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

Liang-Chieh Chen, George Papandreou, *Senior Member, IEEE,* Iasonas Kokkinos, *Member, IEEE,* Kevin Murphy, and Alan L. Yuille, *Fellow, IEEE*

Abstract—In this work we address the task of semantic image segmentation with Deep Learning and make three main contributions that are experimentally shown to have substantial practical merit. *First*, we highlight convolution with upsampled filters, or 'atrous convolution', as a powerful tool in dense prediction tasks. Atrous convolution allows us to explicitly control the resolution at which feature responses are computed within Deep Convolutional Neural Networks. It also allows us to effectively enlarge the field of view of filters to incorporate larger context without increasing the number of parameters or the amount of computation. *Second*, we propose atrous spatial pyramid pooling (ASPP) to robustly segment objects at multiple scales. ASPP probes an incoming convolutional feature layer with filters at multiple sampling rates and effective fields-of-views, thus capturing objects as well as image context at multiple scales. *Third*, we improve the localization of object boundaries by combining methods from DCNNs and probabilistic graphical models. The commonly deployed combination of max-pooling and downsampling in DCNNs achieves invariance but has a toll on localization accuracy. We overcome this by combining the responses at the final DCNN layer with a fully connected Conditional Random Field (CRF), which is shown both qualitatively and quantitatively to improve localization performance. Our proposed "DeepLab" system sets the new state-of-art at the PASCAL VOC-2012 semantic image segmentation task, reaching 79.7% mIOU in the test set, and advances the results on three other datasets: PASCAL-Context, PASCAL-Person-Part, and Cityscapes. All of our code is made publicly available online.

Index Terms—Convolutional Neural Networks, Semantic Segmentation, Atrous Convolution, Conditional Random Fields.

1 Introduction

Deep Convolutional Neural Networks (DCNNs) [1] have pushed the performance of computer vision systems to soaring heights on a broad array of high-level problems, including image classification [2], [3], [4], [5], [6] and object detection [7], [8], [9], [10], [11], [12], where DCNNs trained in an end-to-end manner have delivered strikingly better results than systems relying on hand-crafted features. Essential to this success is the built-in invariance of DCNNs to local image transformations, which allows them to learn increasingly abstract data representations [13]. This invariance is clearly desirable for classification tasks, but can hamper dense prediction tasks such as semantic segmentation, where abstraction of spatial information is undesired.

In particular we consider three challenges in the application of DCNNs to semantic image segmentation: (1) reduced feature resolution, (2) existence of objects at multiple scales, and (3) reduced localization accuracy due to DCNN invariance. Next, we discuss these challenges and our approach to overcome them in our proposed DeepLab system.

The first challenge is caused by the repeated combination of max-pooling and downsampling ('striding') performed at consecutive layers of DCNNs originally designed for image classification [2], [4], [5]. This results in feature maps with significantly reduced spatial resolution when the DCNN is employed in a fully convolutional fashion [14]. In order to overcome this hurdle and efficiently produce denser feature maps, we remove the downsampling operator from the last few max pooling layers of DCNNs and instead *upsample the filters* in subsequent convolutional layers, resulting in feature maps computed at a higher sampling rate. Filter upsampling amounts to inserting holes ('trous' in French) between nonzero filter taps. This technique has a long history in signal processing, originally developed for the efficient computation of the undecimated wavelet transform in a scheme also known as "algorithme à trous" [15]. We use the term *atrous convolution* as a shorthand for convolution with upsampled filters. Various flavors of this idea have been used before in the context of DCNNs by [3], [6], [16]. In practice, we recover full resolution feature maps by a combination of atrous convolution, which computes feature maps more densely, followed by simple bilinear interpolation of the feature responses to the original image size. This scheme offers a simple yet powerful alternative to using deconvolutional layers [13], [14] in dense prediction tasks. Compared to regular convolution with larger filters, atrous convolution allows us to effectively enlarge the field of view of filters without increasing the number of parameters or the amount of computation.

The second challenge is caused by the existence of objects at multiple scales. A standard way to deal with this is to present to the DCNN rescaled versions of the same image and then aggregate the feature or score maps [6], [17], [18]. We show that this approach indeed increases the perfor-

• *L.-C. Chen, G. Papandreou, and K. Murphy are with Google Inc. I. Kokkinos is with University College London. A. Yuille is with the Departments of Cognitive Science and Computer Science, Johns Hopkins University. The first two authors contributed equally to this work.*

DeepLab：基于深度卷积网络、空洞卷积和全连接条件随机场的语义图像分割

梁其琛、乔治·帕潘德里欧、*Senior Member, IEEE,* 亚索纳斯·科基诺斯、*Member, IEEE,* 凯文·墨菲和艾伦·L·尤尔、*Fellow, IEEE*

摘要—在本工作中，我们通过深度学习解决语义图像分割任务，并提出了三个经实验证明具有重要实用价值的主要贡献。*First*，我们重点阐述了采用上采样滤波器的卷积（即“空洞卷积”）作为密集预测任务中的强大工具。空洞卷积使我们能够显式控制深度卷积神经网络中特征响应的计算分辨率，同时在不增加参数量或计算量的前提下，有效扩大滤波器的感受野以融合更广泛的上下文信息。*Second*，我们提出空洞空间金字塔池化（ASPP）模块，以在多尺度下鲁棒地分割目标。ASPP通过多种采样率和有效感受野的滤波器对输入的卷积特征层进行多尺度探测，从而同时捕获多尺度的目标信息与图像上下文。*Third*，我们通过融合深度卷积神经网络与概率图模型的方法提升了目标边界的定位精度。深度卷积神经网络中常用的最大池化与下采样组合虽能保持平移不变性，但会损害定位精度。我们通过将最终深度卷积神经网络层的响应与全连接条件随机场相结合来解决这一问题，定性与定量实验均表明该方法能显著提升边界定位性能。我们提出的“DeepLab”系统在PASCAL VOC-2012语义图像分割任务中取得了79.7%的平均交并比（测试集），刷新了当前最佳性能，并在PASCAL-Context、PASCAL-Person-Part和Cityscapes三个数据集上实现了性能提升。相关代码已全部公开。

索引术语—卷积神经网络，语义分割，空洞卷积，条件随机场。

1 引言

深度卷积神经网络（DCNNs）[1] 在图像分类[2]、[3]、[4]、[5]、[6]和目标检测[7]、[8]、[9]、[10]、[11]、[12]等一系列高层次计算机视觉任务中，将系统性能推向了新的高度。其中，以端到端方式训练的DCNN取得了远超依赖手工特征系统的优异结果。这一成功的关键在于DCNN内置的对局部图像变换的不变性，使其能够学习日益抽象的数据表示[13]。这种不变性对于分类任务显然是有益的，但可能阻碍密集预测任务（如语义分割）的性能，因为此类任务需要保留空间信息而非将其抽象化。

具体而言，我们考虑了将深度卷积神经网络应用于语义图像分割时面临的三个挑战：（1）特征分辨率降低，（2）多尺度物体的存在，以及（3）由于深度卷积神经网络的不变性导致的定位精度下降。接下来，我们将讨论这些挑战，以及在我们提出的DeepLab系统中克服这些挑战的方法。

第一个挑战源于最初为图像分类设计的DCNN连续层中最大池化和下采样（“步进”）的重复组合[2]、[4]、[5]。这导致DCNN生成的特征图空间分辨率显著降低。以全卷积方式采用[14]。为了克服这一障碍并高效生成更密集的特征图，我们从DCNN的最后几个最大池化层中移除了下采样算子，并在后续卷积层中改为*upsamplethe filters*，从而以更高的采样率计算特征图。滤波器上采样相当于在非零滤波器抽头之间插入空洞（法语中称为“trous”）。这项技术在信号处理领域历史悠久，最初是为高效计算未抽取小波变换而开发的，该方案也被称为“algorithme `a trous”[15]。我们使用术语*atrous convolution*作为上采样滤波器卷积的简称。此前在DCNN背景下，[3]、[6]、[16]已运用过这一思想的不同变体。实践中，我们通过结合空洞卷积（以更密集的方式计算特征图）与简单的特征响应双线性插值（恢复至原始图像尺寸），来获取全分辨率特征图。该方案为密集预测任务中使用反卷积层[13][14]提供了一种简单而强大的替代方法。与使用更大滤波器的常规卷积相比，空洞卷积能在不增加参数数量或计算量的情况下，有效扩大滤波器的感受野。

