有效地扩大滤波器的视野，而不增加参数的数量或计算量。

第二个挑战是由多个尺度的物体的存在引起的。解决这个问题的一种标准方法是向 DCNN 呈现相同图像的重新缩放版本，然后聚合特征或得分图[6]，[17]，[18]。我们表明这种方法确实提高了我们系统的性能，但是以输入图像的多个缩放版本的所有 DCNN 层计算特征响应为代价。相反，在空间金字塔汇集[19]，[20]的推动下，我们提出了一种计算有效的方案，在卷积之前以多种速率重新采样给定的特征层。这相当于使用具有互补有效视野的多个过滤器探测原始图像，从而在多个尺度上捕获对象以及有用的图像上下文。我们不是实际重新采样特征，而是使用具有不同采样率的多个并行的迂回卷积层来有效地实现这种映射。我们将所提出的技术称为“巨大的空间金字塔池”（ASPP）。

第三个挑战涉及这样一个事实，即以物体为中心的分类器需要空间变换的不变性，固有地限制了 DCNN 的空间精度。缓解此问题的一种方法是在计算最终分割结果时使用跳过层从多个网络层中提取“超列”特征[14]，[21]。我们的工作探索了一

种我们表现出高度有效的替代方法。特别是，我们通过采用完全连通的条件随机场（CRF）来捕捉细节的能力[22]。 CRF 已被广泛用于语义分割，将多路分类器计算的类别分数与像素和边缘的局部相互作用所捕获的低级信息结合起来[23]，[24]或其他像素[25]。提出了提高复杂度的技术。模拟层次依赖[26]，[27]，[28]和/或段[29]，[30]，[31]，[32]，[33]的高阶依赖关系，我们使用完全连接由[22]提出的成对 CRF，它具有高效的计算能力，能够捕捉细节边缘细节，同时也能满足长距离依赖性。该模型在[22]中显示，以改善基于增强的像素级分类器的性能。在这项工作中，我们证明了当与基于 DCNN 的像素级分类器结合使用时，它可以产生最先进的结果。

图 1 显示了所提出的 DeepLab 模型的高级图示。在图像分类任务中训练的深度卷积神经网络（VGG-16 [4]或 ResNet-101 [11]）是通过以下方式实现语义分割的任务：（1）将所有完全连接的层转换为卷积层（即完全卷积网络[14]）和（2）通过不正常的卷积层增加特征分辨率，允许我们每 8 像素计算特征响应而不是原始网络中的每 32 个像素。然后，我们采用双线性插值将分数图上采样 8 倍，以达到原始图像分辨率，从而产生完全连接的 CRF [22]的输入，该 CRF 确定了分割结果。

从实用的角度来看，我们的 DeepLab 系统的三个主要优点是：（1）速度：由于紊乱的卷积，我们的密集 DCNN 在 NVidia Titan X GPU 上以 8 FPS 运行，而完全连接的 CRF 的平均场推断需要 0.5 秒在 CPU 上。 （2）准确性：我们在几个具有挑战性的数据集上获得了最新的结果，包括 PASCAL VOC 2012 语义分割基准[34]，PASCAL-Context [35]，PASCAL-PersonPart [36]和 Cityscapes [37]。 （3）简单性：我们的系统由两个非常完善的模块，DCNN 和 CRF 组成。

我们在本文中介绍的更新后的 DeepLab 系统与我们原始会议出版物[38]中报告的第一个版本相比有几项改进。我们的新版本可以通过多尺度输入处理[17]，[39]，[40]或建议的 ASPP 更好地分割多个尺度的对象。我们通过调整最新的 ResNet [11]图像分类 DCNN 构建了 DeepLab 的残差网络变体，与基于 VGG-16 的原始模型相比，实现了更好的语义分割性能[4]。最后，我们对多种模型变体进行了更全面的实验评估，并报告了最新结果，不仅是 PASCAL VOC 2012 基准，还有其他具有挑战性的任务。我们通过扩展 Caffe 框架实现了所提出的方法[41]。我们

在配套网站 http://liangchiehchen.com/projects/ DeepLab.html 上分享我们的代码和模型。

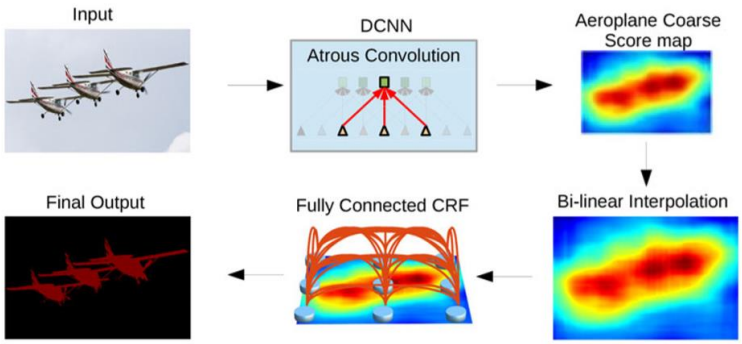


Fig. 1. Model illustration. A deep convolutional neural network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original image resolution. A fully connected CRF is then applied to refine the segmentation result and better capture the object boundaries.

2.related work

在过去十年中开发的大多数成功的语义分割系统依赖于手工制作的特征与平面分类器相结合，例如 Boosting [24]，[42]，Random Forests [43]或支持向量机[44]。通过整合来自背景[45]和结构化预测技术[22]，[26]，[27]，[46]的更丰富的信息，已经取得了实质性的进展，但这些系统的表现一直受到有限的表达能力的影响。特点。在过去几年中，深度学习在图像分类中的突破很快转移到了语义分割任务。由于此任务涉及分段和分类，因此一个核心问题是如何组合这两个任务。

用于语义分割的第一类基于 DCNN 的系统通常采用自下而上的图像分割级联，然后是基于 DCNN 的区域分类。例如，[47]，[48]提供的边界框提议和掩蔽区域在[7]和[49]中用作 DCNN 的输入，以将形状信息合并到分类过程中。同样，[50]的作者依赖于超像素表示。尽管这些方法可以从良好分割所提供的明显边界中获益，但它们也无法从其任何错误中恢复。

第二类工作依赖于使用卷积计算的 DCNN 特征进行密集图像标记，并将它们与独立获得的分割结合起来。首先，[39]以多种图像分辨率应用 DCNN，然后使用分割树来平滑预测结果。最近，[21]提出使用跳过层并将 DCNN 内的计算中间特征映射连接起来用于像素分类。此外，[51]建议按区域提案汇集中间特征地图。这些工作仍采用与 DCNN 分类器结果分离的分段算法，因此有可能承担过早的决策。

第三类作品使用 DCNN 直接提供密集的类别级像素标签，这使得甚至可以完全丢弃分段。 [14]，[52]的无分割方法以完全卷积的方式直接将 DCNN 应用于整个

图像，将 DCNN 的最后完全连接的层转换为卷积层。为了处理引言中概述的空间定位问题，[14]对中间特征图的分数进行上采样和连接，而[52]通过将粗略结果传播到另一个 DCNN 来重新确定从粗到细的预测结果。我们的工作建立在这些工作的基础上，并且如引言中所述通过对特征分辨率施加控制来扩展它们，引入多尺度池技术并将密集连接的 CRF [22]集成在 DCNN 之上。我们证明这会导致显着更好的分割结果，特别是沿着对象边界。 DCNN 和 CRF 的组合当然不是新的，但以前的工作只尝试了本地连接的 CRF 模型。具体而言，[53]使用 CRF 作为基于 DCNN 的重新排名系统的提议机制，而[39]将超像素视为本地成对 CRF 的节点，并使用图形切割进行离散推理。因此，他们的模型受到超像素计算中的错误或忽略的远程依赖性的限制。相反，我们的方法将每个像素视为由 DCNN 接收一元电位的 CRF 节点。至关重要的是，我们采用的完全连接的 CRF 模型[22]中的高斯 CRF 电位可以捕获长程依赖性，同时该模型适用于快速平均场推断。我们注意到，对于传统的图像分割任务，已经广泛研究了平均场推断[54]，[55]，[56]，但这些旧模型通常仅限于短距离连接。在独立工作中，[57]使用非常相似的密集连接的 CRF 模型来重新确定材料分类问题的 DCNN 结果。然而，[57]的 DCNN 模块仅通过稀疏点监督而不是在每个像素处进行密集监督来训练。