第二个挑战源于多尺度物体的存在。处理这一问题的标准方法是向深度卷积神经网络提供同一图像的不同缩放版本，然后聚合特征图或得分图[6], [17], [18]。我们证明这种方法确实能提升性能——

- *L.-C. Chen, G. Papandreou, and K. Murphy are with Google Inc. I. Kokkinos is with University College London. A. Yuille is with the Departments of Cognitive Science and Computer Science, Johns Hopkins University. The first two authors contributed equally to this work.*

mance of our system, but comes at the cost of computing feature responses at all DCNN layers for multiple scaled versions of the input image. Instead, motivated by spatial pyramid pooling [19], [20], we propose a computationally efficient scheme of resampling a given feature layer at multiple rates prior to convolution. This amounts to probing the original image with multiple filters that have complementary effective fields of view, thus capturing objects as well as useful image context at multiple scales. Rather than actually resampling features, we efficiently implement this mapping using multiple parallel atrous convolutional layers with different sampling rates; we call the proposed technique "atrous spatial pyramid pooling" (ASPP).

The third challenge relates to the fact that an object-centric classifier requires invariance to spatial transformations, inherently limiting the spatial accuracy of a DCNN. One way to mitigate this problem is to use skip-layers to extract "hyper-column" features from multiple network layers when computing the final segmentation result [14], [21]. Our work explores an alternative approach which we show to be highly effective. In particular, we boost our model's ability to capture fine details by employing a fully-connected Conditional Random Field (CRF) [22]. CRFs have been broadly used in semantic segmentation to combine class scores computed by multi-way classifiers with the low-level information captured by the local interactions of pixels and edges [23], [24] or superpixels [25]. Even though works of increased sophistication have been proposed to model the hierarchical dependency [26], [27], [28] and/or high-order dependencies of segments [29], [30], [31], [32], [33], we use the fully connected pairwise CRF proposed by [22] for its efficient computation, and ability to capture fine edge details while also catering for long range dependencies. That model was shown in [22] to improve the performance of a boosting-based pixel-level classifier. In this work, we demonstrate that it leads to state-of-the-art results when coupled with a DCNN-based pixel-level classifier.

A high-level illustration of the proposed DeepLab model is shown in Fig. 1. A deep convolutional neural network (VGG-16 [4] or ResNet-101 [11] in this work) trained in the task of image classification is re-purposed to the task of semantic segmentation by (1) transforming all the fully connected layers to convolutional layers (*i.e.*, fully convolutional network [14]) and (2) increasing feature resolution through atrous convolutional layers, allowing us to compute feature responses every 8 pixels instead of every 32 pixels in the original network. We then employ bi-linear interpolation to upsample by a factor of 8 the score map to reach the original image resolution, yielding the input to a fully-connected CRF [22] that refines the segmentation results.

From a practical standpoint, the three main advantages of our DeepLab system are: (1) Speed: by virtue of atrous convolution, our dense DCNN operates at 8 FPS on an NVidia Titan X GPU, while Mean Field Inference for the fully-connected CRF requires 0.5 secs on a CPU. (2) Accuracy: we obtain state-of-art results on several challenging datasets, including the PASCAL VOC 2012 semantic segmentation benchmark [34], PASCAL-Context [35], PASCAL-Person-Part [36], and Cityscapes [37]. (3) Simplicity: our system is composed of a cascade of two very well-established modules, DCNNs and CRFs.

The updated DeepLab system we present in this paper features several improvements compared to its first version reported in our original conference publication [38]. Our new version can better segment objects at multiple scales, via either multi-scale input processing [17], [39], [40] or the proposed ASPP. We have built a residual net variant of DeepLab by adapting the state-of-art ResNet [11] image classification DCNN, achieving better semantic segmentation performance compared to our original model based on VGG-16 [4]. Finally, we present a more comprehensive experimental evaluation of multiple model variants and report state-of-art results not only on the PASCAL VOC 2012 benchmark but also on other challenging tasks. We have implemented the proposed methods by extending the Caffe framework [41]. We share our code and models at a companion web site http://liangchiehchen.com/projects/DeepLab.html.

2 Related Work

Most of the successful semantic segmentation systems developed in the previous decade relied on hand-crafted features combined with flat classifiers, such as Boosting [24], [42], Random Forests [43], or Support Vector Machines [44]. Substantial improvements have been achieved by incorporating richer information from context [45] and structured prediction techniques [22], [26], [27], [46], but the performance of these systems has always been compromised by the limited expressive power of the features. Over the past few years the breakthroughs of Deep Learning in image classification were quickly transferred to the semantic segmentation task. Since this task involves both segmentation and classification, a central question is how to combine the two tasks.

The first family of DCNN-based systems for semantic segmentation typically employs a cascade of bottom-up image segmentation, followed by DCNN-based region classification. For instance the bounding box proposals and masked regions delivered by [47], [48] are used in [7] and [49] as inputs to a DCNN to incorporate shape information into the classification process. Similarly, the authors of [50] rely on a superpixel representation. Even though these approaches can benefit from the sharp boundaries delivered by a good segmentation, they also cannot recover from any of its errors.

The second family of works relies on using convolutionally computed DCNN features for dense image labeling, and couples them with segmentations that are obtained independently. Among the first have been [39] who apply DCNNs at multiple image resolutions and then employ a segmentation tree to smooth the prediction results. More recently, [21] propose to use skip layers and concatenate the computed intermediate feature maps within the DCNNs for pixel classification. Further, [51] propose to pool the intermediate feature maps by region proposals. These works still employ segmentation algorithms that are decoupled from the DCNN classifier's results, thus risking commitment to premature decisions.

The third family of works uses DCNNs to directly provide dense category-level pixel labels, which makes it possible to even discard segmentation altogether. The

我们系统的性能，但代价是需要在多个缩放版本的输入图像上计算所有DCNN层的特征响应。受空间金字塔池化[19]、[20]的启发，我们提出了一种计算高效的方案，即在卷积前以多种速率对给定特征层进行重采样。这相当于使用多个具有互补有效视野的滤波器对原始图像进行探测，从而在多个尺度上捕获物体以及有用的图像上下文。我们并未实际重采样特征，而是通过使用具有不同采样率的多个并行空洞卷积层高效实现这一映射；我们将该技术称为“空洞空间金字塔池化”（ASPP）。

第三个挑战涉及这样一个事实：以对象为中心的分类器需要对空间变换具有不变性，这本质上限制了深度卷积神经网络（DCNN）的空间精度。缓解此问题的一种方法是使用跳跃层在计算最终分割结果时从多个网络层提取“超列”特征[14][21]。我们的研究探索了一种替代方法，并证明其极为有效。具体而言，我们通过采用全连接条件随机场（CRF）[22]来增强模型捕捉细节的能力。CRF已广泛应用于语义分割中，将多路分类器计算的类别分数与像素和边缘局部交互[23][24]或超像素[25]捕获的低层信息相结合。尽管已有更复杂的工作被提出以建模层次依赖性[26][27][28]和/或分段的高阶依赖性[29][30][31][32][33]，我们仍采用[22]提出的全连接成对CRF，因其计算高效，既能捕捉精细边缘细节，又能兼顾长程依赖关系。[22]中表明该模型提升了基于提升算法的像素级分类器性能。在本工作中，我们证明当它与基于DCNN的像素级分类器结合时，能够达到最先进的结果。

所提出的DeepLab模型的高层示意图如图1所示。通过以下方式，将训练用于图像分类任务的深度卷积神经网络（本工作中采用VGG-16 [4] 或 ResNet-101 [11]）改造为语义分割任务：（1）将所有全连接层转换为卷积层（*i.e*，即全卷积网络[14]）；（2）通过空洞卷积层提高特征分辨率，使特征响应计算间隔从原始网络的每32像素缩短至每8像素。随后采用双线性插值将得分图上采样8倍至原始图像分辨率，并输入全连接条件随机场[22]以优化分割结果。

从实践角度来看，我们的DeepLab系统主要有三大优势：(1) 速度：借助空洞卷积，我们的密集DCNN在NVidia Titan X GPU上能以8 FPS运行，而全连接CRF的平均场推断在CPU上仅需0.5秒。(2) 准确性：我们在多个具有挑战性的数据集上取得了最先进的结果，包括PASCAL VOC 2012语义分割基准[34]、PASCAL-Context[35]、PASCAL-Person-Part[36]和Cityscapes[37]。(3) 简洁性：我们的系统由两个成熟模块（DCNN和CRF）级联构成。

我们在本文中提出的更新版DeepLab系统，相较于最初会议论文[38]中报告的第一个版本，具有多项改进。我们的新版本能够通过多尺度输入处理[17][39][40]或提出的ASPP模块，更好地分割多尺度物体。我们通过适配最先进的ResNet[11]图像分类DCNN，构建了DeepLab的残差网络变体，相比基于VGG-16[4]的原始模型实现了更好的语义分割性能。最后，我们对多种模型变体进行了更全面的实验评估，不仅在PASCAL VOC 2012基准测试中取得了领先结果，在其他挑战性任务中也报告了最优性能。我们通过扩展Caffe框架[41]实现了所提出的方法，并在配套网站http://liangchiehchen.com/projects/DeepLab.html上公开了代码与模型。

2 相关工作

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（如Boosting [24]、[42]、随机森林 [43] 或支持向量机 [44]）的结合。通过融入上下文 [45] 和结构化预测技术 [22]、[26]、[27]、[46] 的更丰富信息，系统性能得到了显著提升，但这些系统的表现始终受限于特征表达能力的不足。过去几年，深度学习在图像分类领域的突破迅速被应用于语义分割任务。由于该任务同时涉及分割与分类，一个核心问题是如何将这两项任务结合起来。

首个基于DCNN的语义分割系统家族通常采用自底向上的图像分割级联，随后进行基于DCNN的区域分类。例如，[47]、[48]提出的边界框建议和掩码区域在[7]和[49]中被用作DCNN的输入，以将形状信息融入分类过程。类似地，[50]的作者依赖于超像素表示。尽管这些方法能够受益于优质分割所提供的清晰边界，但它们同样无法修正分割过程中的任何错误。

第二类研究依赖于使用卷积计算的DCNN特征进行密集图像标注，并将其与独立获取的分割结果相结合。其中最早的研究之一是[39]，他们在多个图像分辨率上应用DCNN，然后利用分割树对预测结果进行平滑处理。最近，[21]提出使用跳跃层并将DCNN内计算得到的中间特征图进行拼接，以用于像素分类。此外，[51]提出通过区域建议对中间特征图进行池化。这些研究仍采用与DCNN分类器结果解耦的分割算法，因此可能过早做出决策而带来风险。

第三类工作使用深度卷积神经网络直接提供密集的类别级像素标签，这使得完全放弃分割成为可能。

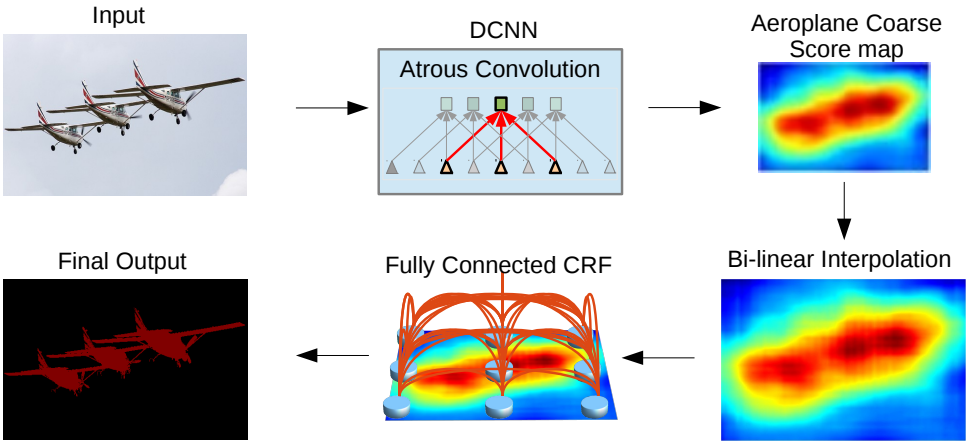


Fig. 1: Model Illustration. A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original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

segmentation-free approaches of [14], [52] directly apply DCNNs to the whole image in a fully convolutional fashion, transforming the last fully connected layers of the DCNN into convolutional layers. In order to deal with the spatial localization issues outlined in the introduction, [14] upsample and concatenate the scores from intermediate feature maps, while [52] refine the prediction result from coarse to fine by propagating the coarse results to another DCNN. Our work builds on these works, and as described in the introduction extends them by exerting control on the feature resolution, introducing multi-scale pooling techniques and integrating the densely connected CRF of [22] on top of the DCNN. We show that this leads to significantly better segmentation results, especially along object boundaries. The combination of DCNN and CRF is of course not new but previous works only tried locally connected CRF models. Specifically, [53] use CRFs as a proposal mechanism for a DCNN-based reranking system, while [39] treat superpixels as nodes for a local pairwise CRF and use graph-cuts for discrete inference. As such their models were limited by errors in superpixel computations or ignored long-range dependencies. Our approach instead treats every pixel as a CRF node receiving unary potentials by the DCNN. Crucially, the Gaussian CRF potentials in the fully connected CRF model of [22] that we adopt can capture long-range dependencies and at the same time the model is amenable to fast mean field inference. We note that mean field inference had been extensively studied for traditional image segmentation tasks [54], [55], [56], but these older models were typically limited to short-range connections. In independent work, [57] use a very similar densely connected CRF model to refine the results of DCNN for the problem of material classification. However, the DCNN module of [57] was only trained by sparse point supervision instead of dense supervision at every pixel.

Since the first version of this work was made publicly available [38], the area of semantic segmentation has progressed drastically. Multiple groups have made important advances, significantly raising the bar on the PASCAL VOC 2012 semantic segmentation benchmark, as reflected to the high level of activity in the benchmark's leaderboard[1] [17], [40], [58], [59], [60], [61], [62], [63]. Interestingly, most top-performing methods have adopted one or both of the key ingredients of our DeepLab system: Atrous convolution for efficient dense feature extraction and refinement of the raw DCNN scores by means of a fully connected CRF. We outline below some of the most important and interesting advances.

End-to-end training for structured prediction has more recently been explored in several related works. While we employ the CRF as a post-processing method, [40], [59], [62], [64], [65] have successfully pursued joint learning of the DCNN and CRF. In particular, [59], [65] unroll the CRF mean-field inference steps to convert the whole system into an end-to-end trainable feed-forward network, while [62] approximates one iteration of the dense CRF mean field inference [22] by convolutional layers with learnable filters. Another fruitful direction pursued by [40], [66] is to learn the pairwise terms of a CRF via a DCNN, significantly improving performance at the cost of heavier computation. In a different direction, [63] replace the bilateral filtering module used in mean field inference with a faster domain transform module [67], improving the speed and lowering the memory requirements of the overall system, while [18], [68] combine semantic segmentation with edge detection.

Weaker supervision has been pursued in a number of papers, relaxing the assumption that pixel-level semantic annotations are available for the whole training set [58], [69], [70], [71], achieving significantly better results than weakly-supervised pre-DCNN systems such as [72]. In another line of research, [49], [73] pursue instance segmentation, jointly tackling object detection and semantic segmentation.