由于这项工作的第一版已公开发表[38]，语义分割领域已经取得了巨大进展。多个小组取得了重大进展，显着提高了 PASCAL VOC 2012 语义分段基准的标准，反映了基准排行榜 1 中的高水平活动[17]，[40]，[58]，[59]，[60]]，[61]，[62]，[63]。有趣的是，大多数表现良好的方法都采用了我们 DeepLab 系统的一个或两个关键要素：通过完全连接的 CRF 进行有效的密集特征提取和原始 DCNN 分数的改进的 Atrous 卷积。我们在下面概述了一些最重要和最有趣的进展。

最近在一些相关的工作中探索了结构化预测的端到端培训。虽然我们使用 CRF 作为后处理方法，[40]，[59]，[62]，[64]，[65]成功地进行了 DCNN 和 CRF 的联合学习。特别是，[59]，[65]展开 CRF 平均场推断步骤，将整个系统转换为端到端可训练的前馈网络，而[62]近似密集 CRF 平均场推断的一次迭代[22]卷积层与可学习的过滤器。 [40]，[66]追求的另一个富有成效的方向是通过 DCNN 学习 CRF 的成对术语，以更重的计算为代价显着提高性能。在不同的方向，[63]用更快的域变换模块[67]取代平均场推理中使用的双边滤波模块，提高速度并降低整

个系统的内存需求，而[18]，[68]结合起来边缘检测的语义分割。
在许多论文中都进行了较弱的监督，放宽了对整个训练集可用的像素级语义注释的假设[58]，[69]，[70]，[71]，得到的结果明显好于弱 - 监督前 DCNN 系统，如[72]。在另一个研究领域，[49]，[73]追求实例分割，共同解决对象检测和语义分割。

我们在这里所谓的无味卷积最初是为了在[15]的“算法？trous”方案中对未抽取小波变换的有效计算而开发的。我们将感兴趣的读者引用到[74]以获得小波文献的早期参考。 Atrous 卷积也与多速率信号处理中的“高贵身份”密切相关，它建立在输入信号和滤波器采样率的相同相互作用的基础上[75]。 Atrous 卷积是我们在[6]中首次使用的术语。同样的操作后来被称为扩张卷积[76]，他们创造的一个术语是由于操作对应于常规卷积与上采样（或在[15]的术语中）扩散的事实。为了在DCNN中进行更密集的特征提取，各种作者使用了相同的操作[3]，[6]，[16]。除了单纯的分辨率增强之外，激烈的卷积使我们能够扩大滤波器的视野范围，以结合更大的背景，我们在[38]中已经证明了它是有益的。 [76]进一步推行了这种方法，他们利用一系列不稳定的卷积层，以更高的速率聚合多尺度环境。这里提出的用于捕获多尺度对象和上下文的浮动空间金字塔池方案也采用具有不同采样率的多个迂回卷积层，然而我们并行地而不是串行地布置。有趣的是，动态卷积技术也被用于更广泛的任务，例如物体检测[12]，[77]，实例级分割[78]，视觉问答[79]和光学流[80]。 。
我们还表明，正如预期的那样，集成到 DeepLab 中的更先进的图像分类 DCNN，例如[11]的残差网络可以产生更好的结果。 [81]也独立观察到了这一点。

3 METHODS

3.1 Atrous Convolution for Dense Feature Extraction and Field-of-View Enlargement

通过以完全卷积的方式部署 DCNN，已经证明 DCNN 用于语义分段或其他密集预测任务是简单而成功地解决的[3]，[14]。然而，在这些网络的连续层处的最大池和跨步的重复组合显着地降低了所得特征图的空间分辨率，通常在最近的 DCNN 中的每个方向上的因子为 32 倍。部分补救措施是使用[14]中的“反卷积”

层，但这需要额外的内存和时间。

我们提倡使用最初为[15]的“算法？trous”方案中的未抽取小波变换的有效计算而开发的 atrous 卷积，并且在 DCNN 背景中使用[3]，[6]，[16。该算法允许我们以任何所需的分辨率计算任何层的响应。一旦网络经过培训，它就可以在事后应用，但也可以与培训无缝集成。

首先考虑一维信号，将具有长度为 K 的滤波器 w [k]的 1-D 输入信号 x [i]的迂回卷积 2 的输出 y [i]定义为

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k]w[k].$$

速率参数 r 对应于我们对输入信号进行采样的步幅。标准卷积是速率 r = 1 的特殊情况。参见图 2 以进行说明。

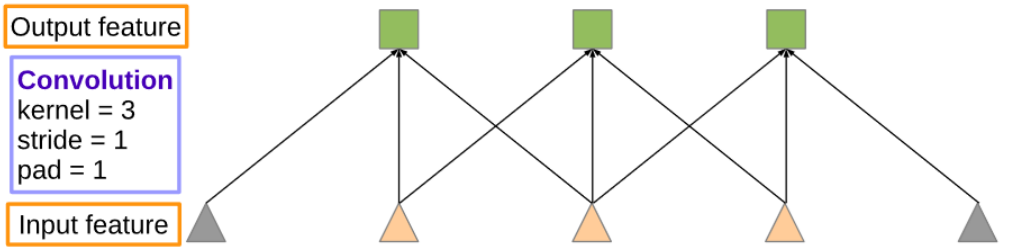
我们通过图 3 中的一个简单示例来说明算法在 2-D 中的操作：给定图像，我们假设我们首先进行下采样操作，将分辨率降低 2 倍，然后使用内核执行卷积 - 这里，垂直高斯导数。如果在原始图像坐标中植入生成的特征图，我们意识到我们仅在图像位置的 1/4 处获得了响应。相反，我们可以计算所有图像位置的响应，如果我们将全分辨率图像与过滤器'带孔'进行卷积，其中我们将原始滤波器上采样 2 倍，并在滤波器值之间引入零。虽然有效滤波器尺寸增加，但我们只需要考虑非零滤波器值，因此滤波器参数的数量和每个位置的操作数量保持不变。由此产生的方案使我们能够轻松，明确地控制神经网络特征响应的空间分辨率。

在 DCNN 的上下文中，可以在层链中使用迂回卷积，有效地允许我们以任意高的分辨率计算最终的 DCNN 网络响应。例如，为了使 VGG-16 或 ResNet-101 网络中计算出的特征响应的空间密度加倍，我们找到了降低分辨率的最后一个汇集或卷积层（分别为“pool5”或“conv5_1”），设置了它的步幅为了避免信号抽取，并用速率为 r = 2 的迂回卷积层替换所有后续卷积层。在整个网络中一直推动这种方法可以让我们以原始图像分辨率计算特征响应，但这最终成为太贵了。我们采