What we call here *atrous convolution* was originally developed for the efficient computation of the undecimated wavelet transform in the "algorithme à trous" scheme of [15]. We refer the interested reader to [74] for early references from the wavelet literature. Atrous convolution is also intimately related to the "noble identities" in multi-rate signal processing, which builds on the same interplay of input

1. http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6

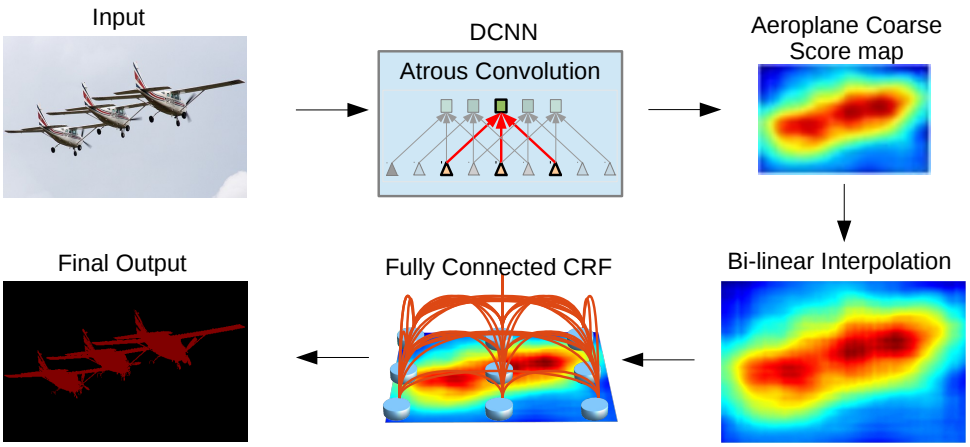


图1：模型示意图。采用深度卷积神经网络（如VGG-16或ResNet-101）以全卷积方式运行，利用空洞卷积降低信号下采样程度（从32倍降至8倍）。通过双线性插值阶段将特征图放大至原始图像分辨率。随后应用全连接条件随机场优化分割结果，以更精确地捕捉物体边界。

[14]、[52]的无分割方法以全卷积方式直接将DCNN应用于整个图像，将DCNN最后的全连接层转换为卷积层。为了处理引言中概述的空间定位问题，[14]对中间特征图的得分进行上采样和拼接，而[52]通过将粗略结果传播到另一个DCNN中，实现从粗到精的预测结果细化。我们的工作基于这些研究，并如引言所述，通过控制特征分辨率、引入多尺度池化技术以及在DCNN之上集成[22]的密集连接CRF对其进行了扩展。我们证明这能显著改善分割结果，尤其是沿物体边界的分割效果。当然，DCNN与CRF的结合并非首创，但先前的研究仅尝试了局部连接的CRF模型。具体而言，[53]将CRF用作基于DCNN的重排序系统的建议机制，而[39]将超像素视为局部成对CRF的节点，并使用图割进行离散推理。因此，他们的模型受限于超像素计算误差或忽略了长程依赖关系。我们的方法则将每个像素视为接收DCNN一元势能的CRF节点。关键在于，我们所采用的[22]全连接CRF模型中的高斯CRF势能可以捕获长程依赖关系，同时该模型适用于快速均值场推理。我们注意到，均值场推理在传统图像分割任务[54]、[55]、[56]中已被广泛研究，但这些早期模型通常仅限于短程连接。在独立工作中，[57]使用非常相似的密集连接CRF模型来细化DCNN在材料分类问题上的结果。然而，[57]的DCNN模块仅通过稀疏点监督而非逐像素的密集监督进行训练。

自本研究的首个版本公开以来[38]，语义分割领域取得了巨大进展。多个研究团队实现了重要突破，显著提升了PASCAL VOC 2012语义分割基准的标杆水平，这体现在基准测试排行榜上活跃度很高[1] [17], [40], [58], [59], [60], [61], [62], [63]。有趣的是，大多数表现最佳的方法都采用了我们DeepLab系统的一个或两个关键要素：用于高效密集特征提取的空洞卷积，以及通过全连接CRF优化原始DCNN得分。下文我们将概述其中一些最重要且有趣的进展。

End-to-end training for structured prediction 最近，在几项相关工作中，这一方向得到了进一步探索。尽管我们将CRF用作后处理方法，但[40]、[59]、[62]、[64]、[65]已成功实现了DCNN与CRF的联合学习。特别是[59]和[65]通过展开CRF均值场推断步骤，将整个系统转换为端到端可训练的前馈网络；而[62]则利用可学习滤波器的卷积层来近似密集CRF均值场推断[22]的单次迭代。另一条由[40]和[66]探索的有效路径是通过DCNN学习CRF的成对项，虽以更高计算成本为代价，但显著提升了性能。在另一方向上，[63]用更快的域变换模块[67]替代均值场推断中使用的双边滤波模块，从而提高了整体系统的速度并降低了内存需求；同时[18]和[68]将语义分割与边缘检测相结合。

Weaker supervision 已在多篇论文中得到探索，这些研究放宽了要求整个训练集具备像素级语义标注的假设[58]、[69]、[70]、[71]，取得了比弱监督的预深度卷积网络系统（如[72]）显著更好的结果。在另一研究方向中，[49]、[73]致力于实例分割，将目标检测与语义分割任务共同解决。

我们这里所称的*atrous convolution*最初是为[15]提出的“algorithme à trous”方案中高效计算非抽取小波变换而开发的。关于小波文献中的早期参考文献，我们建议感兴趣的读者查阅[74]。空洞卷积也与多速率信号处理中的“贵族恒等式”密切相关，二者都建立在输入信号相同交织关系的基础上。

1. http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6

signal and filter sampling rates [75]. Atrous convolution is a term we first used in [6]. The same operation was later called dilated convolution by [76], a term they coined motivated by the fact that the operation corresponds to regular convolution with upsampled (or dilated in the terminology of [15]) filters. Various authors have used the same operation before for denser feature extraction in DCNNs [3], [6], [16]. Beyond mere resolution enhancement, atrous convolution allows us to enlarge the field of view of filters to incorporate larger context, which we have shown in [38] to be beneficial. This approach has been pursued further by [76], who employ a series of atrous convolutional layers with increasing rates to aggregate multiscale context. The atrous spatial pyramid pooling scheme proposed here to capture multiscale objects and context also employs multiple atrous convolutional layers with different sampling rates, which we however lay out in parallel instead of in serial. Interestingly, the atrous convolution technique has also been adopted for a broader set of tasks, such as object detection [12], [77], instance-level segmentation [78], visual question answering [79], and optical flow [80].

We also show that, as expected, integrating into DeepLab more advanced image classification DCNNs such as the residual net of [11] leads to better results. This has also been observed independently by [81].

3 Methods

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

The use of DCNNs for semantic segmentation, or other dense prediction tasks, has been shown to be simply and successfully addressed by deploying DCNNs in a fully convolutional fashion [3], [14]. However, the repeated combination of max-pooling and striding at consecutive layers of these networks reduces significantly the spatial resolution of the resulting feature maps, typically by a factor of 32 across each direction in recent DCNNs. A partial remedy is to use 'deconvolutional' layers as in [14], which however requires additional memory and time.

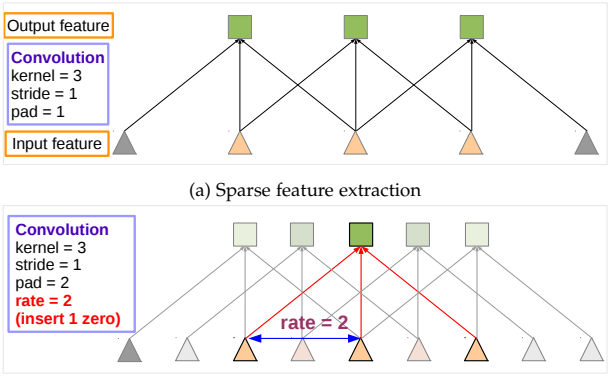
We advocate instead the use of atrous convolution, originally developed for the efficient computation of the undecimated wavelet transform in the "algorithme à trous" scheme of [15] and used before in the DCNN context by [3], [6], [16]. This algorithm allows us to compute the responses of any layer at any desirable resolution. It can be applied post-hoc, once a network has been trained, but can also be seamlessly integrated with training.

Considering one-dimensional signals first, the output $y[i]$ of atrous convolution [2] of a 1-D input signal $x[i]$ with a filter $w[k]$ of length K is defined as:

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

The *rate* parameter r corresponds to the stride with which we sample the input signal. Standard convolution is a special case for rate $r = 1$. See Fig. 2 for illustration.

2. We follow the standard practice in the DCNN literature and use non-mirrored filters in this definition.



(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.

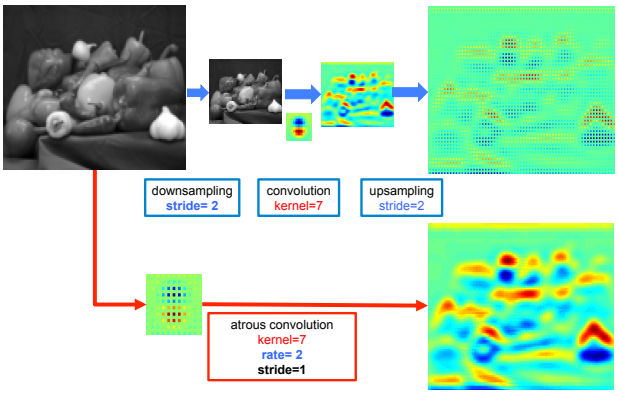
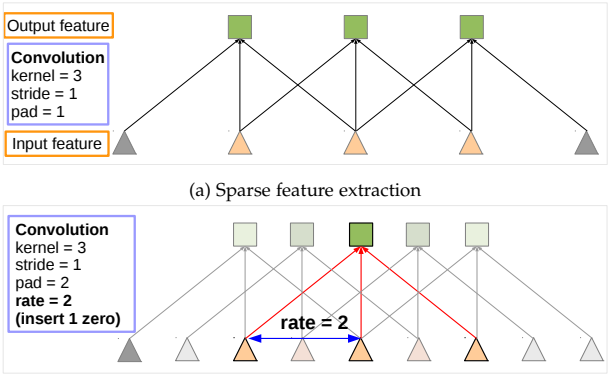


Fig. 3: Illustration of atrous convolution in 2-D. Top row: sparse feature extraction with standard convolution on a low resolution input feature map. Bottom row: Dense feature extraction with atrous convolution with rate $r = 2$, applied on a high resolution input feature map.

We illustrate the algorithm's operation in 2-D through a simple example in Fig. 3: Given an image, we assume that we first have a downsampling operation that reduces the resolution by a factor of 2, and then perform a convolution with a kernel - here, the vertical Gaussian derivative. If one implants the resulting feature map in the original image coordinates, we realize that we have obtained responses at only 1/4 of the image positions. Instead, we can compute responses at all image positions if we convolve the full resolution image with a filter 'with holes', in which we upsample the original filter by a factor of 2, and introduce zeros in between filter values. Although the effective filter size increases, we only need to take into account the non-zero filter values, hence both the number of filter parameters and the number of operations per position stay constant. The resulting scheme allows us to easily and explicitly control the spatial resolution of neural network feature responses.

In the context of DCNNs one can use atrous convolution in a chain of layers, effectively allowing us to compute the

信号和滤波器的采样率[75]。空洞卷积是我们首次在[6]中使用的术语。同一操作后来被[76]称为膨胀卷积，他们创造这一术语的动机在于该操作对应于使用上采样（或按[15]的术语称为膨胀）滤波器进行常规卷积。多位学者此前已在深度卷积神经网络中将该操作用于更密集的特征提取[3]、[6]、[16]。除了单纯提升分辨率外，空洞卷积还能扩大滤波器的感受野以融合更广泛的上下文信息，我们在[38]中已证明这具有显著优势。[76]进一步推进了该方法，他们采用一系列采样率递增的空洞卷积层来聚合多尺度上下文。本文提出的用于捕获多尺度目标及上下文的空洞空间金字塔池化方案，同样采用了多个具有不同采样率的空洞卷积层，但我们将其并行排列而非串联布置。值得注意的是，空洞卷积技术已被更广泛地应用于各类任务，如目标检测[12]、[77]、实例级分割[78]、视觉问答[79]以及光流估计[80]。



(a) Sparse feature extraction

(b) Dense feature extraction

图2：一维空洞卷积示意图。(a) 在低分辨率输入特征图上使用标准卷积进行稀疏特征提取。(b) 在高分辨率输入特征图上应用空洞率为$r = 2$的空洞卷积进行密集特征提取。

我们还表明，正如预期的那样，将更先进的图像分类深度卷积神经网络（如[11]的残差网络）集成到DeepLab中会带来更好的结果。[81]的研究也独立观察到了这一点。

3 种方法

3.1 用于密集特征提取和视野扩大的空洞卷积

将深度卷积神经网络（DCNNs）应用于语义分割或其他密集预测任务时，通过以全卷积方式部署DCNNs已被证明是简单且有效的解决方案[3], [14]。然而，在这些网络的连续层中反复结合最大池化和步进操作，会显著降低最终特征图的空间分辨率——在近期的DCNNs中，通常每个方向的分辨率会缩减至原来的1/32。一种部分解决方案是采用如[14]所述的“反卷积”层，但这需要额外的内存和时间开销。

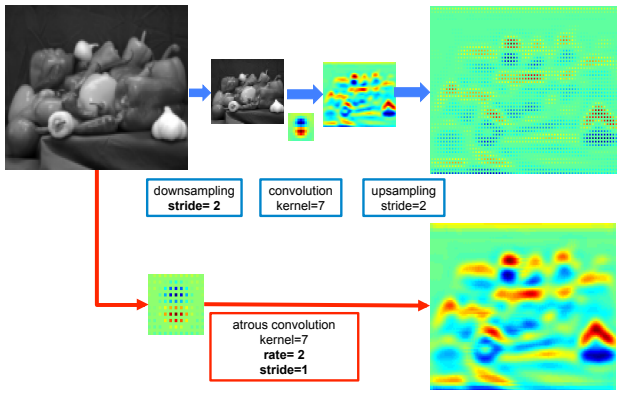


图3：二维空洞卷积示意图。上行：在低分辨率输入特征图上使用标准卷积进行稀疏特征提取。下行：在高分辨率输入特征图上应用空洞率为$r = 2$的空洞卷积进行密集特征提取。

我们转而提倡使用空洞卷积，该技术最初在[15]的“algorithme {v*}à trous”方案中为高效计算非抽取小波变换而开发，并曾被[3]、[6]、[16]应用于DCNN领域。此算法允许我们以任意期望的分辨率计算任何层的响应。它可以在网络训练完成后事后应用，也能无缝集成到训练过程中。

首先考虑一维信号，对于长度为K的滤波器$w[k]$，一维输入信号$x[i]$的空洞卷积[2]输出$y[i]$定义为：

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

*rate*参数r对应于我们对输入信号进行采样的步长。标准卷积是速率$r =$为1时的特例。图示请参见图2。

我们通过图3中的一个简单示例在二维空间中说明算法的操作：给定一张图像，我们假设首先进行下采样操作，将分辨率降低2倍，然后使用一个核进行卷积——此处为垂直高斯导数核。若将得到的特征图嵌入原始图像坐标系中，我们会发现仅在1/4的图像位置上获得了响应。相反，如果我们将全分辨率图像与一个“带孔”的滤波器进行卷积，则可以在所有图像位置上计算响应——该滤波器将原始滤波器上采样2倍，并在滤波器值之间插入零值。虽然有效滤波器尺寸增大，但我们只需考虑非零滤波器值，因此滤波器参数数量和每个位置的计算量均保持不变。这种方案使我们能够轻松且显式地控制神经网络特征响应的空间分辨率。

在深度卷积神经网络（DCNN）的背景下，可以在层链中使用空洞卷积，这使我们能够有效地计算

2. 我们遵循DCNN文献中的标准做法，在此定义中使用非镜像滤波器。

final DCNN network responses at an arbitrarily high resolution. For example, in order to double the spatial density of computed feature responses in the VGG-16 or ResNet-101 networks, we find the last pooling or convolutional layer that decreases resolution ('pool5' or 'conv5_1' respectively), set its stride to 1 to avoid signal decimation, and replace all subsequent convolutional layers with atrous convolutional layers having rate $r = 2$. Pushing this approach all the way through the network could allow us to compute feature responses at the original image resolution, but this ends up being too costly. We have adopted instead a hybrid approach that strikes a good efficiency/accuracy trade-off, using atrous convolution to increase by a factor of 4 the density of computed feature maps, followed by fast bilinear interpolation by an additional factor of 8 to recover feature maps at the original image resolution. Bilinear interpolation is sufficient in this setting because the class score maps (corresponding to log-probabilities) are quite smooth, as illustrated in Fig. 5. Unlike the deconvolutional approach adopted by [14], the proposed approach converts image classification networks into dense feature extractors without requiring learning any extra parameters, leading to faster DCNN training in practice.