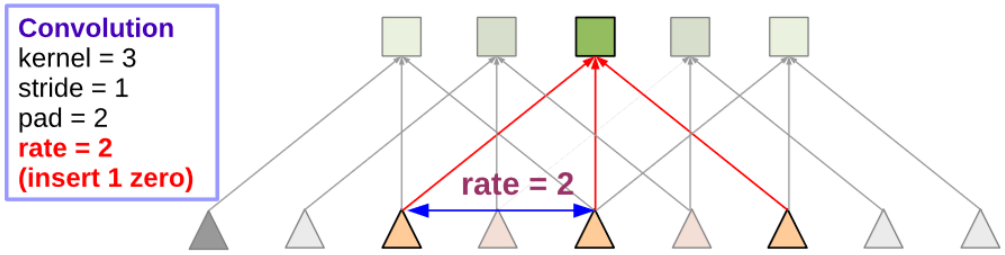
用了混合方法来实现良好的效率/准确性权衡，使用迂回卷积将计算特征图的密度增加 4 倍，然后通过额外因子 8 快速双线性插值来恢复特征图在原始图像分辨率。双线性插值在此设置中是足够的，因为类别得分图（对应于对数概率）非常平滑，如图 5 所示。与[14]采用的去卷积方法不同，所提出的方法将图像分类网络转换为密集特征提取器无需学习任何额外参数，从而在实践中实现更快的 DCNN 培训。

Atrous 卷积还允许我们在任何 DCNN 层任意放大滤波器的视野。现有技术的 DCNN 通常采用空间上小的卷积核（通常为 3×3），以便保持计算和包含的参数数量。利率 r 的剧烈卷积在连续滤波器值之间引入 r-1 个零，有效地将 ak×k 滤波器的内核大小扩大到 ke＝k＋（k-1）×（r-1），而不增加参数的数量或者计算。因此，它提供了一种有效的机制来控制视野，并在精确定位（小视场）和上下文同化（大视场）之间找到最佳平衡点。我们已经成功地尝试了这种技术：我们的 DeepLab-LargeFOV 模型变体[38]采用了萎缩卷积，在 VGG-16'fc6'层中速率 r = 12，具有显着的性能增益，详见第 4 节。

转向实现方面，有两种有效的方法来执行痛苦的卷积。首先是通过插入孔（零）来隐式地对过滤器进行上采样，或者等效地稀疏地对输入特征图进行采样[15]。我们在早期的工作[6]，[38]中实现了这一点，然后在[76]中，在 Caffe 框架[41]中通过添加到 im2col 函数（它从多通道特征映射中提取矢量化补丁）稀疏地实现了这个选项。对基础要素图进行采样。最初由[82]提出并在[3]，[16]中使用的第二种方法是通过等于迂回卷积率 r 的因子对输入特征图进行子采样，对其进行解交织以产生 r2 降低分辨率的映射，每个映射一个 r×r 可能的变化。然后将标准卷积应用于这些中间特征图并将它们重新交换为原始图像分辨率。通过将 atrous 卷积减少为常规卷积，它允许我们使用现成的高度优化的卷积例程。我们已经在 TensorFlow 框架中实现了第二种方法[83]。



(a) Sparse feature extraction



(b) Dense feature extraction

Fig. 2. Illustration of atrous convolution in 1-D. (a) Sparse feature extraction with standard convolution on a low resolution input feature map. (b) Dense feature extraction with atrous convolution with rate $r = 2$, applied on a high resolution input feature map.

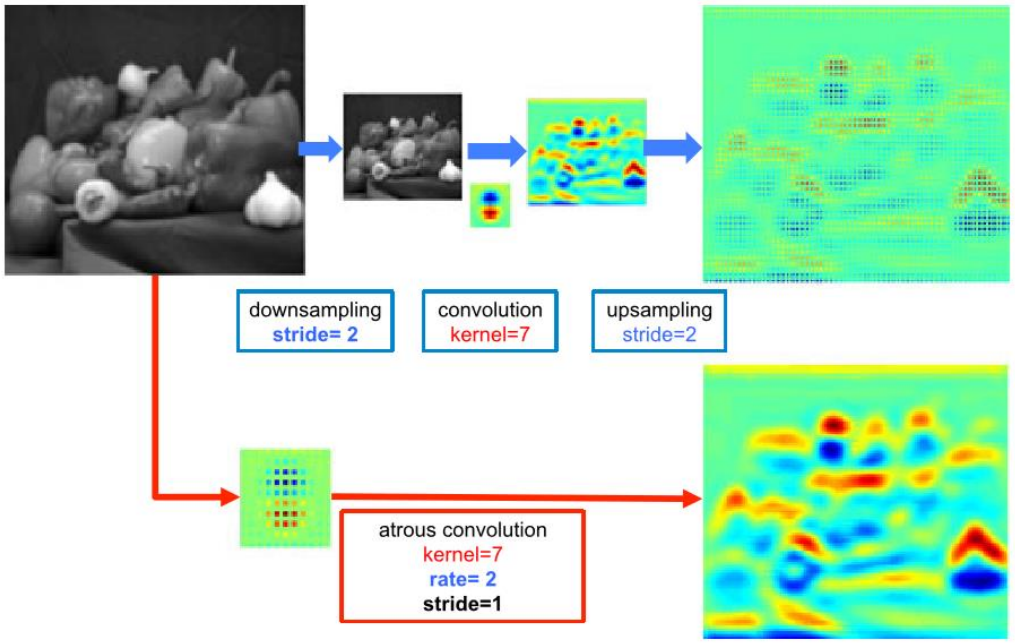


Fig. 3. Illustration of atrous convolution in 2-D. Top row: sparse feature extraction with standard convolution on a low resolution input feature map. Bottom row: Dense feature extraction with atrous convolution with rate $r = 2$, applied on a high resolution input feature map.

3.2 Multiscale Image Representations Using Atrous Spatial Pyramid Pooling

DCNN 已显示出显着的隐式表示比例的能力，只需在包含不同大小的对象的数据集上进行训练即可。尽管如此，明确考虑对象规模可以提高 DCNN 成功处理大型和小型对象的能力[6]。

我们已经尝试了两种处理语义分割中的尺度可变性的方法。第一种方法相当于标准的多尺度处理[17]，[18]。我们使用共享相同参数的并行 DCNN 分支从多个（在我们的实验中）三个重新缩放版本的原始图像中提取 DCNN 得分图。为了产生最终结果，我们将并行 DCNN 分支的特征映射双线性插值到原始图像分辨率并融合它们，通过在每个位置获取不同尺度上的最大响应。我们在培训和测试期间都这样做。多尺度处理显着提高了性能，但代价是计算所有 DCNN 层的特征响应，以满足多种输入规模。

第二种方法的灵感来自于[20]的 R-CNN 空间金字塔合并方法的成功，该方法表明，通过重新采样在单一尺度上提取的卷积特征，可以准确且有效地对任意尺度的区域进行分类。我们已经实现了他们的方案的变体，其使用具有不同采样率的多个并行的 atrous 卷积层。为每个采样率提取的特征在不同的分支中进一步处理并融合以产生最终结果。提出的“atrous 空间金字塔池”（DeepLabASPP）方法概括了我们的 DeepLab-LargeFOV 变体，如图 4 所示。

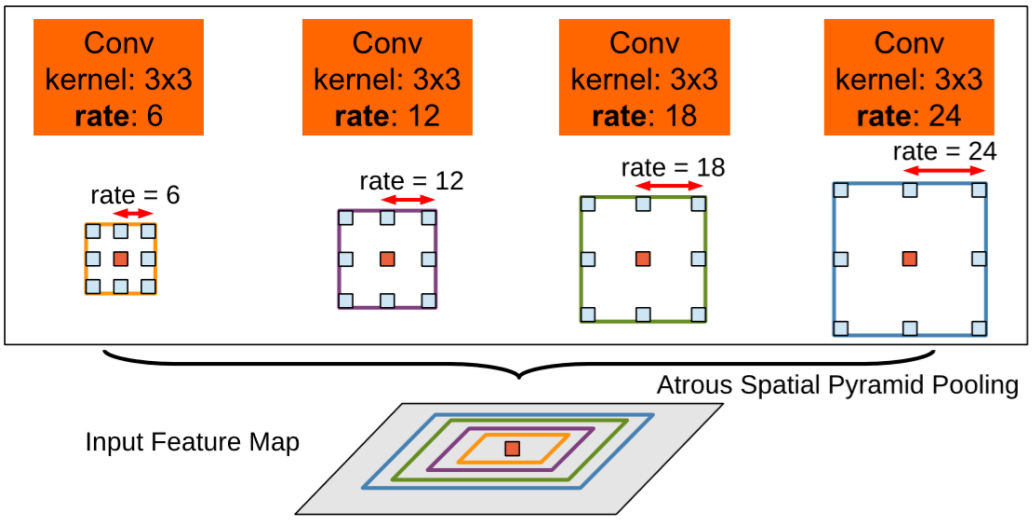


Fig. 4. Atrous Spatial Pyramid Pooling (ASPP). To classify the center pixel (orange), ASPP exploits multi-scale features by employing multiple parallel filters with different rates. The effective Field-Of-Views are shown in different colors.