Atrous convolution also allows us to arbitrarily enlarge the *field-of-view* of filters at any DCNN layer. State-of-the-art DCNNs typically employ spatially small convolution kernels (typically 3×3) in order to keep both computation and number of parameters contained. Atrous convolution with rate r introduces $r - 1$ zeros between consecutive filter values, effectively enlarging the kernel size of a $k \times k$ filter to $k_e = k + (k - 1)(r - 1)$ without increasing the number of parameters or the amount of computation. It thus offers an efficient mechanism to control the field-of-view and finds the best trade-off between accurate localization (small field-of-view) and context assimilation (large field-of-view). We have successfully experimented with this technique: Our DeepLab-LargeFOV model variant [38] employs atrous convolution with rate $r = 12$ in VGG-16 'fc6' layer with significant performance gains, as detailed in Section 4.

Turning to implementation aspects, there are two efficient ways to perform atrous convolution. The first is to implicitly upsample the filters by inserting holes (zeros), or equivalently sparsely sample the input feature maps [15]. We implemented this in our earlier work [6], [38], followed by [76], within the Caffe framework [41] by adding to the *im2col* function (it extracts vectorized patches from multi-channel feature maps) the option to sparsely sample the underlying feature maps. The second method, originally proposed by [82] and used in [3], [16] is to subsample the input feature map by a factor equal to the atrous convolution rate r, deinterlacing it to produce r^2 reduced resolution maps, one for each of the $r \times r$ possible shifts. This is followed by applying standard convolution to these intermediate feature maps and reinterlacing them to the original image resolution. By reducing atrous convolution into regular convolution, it allows us to use off-the-shelf highly optimized convolution routines. We have implemented the second approach into the TensorFlow framework [83].

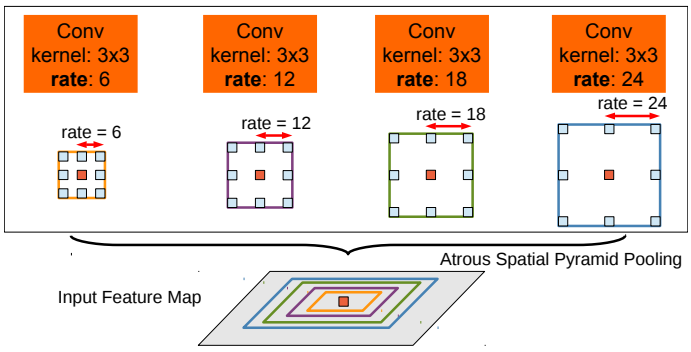


Fig. 4: Atrous Spatial Pyramid Pooling (ASPP). To classify the center pixel (orange), ASPP exploits multi-scale features by employing multiple parallel filters with different rates. The effective Field-Of-Views are shown in different colors.

3.2 Multiscale Image Representations using Atrous Spatial Pyramid Pooling

DCNNs have shown a remarkable ability to implicitly represent scale, simply by being trained on datasets that contain objects of varying size. Still, explicitly accounting for object scale can improve the DCNN's ability to successfully handle both large and small objects [6].

We have experimented with two approaches to handling scale variability in semantic segmentation. The first approach amounts to standard multiscale processing [17], [18]. We extract DCNN score maps from multiple (three in our experiments) rescaled versions of the original image using parallel DCNN branches that share the same parameters. To produce the final result, we bilinearly interpolate the feature maps from the parallel DCNN branches to the original image resolution and fuse them, by taking at each position the maximum response across the different scales. We do this both during training and testing. Multiscale processing significantly improves performance, but at the cost of computing feature responses at all DCNN layers for multiple scales of input.

The second approach is inspired by the success of the R-CNN spatial pyramid pooling method of [20], which showed that regions of an arbitrary scale can be accurately and efficiently classified by resampling convolutional features extracted at a single scale. We have implemented a variant of their scheme which uses multiple parallel atrous convolutional layers with different sampling rates. The features extracted for each sampling rate are further processed in separate branches and fused to generate the final result. The proposed "atrous spatial pyramid pooling" (DeepLab-ASPP) approach generalizes our DeepLab-LargeFOV variant and is illustrated in Fig. 4.

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

A trade-off between localization accuracy and classification performance seems to be inherent in DCNNs: deeper models with multiple max-pooling layers have proven most successful in classification tasks, however the increased invariance and the large receptive fields of top-level nodes can only yield smooth responses. As illustrated in Fig. 5, DCNN

最终DCNN网络响应可在任意高分辨率下获得。例如，为了将VGG-16或ResNet-101网络中计算特征响应的空间密度提升一倍，我们找到最后一个降低分辨率的池化层或卷积层（分别为'pool5'或'conv5_1'），将其步幅设为1以避免信号抽取，并将后续所有卷积层替换为膨胀率为$r = 2$的膨胀卷积层。若将此方法贯穿整个网络，理论上可在原始图像分辨率下计算特征响应，但这会导致计算成本过高。因此我们采用了一种混合策略，在效率与精度间取得良好平衡：先使用膨胀卷积将计算特征图的密度提升4倍，再通过快速双线性插值额外放大8倍，以恢复原始图像分辨率的特征图。在此场景下双线性插值已足够适用，因为类别得分图（对应对数概率）相当平滑，如图5所示。与[14]采用的反卷积方法不同，本方法无需学习额外参数即可将图像分类网络转换为密集特征提取器，从而在实践中实现更快的DCNN训练。

空洞卷积还允许我们在任何DCNN层任意放大滤波器的*field-of-view*。最先进的DCNN通常采用空间尺寸较小的卷积核（通常为3×3），以控制计算量和参数数量。使用r倍率的空洞卷积会在连续滤波器值之间插入$r-$个零，从而在不增加参数数量或计算量的情况下，将$k \times k$滤波器的核尺寸有效扩大至$k_e = k + (k - \times 1)(r - \times 1)$。这提供了一种高效的控制感受野的机制，并在精确定位（小感受野）与上下文融合（大感受野）之间找到了最佳平衡。我们已成功应用该技术：如第4节详述，我们的DeepLab-LargeFOV模型变体[38]在VGG-16的'fc6'层采用r =倍率的空洞卷积，取得了显著的性能提升。

转向实现方面，有两种高效执行空洞卷积的方法。第一种是通过插入孔洞（零值）隐式上采样滤波器，或等效地稀疏采样输入特征图[15]。我们在早期工作[6]、[38]中实现了这种方法，随后[76]在Caffe框架[41]中通过为im2col函数（该函数从多通道特征图中提取向量化图像块）增加稀疏采样底层特征图的选项来实现。第二种方法最初由[82]提出并在[3]、[16]中使用，即按空洞卷积速率r对输入特征图进行下采样，通过解交织生成r^2个降低分辨率的特征图，每个图对应$r \times r$种可能的偏移之一。随后对这些中间特征图应用标准卷积，再通过重新交织恢复到原始图像分辨率。通过将空洞卷积转化为常规卷积，这种方法允许我们使用现成的高度优化的卷积例程。我们已在TensorFlow框架[83]中实现了第二种方法。