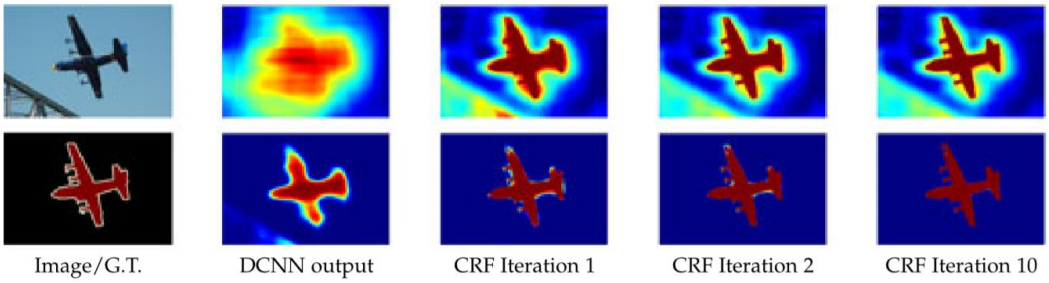


Fig. 5. Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. We show the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of last DCNN layer is used as input to the mean field inference.

3.3 Structured Prediction with Fully-Connected Conditional Random Fields for Accurate Boundary Recovery

在定位精度和分类性能之间的权衡似乎是 DCNN 中固有的：具有多个最大池层的更深层模型已经证明在分类任务中最成功，但是增加的不变性和顶级节点的大型接收场只能产生平滑的响应。如图 5 所示，DCNN 得分图可以预测对象的存在和粗略位置，但不能真正描绘其边界。

以前的工作追求两个方向来解决这一本地化挑战。第一种方法是利用卷积网络中多层的信息，以便更好地估计物体边界[14]，[21]，[52]。第二种是采用超像素表示，基本上将本地化任务委托给低级分割方法[50]。

我们在 DCNN 的识别能力与完全连接的 CRF 的细粒度定位精度的耦合基础上寻求替代方向，并表明它在解决定位挑战方面非常成功，产生了准确的语义分割结果并在一定程度上恢复了对象边界。细节远远超出了现有方法的范围。几个后续文件[17]，[40]，[58]，[59]，[60]，[61]，[62]，[63]，[65]已经扩展了这个方向，因为第一个我们的工作版本已发表[38]。

传统上，条件随机场（CRF）已被用于平滑噪声分割图[23]，[31]。通常，这些模型耦合相邻节点，有利于相同标签分配到空间上近端像素。定性地说，这些短程 CRF 的主要功能是清理基于本地手工设计功能构建的弱分类器的虚假预测。

与这些较弱的分类器相比，现代 DCNN 架构（例如我们在此工作中使用的架构）产生的得分图和语义标签预测在质量上是不同的。如图 5 所示，得分图通常非常平滑并产生均匀的分类结果。在这种情况下，使用短程 CRF 可能是有害的，因

为我们的目标应该是恢复详细的局部结构而不是进一步平滑它。将对比敏感电位[23]与局部范围 CRF 结合使用可以潜在地改善定位，但仍然缺少薄结构，并且通常需要解决昂贵的离散优化问题。

为了克服短距离 CRF 的这些限制，我们将[22]的完全连接的 CRF 模型集成到我们的系统中。该模型采用能量函数

$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j),$$

其中 x 是像素的标签分配。 我们使用一元势 o（xi）= - logP（xi），其中 Pðxi 是由 DCNN 计算的像素 i 处的标签分配概率。 成对电位具有允许在使用完全连接的图形时进行有效推断的形式，即，当连接所有图像像素对时，i; j。 特别是，如[22]中所述，我们使用以下表达式

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j)\left[w_1 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}\right) + w_2 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}\right)\right],$$

其中 u（xi; xj）= 1，如果 xi！= xj，否则为零，如 Potts 模型中所示，这意味着只有具有不同标签的节点才会受到惩罚。剩下的表达式在不同的特征空间中使用两个高斯核;第一个“双边”内核取决于像素位置（表示为 p）和 RGB 颜色（表示为 I），第二个内核仅取决于像素位置。超参数 sa，sb 和 sg 控制高斯核的尺度。第一个内核强制具有相似颜色和位置的像素具有相似的标签，而第二个内核仅在强制平滑时考虑空间接近度。

至关重要的是，该模型适用于有效的近似概率推理[22]。在完全可分解的平均场近似 bx = 累乘 bi（xi）下传递更新的消息可以表示为双边空间中的高斯卷积。高维滤波算法[84]显着加速了这种计算，导致算法在实践中非常快，使用[22]的公开实现，Pascal VOC 图像平均需要少于 0.5 秒。

4 EXPERIMENTAL RESULTS

我们按照[14]的程序，通过 Imagenet 预训练的 VGG-16 或 ResNet-101 网络的模型权重，以简单的方式使它们适应语义分割任务。我们用一个分类器替换最后一层中的 1000 路 Imagenet 分类器，该分类器具有与我们任务的语义类数量（包括背景，如果适用）一样多的目标。我们的损失函数是 CNN 输出图中每个空间位置的交叉熵项的总和（与原始图像相比，由 8 个子采样）。所有位置和标签在整体损失函数中均等加权（忽略的未标记像素除外）。我们的目标是基本事实标签（由 8 个子样本）。我们通过[2]的标准 SGD 程序针对所有网络层的权重优化目标函数。我们解耦 DCNN 和 CRF 训练阶段，假设在设置 CRF 参数时固定 DCNN 一元项。

我们在四个具有挑战性的数据集上评估所提出的模型：PASCAL VOC 2012，PASCAL-Context，PASCALPerson-Part 和 Cityscapes。我们首先报告了关于 PASCAL VOC 2012 的会议版本[38]的主要结果，并向前推进了所有数据集的最新结果。

4.1 PASCAL VOC 2012

数据集。PASCAL VOC 2012 分割基准[34]涉及 20 个前景对象类和一个背景类。原始数据集包含分别用于训练，验证和测试的 1,464（训练），1,449（val）和 1,456（测试）像素级标记图像。 数据集由[85]提供的额外注释增强，产生 10,582（trainaug）训练图像。 性能是根据 21 个类别的平均像素交叉（IOU）来衡量的。

4.1.1 Results from Our Conference Version

我们使用在 Imagenet 上预训练的 VGG-16 网络，适用于 3.1 节中描述的语义分割。我们使用 20 个图像的小批量和 0.001 的初始学习率（最终分类层为 0.01），每 2000 次迭代将学习率乘以 0.1。我们使用 0.9 的动量和 0.0005 的重量衰减。

在 DCNN 已经在 trainaug 上进行了调整后，我们沿着[22]的线交叉验证 CRF 参数。我们使用 w2 = 3 和 sg = 3 的默认值，我们通过对来自 val 的 100 个图像的交叉验证来搜索 w1，sa 和 sb 的最佳值。我们采用粗略搜索方案。参数的初始搜索范围是 w1 [3：6] ?, sa [30：10：100]和 sb [3：6]（MATLAB 表示法），然后我们

围绕第一轮的最佳值重新确定搜索步长。我们采用 10 次平均场迭代。

视野和 CRF。在表 1 中，我们报告使用 DeepLab 模型变体的实验，这些变体使用不同的视野尺寸，通过调整'fc6'层中的内核大小和不稳定采样率 r 获得，如 3.1 节所述。我们首先直接调整 VGG16 网络，使用原始的 7×7 内核大小和 r＝4（因为我们对最后两个最大池化层没有任何步幅）。该模型在 CRF 后产生 67.64％的性能，但相对较慢（在训练期间每秒 1.44 幅图像）。通过将内核大小减小到 4×4，我们将模型速度提高到每秒 2.9 个图像。我们已经尝试了两种具有较小（r = 4）和较大（r = 8）FOV 大小的网络变体;后者表现更好。最后，我们采用内核大小 3×3 甚至更大的采样率（r＝12），通过在层'fc6'和'fc7'中保留 4,096 个滤波器中的 1,024 个随机子集，使网络更薄。得到的模型 DeepLab-CRF-LargeFOV 与直接 VGG-16 自适应（7×7 内核大小，r = 4）的性能相匹配。与此同时，DeepLab-LargeFOV 的速度提高了 3.36 倍，参数显着减少（20.5 M 而不是 134.3 M）。CRF 大大提升了所有型号的性能，平均 IOU 绝对增加了 3-5％。

TABLE 1
Effect of Field-Of-View by Adjusting the Kernel Size and Atrous sampling Rate r at 'fc6' Layer

Kernel	Rate	FOV	Params	Speed	bef/aft CRF
7×7	4	224	134.3M	1.44	64.38 / 67.64
4×4	4	128	65.1M	2.90	59.80 / 63.74
4×4	8	224	65.1M	2.90	63.41 / 67.14
3×3	12	224	20.5M	4.84	62.25 / 67.64

We show number of model parameters, training speed (img/sec), and val *set mean IOU before and after CRF. DeepLab-LargeFOV (kernel size* 3×3, $r = 12$*) strikes the best balance.*

4.1.2 Improvements after Conference Version of This Work

在这项工作的会议版本之后[38]，我们对模型进行了三次主要改进，我们将在下面讨论：（1）培训期间的不同学习政策，（2）不稳定的空间金字塔汇集，以及（3）更深层次的就业网络和多尺度处理。

学习率政策。我们在培训 DeepLab-LargeFOV 时探索了不同的学习率政策。类似

于[86]，我们还发现采用“多边形”学习率政策（学习率乘以 ð1？ iter maxiterÞpower）比“步进”学习率更有效（降低固定步长的学习率）)。如表 2 所示，采用“poly”（功率为 0：9）并使用相同的批量大小和相同的训练迭代，比使用“步骤”策略产生 1.17％的性能。修复批量大小并将训练迭代次数增加到 10 K 可将性能提高到 64.90％（增益为 1.48％）;但是，由于更多的训练迭代，总训练时间增加了。然后我们将批量大小减少到 10，发现仍然保持了相当的性能（64.90％对比 64.71％）。最后，我们采用批量大小= 10 和 20 K 迭代，以保持与先前“步骤”策略类似的训练时间。令人惊讶的是，这使得我们在 val 上的性能提高了 65.88％（相对于“步骤”提高了 3.63％），在测试中提高了 67.7％，而在 CRF 之前，DeepLabLargeFOV 的原始“步骤”设置为 65.1％。我们对本文其余部分报告的所有实验采用“多边”学习率政策。

Atrous 空间金字塔池。我们已经尝试了第 3.1 节中描述的提议的 Atrous 空间金字塔池（ASPP）方案。如图 7 所示，VGG-16 的 ASPP 采用几个并行的 fc6-fc7-fc8 分支。他们都在’fc6’中使用 3？ 3 个内核但不同的 atrous rate r 来捕获不同大小的对象。在表 3 中，我们报告了几个设置的结果：（1）我们的基线 LargeFOV 模型，具有 r = 12 的单个分支，（2）ASPP-S，具有四个分支和较小的动力率（r = {2,4， 8,12}，和（3）ASPP-L，具有四个分支和更大的速率（r = {6,12,18,24}）。Foreach 变体我们报告 CRF 之前和之后的结果。如表中所示，ASPP-S 比 CRF 之前的基线 LargeFOV 提高了 1.22％。然而，在 CRF 之后，LargeFOV 和 ASPP-S 的表现相似。另一方面，ASPP-L 在 CRF 之前和之后产生了对基线 LargeFOV 的一致改进。我们在测试中评估了拟议的 ASPP-L + CRF 模型，达到了 72.6％。我们想象了图 8 中不同方案的效果。

更深入的网络和多尺度处理。我们已经尝试围绕最近提出的剩余网络 ResNet-101 [11]而不是 VGG-16 构建 DeepLab。与我们对 VGG-16 网络所做的类似，我们通过激烈的卷积重新使用 ResNet-101，如第 3.1 节所述。最重要的是，我们采用了[17]，[18]，[39]，[40]，[58]，[59]，[62]的最新研究成果：（1）多尺度输入：我们分别以比例= {0.5,0.75,1}对 DCNN 图像进行馈送，通过分别对每个位置的比例进行最大响应来融合他们的得分图[17]。 （2）在 MS-COCO 上预训练的模型[87]。 （3）通过在训练期间随机缩放输入图像（从 0.5 到 1.5）来增加

数据。在表 4 中，我们评估了这些因素以及 LargeFOV 和 atrous 空间金字塔池（ASPP）如何影响 val 集性能。采用 ResNet-101 而不是 VGG-16 显着提高了 DeepLab 的性能（例如，我们最简单的基于 ResNet-101 的型号达到 68.72％，而基于 DeepLab-LargeFOV VGG-16 的型号为 65.76％，均为 CRF 之前）。多尺度融合[17]带来了额外的 2.55％的改进，而在 MS-COCO 上预训练模型又获得了 2.01％的增益。培训期间的数据增加是有效的（大约 1.6％的改进）。使用 LargeFOV（在 ResNet 之上添加一个充满紊乱的卷积层，具有 3×3 内核并且速率= 12）是有益的（大约 0.6％的改进）。通过剧烈的空间金字塔池（ASPP）实现了 0.8％的改进。通过密集的 CRF 对我们的最佳模型进行后处理可以获得 77.69％的性能。

定性结果。我们提供了图 6 中 CRF 之前和之后 DeepLab 结果（我们的最佳模型变体）的定性视觉比较。在 CRF 之前由 DeepLab 获得的可视化结果已经产生了出色的分割结果，而使用 CRF 通过消除误报和进一步提高了性能重新定义对象边界。

测试集结果。我们已经将最终最佳模型的结果提交给了官方服务器，获得了 79.7％的测试集性能，如表 5 所示。该模型大大优于以前的 DeepLab 变体（例如，带有 VGG-16 网络的 DeepLab-LargeFOV），并且目前是 PASCAL VOC 2012 细分排行榜的最佳表现方法。

VGG-16 与 ResNet-101。我们观察到基于 ResNet-101 [11]的 DeepLab 在物体边界上提供了比使用 VGG-16 更好的分割结果[4]，如图 9 所示。我们认为 ResNet-101 的身份映射[94]具有相似性影响作为超列特征[21]，它利用中间层的特征来更好地定位边界。我们在“trimap”[22]，[31]（沿物体边界的窄带）内进一步量化了图 10 中的这种效应。如图所示，在 CRF 之前使用 ResNet-101 在物体边界上具有与将 VGG-16 与 CRF 结合使用时几乎相同的精度。使用 CRF 对 ResNet-101 结果进行后处理可进一步改善分割结果。

TABLE 2
PASCAL VOC 2012 *val* Set Results (%) (before CRF)
as Different Learning Hyper Parameters Vary

Learning policy	Batch size	Iteration	mean IOU
step	30	6K	62.25
poly	30	6K	63.42
poly	30	10K	64.90
poly	10	10K	64.71
poly	10	20K	65.88

Employing "poly" learning policy is more effective than "step" when training DeepLab-LargeFOV.