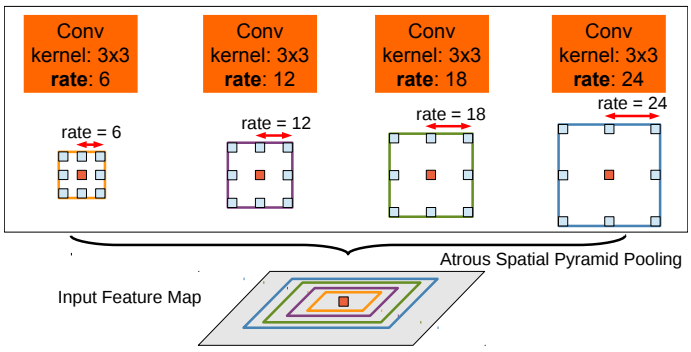


图4：空洞空间金字塔池化（ASPP）。为对中心像素（橙色）进行分类，ASPP通过采用多个不同膨胀率的并行滤波器来利用多尺度特征。有效感受野以不同颜色标示。

3.2 使用空洞空间金字塔池化的多尺度图像表示

深度卷积神经网络（DCNN）已展现出一种显著的能力，能够通过仅在包含不同尺寸物体的数据集上进行训练，来隐式地表示尺度。尽管如此，显式地考虑物体尺度仍能提升DCNN成功处理大小物体的能力[6]。

我们尝试了两种方法来处理语义分割中的尺度变化问题。第一种方法采用标准的多尺度处理技术[17][18]。我们通过共享参数的并行DCNN分支，从原始图像的多个重缩放版本（实验中为三个尺度）提取DCNN得分图。为生成最终结果，我们将并行DCNN分支的特征图双线性插值至原始图像分辨率，并通过逐位置选取不同尺度间的最大响应值进行融合。该方法在训练和测试阶段均被采用。多尺度处理显著提升了性能，但代价是需要为多个输入尺度计算所有DCNN层的特征响应。

第二种方法受到[20]中R-CNN空间金字塔池化方法成功的启发，该方法表明通过重采样在单一尺度提取的卷积特征，可以准确高效地对任意尺度的区域进行分类。我们实现了该方案的一个变体，该变体使用多个具有不同采样率的并行空洞卷积层。针对每个采样率提取的特征会在独立分支中进一步处理，并融合以生成最终结果。所提出的“空洞空间金字塔池化”（DeepLab-ASPP）方法推广了我们的DeepLab-LargeFOV变体，如图4所示。

3.3 基于全连接条件随机场的结构化预测以实现精确边界恢复

在深度卷积神经网络（DCNN）中，定位精度与分类性能之间似乎存在一种固有的权衡：具有多个最大池化层的更深层模型在分类任务中被证明最为成功，然而顶层节点增强的不变性和大感受野只能产生平滑的响应。如图5所示，DCNN

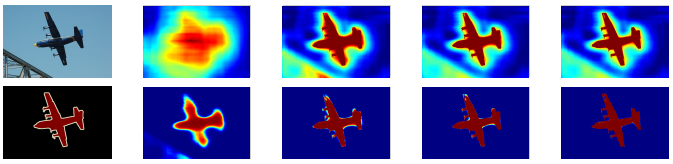


Fig. 5: Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. We show the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of last DCNN layer is used as input to the mean field inference.

score maps can predict the presence and rough position of objects but cannot really delineate their borders.

Previous work has pursued two directions to address this localization challenge. The first approach is to harness information from multiple layers in the convolutional network in order to better estimate the object boundaries [14], [21], [52]. The second is to employ a super-pixel representation, essentially delegating the localization task to a low-level segmentation method [50].

We pursue an alternative direction based on coupling the recognition capacity of DCNNs and the fine-grained localization accuracy of fully connected CRFs and show that it is remarkably successful in addressing the localization challenge, producing accurate semantic segmentation results and recovering object boundaries at a level of detail that is well beyond the reach of existing methods. This direction has been extended by several follow-up papers [17], [40], [58], [59], [60], [61], [62], [63], [65], since the first version of our work was published [38].

Traditionally, conditional random fields (CRFs) have been employed to smooth noisy segmentation maps [23], [31]. Typically these models couple neighboring nodes, favoring same-label assignments to spatially proximal pixels. Qualitatively, the primary function of these short-range CRFs is to clean up the spurious predictions of weak classifiers built on top of local hand-engineered features.

Compared to these weaker classifiers, modern DCNN architectures such as the one we use in this work produce score maps and semantic label predictions which are qualitatively different. As illustrated in Fig. 5, the score maps are typically quite smooth and produce homogeneous classification results. In this regime, using short-range CRFs can be detrimental, as our goal should be to recover detailed local structure rather than further smooth it. Using contrast-sensitive potentials [23] in conjunction to local-range CRFs can potentially improve localization but still miss thin-structures and typically requires solving an expensive discrete optimization problem.

To overcome these limitations of short-range CRFs, we integrate into our system the fully connected CRF model of [22]. The model employs the energy function

$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \tag{2}$$

where $\boldsymbol{x}$ is the label assignment for pixels. We use as unary potential $\theta_i(x_i) = -\log P(x_i)$, where $P(x_i)$ is the label assignment probability at pixel i as computed by a DCNN. The pairwise potential has a form that allows for efficient inference while using a fully-connected graph, i.e. when connecting all pairs of image pixels, i, j. In particular, as in [22], we use the following expression:

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j) \left[w_1 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2} \right) + w_2 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2} \right) \right] \tag{3}$$

where $\mu(x_i, x_j) = 1$ if $x_i \neq x_j$, and zero otherwise, which, as in the Potts model, means that only nodes with distinct labels are penalized. The remaining expression uses two Gaussian kernels in different feature spaces; the first, 'bilateral' kernel depends on both pixel positions (denoted as p) and RGB color (denoted as I), and the second kernel only depends on pixel positions. The hyper parameters σ_α, σ_β and σ_γ control the scale of Gaussian kernels. The first kernel forces pixels with similar color and position to have similar labels, while the second kernel only considers spatial proximity when enforcing smoothness.

Crucially, this model is amenable to efficient approximate probabilistic inference [22]. The message passing updates under a fully decomposable mean field approximation $b(\boldsymbol{x}) = \prod_i b_i(x_i)$ can be expressed as Gaussian convolutions in bilateral space. High-dimensional filtering algorithms [84] significantly speed-up this computation resulting in an algorithm that is very fast in practice, requiring less that 0.5 sec on average for Pascal VOC images using the publicly available implementation of [22].

4 Experimental Results

We finetune the model weights of the Imagenet-pretrained VGG-16 or ResNet-101 networks to adapt them to the semantic segmentation task in a straightforward fashion, following the procedure of [14]. We replace the 1000-way Imagenet classifier in the last layer with a classifier having as many targets as the number of semantic classes of our task (including the background, if applicable). Our loss function is the sum of cross-entropy terms for each spatial position in the CNN output map (subsampled by 8 compared to the original image). All positions and labels are equally weighted in the overall loss function (except for unlabeled pixels which are ignored). Our targets are the ground truth labels (subsampled by 8). We optimize the objective function with respect to the weights at all network layers by the standard SGD procedure of [2]. We decouple the DCNN and CRF training stages, assuming the DCNN unary terms are fixed when setting the CRF parameters.

We evaluate the proposed models on four challenging datasets: PASCAL VOC 2012, PASCAL-Context, PASCAL-Person-Part, and Cityscapes. We first report the main results of our conference version [38] on PASCAL VOC 2012, and move forward to latest results on all datasets.