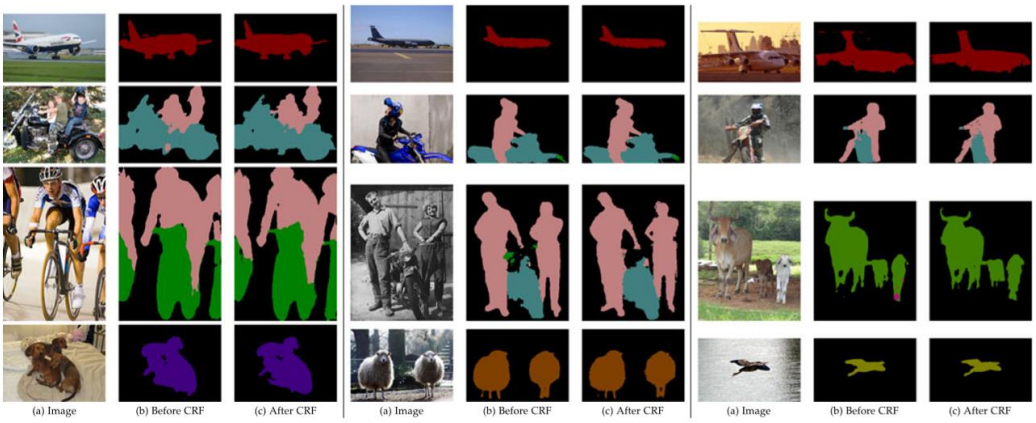
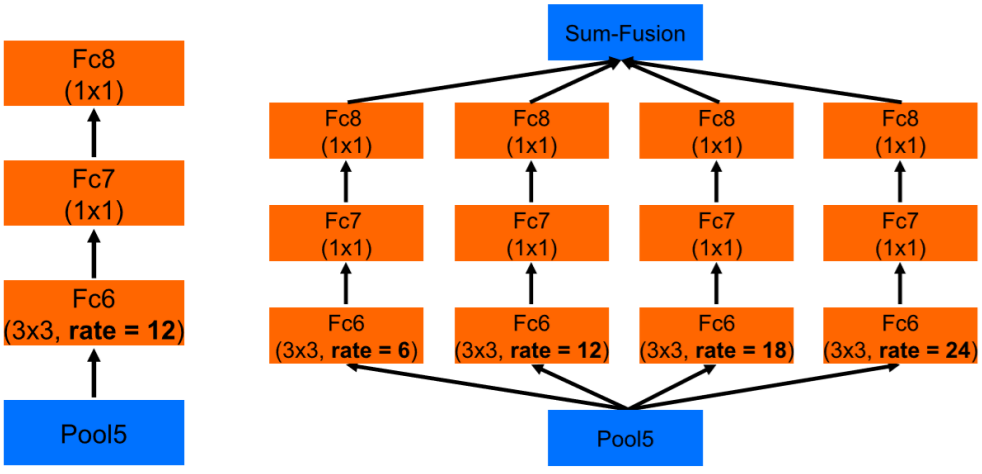
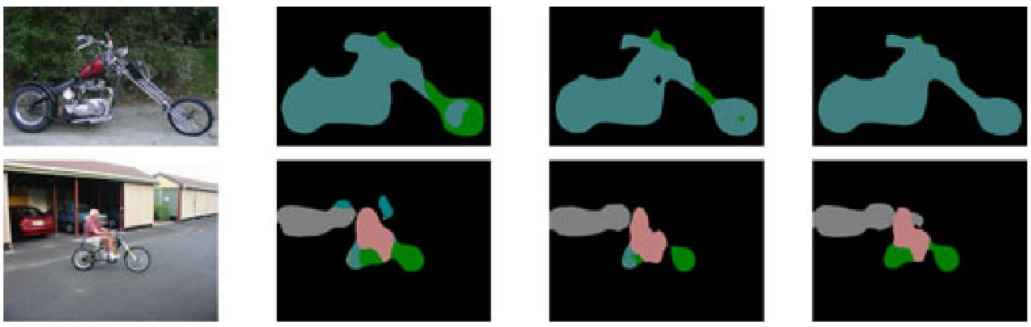


Fig. 6. PASCAL VOC 2012 *val* results. Input image and our DeepLab results before/after CRF.



(a) DeepLab-LargeFOV (b) DeepLab-ASPP

Fig. 7. DeepLab-ASPP employs multiple filters with different rates to capture objects and context at multiple scales.



(a) Image (b) LargeFOV (c) ASPP-S (d) ASPP-L

Fig. 8. Qualitative segmentation results with ASPP compared to the baseline LargeFOV model. The **ASPP-L** model, employing multiple *large* FOVs can successfully capture objects as well as image context at multiple scales.

TABLE 3
Effect of ASPP on PASCAL VOC 2012 *val* Set Performance (Mean IOU) for VGG-16 Based DeepLab Model

Method	before CRF	after CRF
LargeFOV	65.76	69.84
ASPP-S	66.98	69.73
ASPP-L	68.96	71.57

LargeFOV: *single branch,* $r = 12$. ***ASPP-S***: *four branches,* $r = \{2, 4, 8, 12\}$. ***ASPP-L***: *four branches,* $r = \{6, 12, 18, 24\}$.

TABLE 4
Employing ResNet-101 for DeepLab on PASCAL VOC 2012 *val* set

MSC	COCO	Aug	LargeFOV	ASPP	CRF	mIOU
						68.72
✓						71.27
✓	✓					73.28
✓	✓	✓				74.87
✓	✓	✓	✓			75.54
✓	✓	✓		✓		76.35
✓	✓	✓		✓	✓	77.69

MSC: *Employing mutli-scale inputs with max fusion* ***COCO***: *Models pre-trained on MS-COCO.* ***Aug***: *Data augmentation by randomly rescaling inputs.*

TABLE 5
Performance on PASCAL VOC 2012 *test* Set

Method	mIOU
DeepLab-CRF-LargeFOV-COCO [58]	72.7
MERL_DEEP_GCRF [88]	73.2
CRF-RNN [59]	74.7
POSTECH_DeconvNet_CRF_VOC [61]	74.8
BoxSup [60]	75.2
Context + CRF-RNN [76]	75.3
QO_4^{mres} [66]	75.5
DeepLab-CRF-Attention [17]	75.7
CentraleSuperBoundaries++ [18]	76.0
DeepLab-CRF-Attention-DT [63]	76.3
H-ReNet + DenseCRF [89]	76.8
LRR_4x_COCO [90]	76.8
DPN [62]	77.5
Adelaide_Context [40]	77.8
Oxford_TVG_HO_CRF [91]	77.9
Context CRF + Guidance CRF [92]	78.1
Adelaide_VeryDeep_FCN_VOC [93]	79.1
DeepLab-CRF (ResNet-101)	79.7

We have added some results from recent arXiv papers on top of the official leadearboard results.

TABLE 6
Comparison with Other State-of-Art Methods
on PASCAL-Context Dataset

Method	MSC	COCO	Aug	LargeFOV	ASPP	CRF	mIOU
VGG-16							
DeepLab [38]				✓			37.6
DeepLab [38]				✓		✓	39.6
ResNet-101							
DeepLab							39.6
DeepLab	✓		✓				41.4
DeepLab	✓	✓	✓				42.9
DeepLab	✓	✓	✓	✓			43.5
DeepLab	✓	✓	✓		✓		44.7
DeepLab	✓	✓	✓		✓	✓	45.7
O_2P [45]							18.1
CFM [51]							34.4
FCN-8s [14]							37.8
CRF-RNN [59]							39.3
ParseNet [86]							40.4
BoxSup [60]							40.5
HO_CRF [91]							41.3
Context [40]							43.3
VeryDeep [93]							44.5

Fig. 9. DeepLab results based on VGG-16 net or ResNet-101 before and after CRF. The CRF is critical for accurate prediction along object boundaries with VGG-16, whereas ResNet-101 has acceptable performance even before CRF.

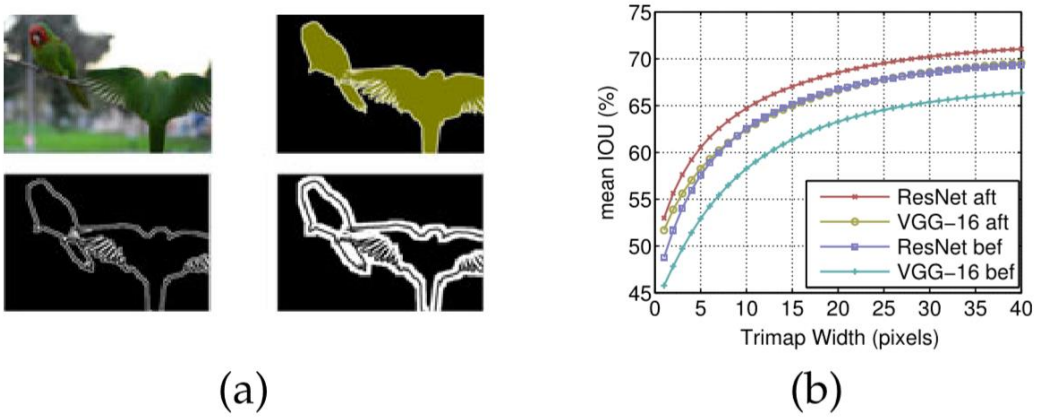


Fig. 10. (a) Trimap examples (top-left: image. top-right: ground-truth. bottom-left: trimap of 2 pixels. bottom-right: trimap of 10 pixels). (b) Pixel mean IOU as a function of the band width around the object boundaries when employing VGG-16 or ResNet-101 before and after CRF.

4.2 PASCAL-Context

数据集。 PASCAL-Context 数据集[35]为整个场景提供详细的语义标签，包括对象（例如人）和东西（例如天空）。在[35]之后，所提出的模型在最频繁的 59 个类别和一个背景类别上进行评估。训练集和验证集包含 4,998 和 5,105 个图像。评价。我们在表 6 中报告评估结果。我们的基于 VGG-16 的 LargeFOV 变体在 CRF 之前产生 37.6%和 39.6%。为 DeepLab 重新利用 ResNet101 [11]比 VGG-16 LargeFOV 提高了 2%。与[17]类似，采用多尺度输入和最大合并来合并结果可将性能提高到 41.4%。在 MS-COCO 上预加工模型可带来额外 1.5%的改进。使用不稳定的空间金字塔池比 LargeFOV 更有效。在进一步采用密集 CRF 作为后处理之后，我们的最终模型产生了 45.7%，在不使用非线性成对项的情况下，优于当前最先进的方法[40] 2.4%。我们的最终模型略微优于并发工作[93] 1.2%，它还采用了迂回卷积来重新利用[11]的剩余网络进行语义分割。

定性结果。我们将有和没有 CRF 的最佳模型的分割结果可视化为图 11 中的后处理.CRF 之前的 DeepLab 已经可以高精度地预测大多数对象/东西。使用 CRF，我们的模型能够进一步消除孤立的误报，并改善沿对象/填充边界的预测。

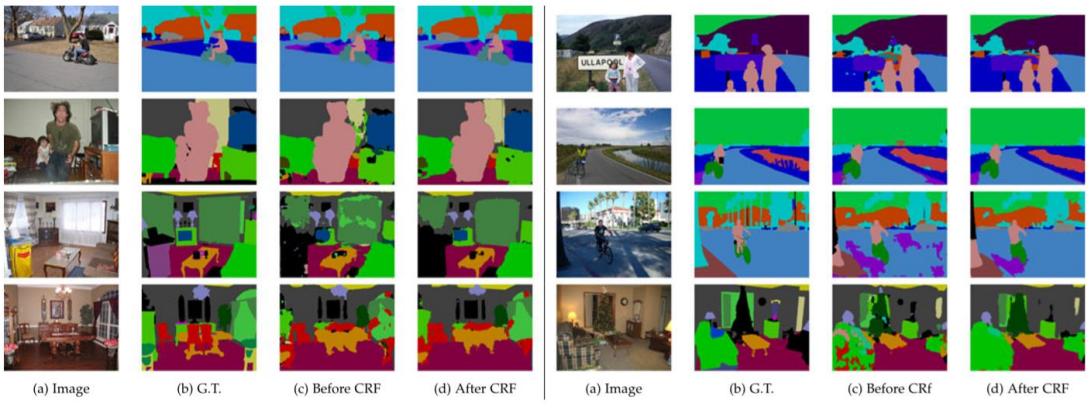


Fig. 11. PASCAL-Context results. Input image, ground-truth, and our DeepLab results before/after CRF.

4.3 PASCAL-Person-Part

数据集。我们进一步使用额外的 PASCAL VOC 2010 注释[36]进行语义部分分割[98]，[99]的实验。我们关注数据集的人物部分，其中包含更多的训练数据以及对象尺度和人体姿势的大变化。具体而言，数据集包含每个人的详细部分注释，例如眼睛，鼻子。我们将注释合并为头部，躯干，上/下臂和上/下腿，从而产生六个人的部分类和一个背景类。我们仅使用包含人员的图像进行训练（1,716 张图像）和验证（1,817 张图像）。

评价。上 PASCAL 人称部分人体部分的分割结果被报告在表 7 [17]已经在此数据集重新旨意 VGG-16 网为 DeepLab 进行的实验中，获得 56.39％的（与多尺度输入）。因此，在本部分中，我们主要关注将 ResNet-101 重新用于 DeepLab 的效果。与 RESNET-101，单独 DeepLab 产生 58.9％的显着地由分别为约 7％和 2.5％，表现优于 DeepLab-LargeFOV（VGG-16 网）和 DeepLab-注意（VGG-16 网）。通过最大池化结合多尺度输入和融合进一步将性能提高到 63.1％。此外，在 MS-COCO 上预先训练模型还可以提高 1.3％。但是，在此数据集上采用 LargeFOV 或 ASPP 时，我们没有观察到任何改进。采用密集的 CRF 进行后处理我们的最终输出大大优于同时工作[97] 4.78％。

定性结果。我们将结果可视化为图 12。

TABLE 7
Comparison with Other State-of-Art Methods
on PASCAL-Person-Part Dataset

Method	MSC	COCO	Aug	LFOV	ASPP	CRF	mIOU
ResNet-101							
DeepLab							58.90
DeepLab	✓		✓				63.10
DeepLab	✓	✓	✓				64.40
DeepLab	✓	✓	✓			✓	64.94
DeepLab	✓	✓	✓	✓			62.18
DeepLab	✓	✓	✓		✓		62.76
Attention [17]							56.39
HAZN [95]							57.54
LG-LSTM [96]							57.97
Graph LSTM [97]							60.16

Fig. 12. PASCAL-Person-Part results. Input image, ground-truth, and our DeepLab results before/after CRF.

4.4 Cityscapes

数据集。 Cityscapes [37]是最近发布的大型数据集，其中包含来自 50 个不同城市的街景中收集的 5,000 张图像的高质量像素级注释。根据评估协议[37]，19 个语义标签（属于 7 个超级类别：地面，建筑，对象，自然，天空，人类和车辆）用于评估（空白标签不被考虑用于评估）。训练，验证和测试集分别包含 2,975,500 和 1,525 个图像。

预发布的测试集结果。我们参与了 Cityscapes 数据集预发布的基准测试。如表 8 顶部所示，我们的模型获得第三名，表现为 63.1％和 64.8％（附加粗略注释图像的训练）。

Val Set 结果。在初始发布之后，我们进一步探索了表 9 中的验证集。城市景观的图像具有分辨率 2; 048？1; 024，使得训练具有有限 GPU 内存的更深网络成为具有挑战性的问题。在对数据集的预发布进行基准测试期间，我们将图像下采样为 2.但是，我们发现以原始分辨率处理图像是有益的。使用相同的训练协议，使用原始分辨率的图像分别在 CRF 之前和之后显着地带来 1.9 和 1.8％的改进。为了使用高分辨率图像对该数据集进行推断，我们将每个图像分割成重叠区域，类似于[37]。我们还用 ResNet-101 取代了 VGG-16 网。由于手头的 GPU 内存有限，我们不会利用多尺度输入。相反，我们只探索（1）更深的网络（即 ResNet-101），（2）数据增强，（3）LargeFOV 或 ASPP，以及（4）CRF 作为此数据集的后处理。我们首先发现单独使用 ResNet101 比使用 VGG-16 网更好。使用 LargeFOV 带来 2.6％的改进，使用 ASPP 进一步提高了 1.2％的结果。采用数据增强和 CRF 作为后处理分别带来了 0.6％和 0.4％。

目前的测试结果。我们已将最佳模型上传到评估服务器，获得 70.4％的性能。请注意，我们的模型仅在列车组上进行训练。

定性结果。我们将结果可视化为图 13。

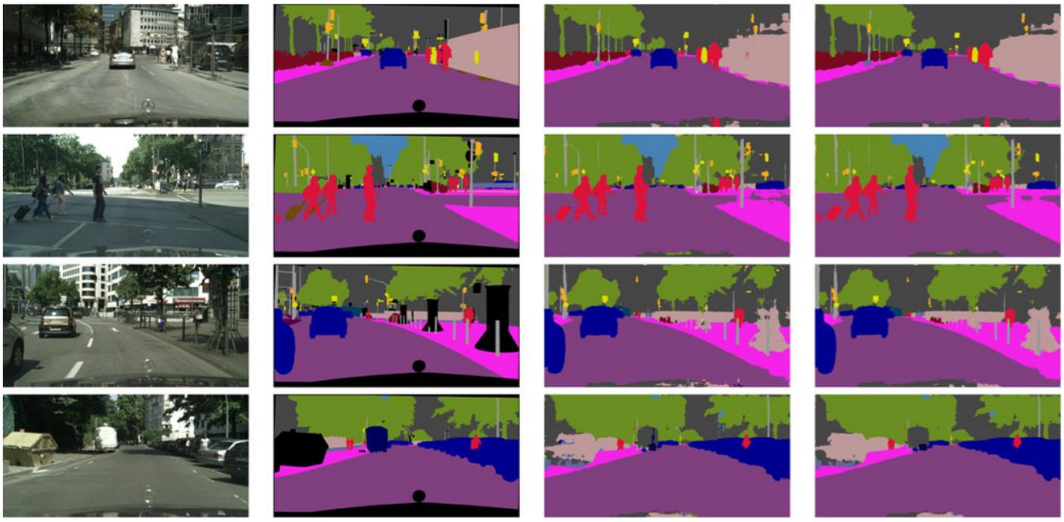
TABLE 8
Test Set Results on the Cityscapes Dataset, Comparing Our DeepLab System with Other State-of-Art Methods

Method	mIOU
pre-release version of dataset	
Adelaide_Context [40]	66.4
FCN-8s [14]	65.3
DeepLab-CRF-LargeFOV-StrongWeak [58]	64.8
DeepLab-CRF-LargeFOV [38]	63.1
CRF-RNN [59]	62.5
DPN [62]	59.1
Segnet basic [100]	57.0
Segnet extended [100]	56.1
official version	
Adelaide_Context [40]	71.6
Dilation10 [76]	67.1
DPN [62]	66.8
Pixel-level Encoding [101]	64.3
DeepLab-CRF (ResNet-101)	70.4

TABLE 9
Val Set Results on Cityscapes Dataset

Full	Aug	LargeFOV	ASPP	CRF	mIOU
VGG-16					
		✓			62.97
		✓		✓	64.18
✓		✓			64.89
✓		✓		✓	65.94
ResNet-101					
✓					66.6
✓		✓			69.2
✓			✓		70.4
✓	✓		✓		71.0
✓	✓		✓	✓	71.4

***Full**: model trained with full resolution images.*

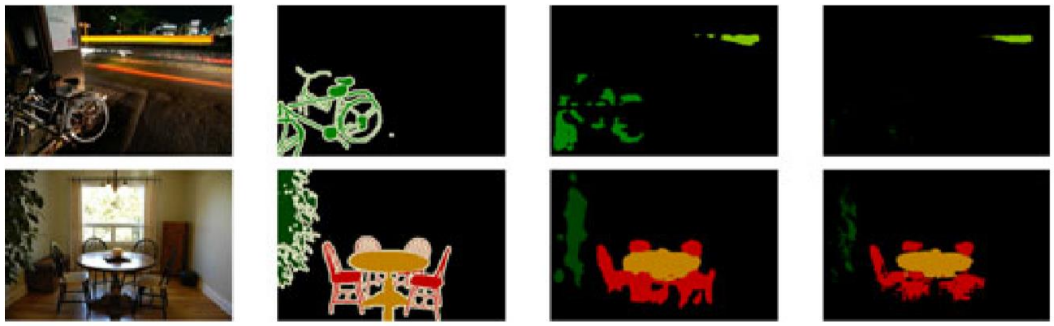


(a) Image (b) G.T. (c) Before CRF (d) After CRF

Fig. 13. Cityscapes results. Input image, ground-truth, and our DeepLab results before/after CRF.

4.5 Failure Modes

我们进一步定性分析了 PASCAL VOC 2012 val set 中我们最佳模型变体的一些失效模式。 如图 14 所示，我们提出的模型无法捕捉到物体的微妙边界，例如自行车和椅子。 CRF 后期处理甚至无法恢复细节，因为一元期限不够充分。 我们假设[100]，[102]的编码器 - 解码器结构可以通过利用解码器路径中的高分辨率特征图来缓解该问题。 如何有效地合并该方法是一项未来的工作。



(a) Image (b) G.T. (c) Before CRF (d) After CRF

Fig. 14. Failure modes. Input image, ground-truth, and our DeepLab results before/after CRF.

5 CONCLUSION

我们提出的“DeepLab”系统通过将“atrous convolution”与上采样滤波器应用于密集特征提取，将图像分类训练的网络重新用于语义分割任务。 我们进一步将其扩展到有害的空间金字塔池，它在多个尺度上编码对象以及图像上下文。 为了沿对象边界生成语义准确的预测和详细的分割图，我们还结合了深度卷积神经网络和完全连接的条件随机场的思想。 我们的实验结果表明，所提出的方法在几个具有挑战性的数据集中显着提升了现有技术水平，包括 PASCAL VOC 2012 语义图像分割基准，PASCAL-Context，PASCAL-Person-Part 和 Cityscapes 数据集。

ACKNOWLEDGMENTS

这项工作部分得到了 ARO 62250-CS，FP7RECONFIG，FP7-MOBOT 和 H2020-ISUPPORT EU 项目的支持。我们非常感谢 NVIDIA 公司支持捐赠用于本研究的 GPU。 前两位作者对这项工作的贡献相同。