4.1 PASCAL VOC 2012

Dataset: The PASCAL VOC 2012 segmentation benchmark [34] involves 20 foreground object classes and one background class. The original dataset contains $1,464$ (*train*),

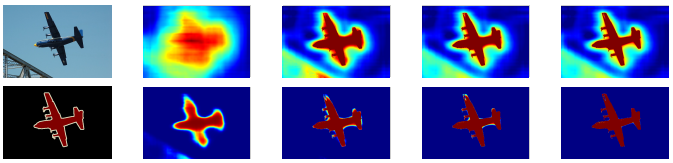


图5：飞机类别的得分图（softmax函数前的输入）与置信度图（softmax函数后的输出）。我们展示了每次均值场迭代后的得分图（第一行）与置信度图（第二行）。最后一个DCNN层的输出被用作均值场推理的输入。

得分图可以预测物体的存在和大致位置，但无法真正勾勒出它们的边界。

先前的研究为应对这一定位挑战，已沿着两个方向展开。第一种方法是利用卷积网络中多层的信息，以更精确地估计物体边界[14], [21], [52]。第二种方法则采用超像素表示，实质上将定位任务交由低层分割方法处理[50]。

我们探索了一种基于深度卷积神经网络（DCNNs）识别能力与全连接条件随机场（CRFs）细粒度定位精度相结合的新方向，并证明其在应对定位挑战方面成效显著，不仅能生成精确的语义分割结果，还能恢复物体边界的细节层次，其精细程度远超现有方法。自我们工作的初版[38]发表以来，后续多篇研究[17]、[40]、[58]、[59]、[60]、[61]、[62]、[63]、[65]已对此方向进行了拓展。

传统上，条件随机场（CRFs）被用于平滑噪声分割图[23]，[31]。这类模型通常耦合相邻节点，倾向于为空间上邻近的像素分配相同标签。从性质上看，这些短程CRFs的主要功能是清理基于局部手工设计特征构建的弱分类器所产生的虚假预测。

与这些较弱的分类器相比，现代深度卷积神经网络（DCNN）架构（例如我们在本工作中使用的架构）生成的得分图和语义标签预测在性质上有所不同。如图5所示，得分图通常相当平滑，并产生同质的分类结果。在这种情况下，使用短程条件随机场（CRF）可能适得其反，因为我们的目标应是恢复详细的局部结构，而非进一步平滑它。在局部范围CRF中结合使用对比敏感势能[23]可能改善定位效果，但仍可能忽略细薄结构，并且通常需要解决计算成本高昂的离散优化问题。

为了克服短程条件随机场的这些限制，我们将[22]中的全连接条件随机场模型集成到我们的系统中。该模型采用能量函数

$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \tag{2}$$

其中$\boldsymbol{x}$是像素的标签分配。我们使用$\theta_i(x_i) = -\log P(x_i)$作为一元势能，其中$P(x_i)$是由DCNN计算的像素$i$处的标签分配概率。

成对势函数的形式允许在完全连接图的情况下进行高效推断，即当连接所有图像像素对时，{v*}。具体而言，如[22]所述，我们采用以下表达式：

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j)\left[w_1 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}\right) + w_2 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}\right)\right] \tag{3}$$

其中，当$x_i \neq x_j$时$\mu(x_i, x_j) =$为1，否则为零，这意味着与Potts模型一样，只有标签不同的节点会受到惩罚。其余表达式使用了两个不同特征空间中的高斯核；第一个“双边”核同时依赖于像素位置（记为p）和RGB颜色（记为I），而第二个核仅依赖于像素位置。超参数σ_α、σ_β和σ_γ控制着高斯核的尺度。第一个核强制颜色和位置相似的像素具有相似的标签，而第二个核在施加平滑性时仅考虑空间邻近性。

关键的是，该模型适用于高效的近似概率推断[22]。在完全可分解的平均场近似{v*}下，消息传递更新可表示为双边空间中的高斯卷积。高维滤波算法[84]显著加速了这一计算过程，使得算法在实际应用中非常快速——使用公开可用的[22]实现，处理Pascal VOC图像平均仅需不到0.5秒。

4 实验结果

我们按照[14]的方法，以直接的方式对ImageNet预训练的VGG-16或ResNet-101网络的模型权重进行微调，使其适应语义分割任务。我们将最后一层的1000路ImageNet分类器替换为与任务语义类别数量（包括背景，如适用）相同目标数的分类器。损失函数是CNN输出图中每个空间位置（相对于原始图像下采样8倍）的交叉熵项之和。在整体损失函数中，所有位置和标签均被平等加权（未标记像素除外）。我们的目标是真实标签（下采样8倍）。我们通过[2]的标准SGD程序优化所有网络层权重的目标函数。我们将DCNN和CRF训练阶段解耦，假设在设置CRF参数时DCNN的一元项是固定的。

我们在四个具有挑战性的数据集上评估了所提出的模型：PASCAL VOC 2012、PASCAL-Context、PASCAL-Person-Part和 Cityscapes。我们首先报告了会议版本[38]在 PASCAL VOC 2012 上的主要结果，随后展示了所有数据集上的最新结果。

4.1 PASCAL VOC 2012

数据集：PASCAL VOC 2012分割基准[34]包含20个前景对象类别和一个背景类别。原始数据集包含1, 464 (*train*),

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

TABLE 1: Effect of Field-Of-View by adjusting the kernel size and atrous sampling rate r at 'fc6' layer. 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.

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

TABLE 2: PASCAL VOC 2012 *val* set results (%) (before CRF) as different learning hyper parameters vary. Employing "poly" learning policy is more effective than "step" when training DeepLab-LargeFOV.

$1,449$ (*val*), and $1,456$ (*test*) pixel-level labeled images for training, validation, and testing, respectively. The dataset is augmented by the extra annotations provided by [85], resulting in $10,582$ (*trainaug*) training images. The performance is measured in terms of pixel intersection-over-union (IOU) averaged across the 21 classes.

4.1.1 Results from our conference version

We employ the VGG-16 network pre-trained on Imagenet, adapted for semantic segmentation as described in Section 3.1. We use a mini-batch of 20 images and initial learning rate of 0.001 (0.01 for the final classifier layer), multiplying the learning rate by 0.1 every 2000 iterations. We use momentum of 0.9 and weight decay of 0.0005.

After the DCNN has been fine-tuned on *trainaug*, we cross-validate the CRF parameters along the lines of [22]. We use default values of $w_2 = 3$ and $\sigma_\gamma = 3$ and we search for the best values of w_1, σ_α, and σ_β by cross-validation on 100 images from *val*. We employ a coarse-to-fine search scheme. The initial search range of the parameters are $w_1 \in [3 : 6]$, $\sigma_\alpha \in [30 : 10 : 100]$ and $\sigma_\beta \in [3 : 6]$ (MATLAB notation), and then we refine the search step sizes around the first round's best values. We employ 10 mean field iterations.

Field of View and CRF: In Tab. 1, we report experiments with DeepLab model variants that use different field-of-view sizes, obtained by adjusting the kernel size and atrous sampling rate r in the 'fc6' layer, as described in Sec. 3.1. We start with a direct adaptation of VGG-16 net, using the original 7×7 kernel size and $r = 4$ (since we use no stride for the last two max-pooling layers). This model yields performance of 67.64% after CRF, but is relatively slow (1.44 images per second during training). We have improved model speed to 2.9 images per second by reducing the kernel size to 4×4. We have experimented with two such network variants with smaller ($r = 4$) and larger ($r = 8$) FOV sizes; the latter one performs better. Finally, we employ kernel size 3×3 and even larger atrous sampling rate ($r = 12$), also making the network thinner by retaining a random subset of 1,024 out of the 4,096 filters in layers 'fc6' and 'fc7'. The resulting model, DeepLab-CRF-LargeFOV, matches the performance of the direct VGG-16 adaptation (7×7 kernel size, $r = 4$). At the same time, DeepLab-LargeFOV is 3.36 times faster and has significantly fewer parameters (20.5M instead of 134.3M).

The CRF substantially boosts performance of all model variants, offering a 3-5% absolute increase in mean IOU.

Test set evaluation: We have evaluated our DeepLab-CRF-LargeFOV model on the PASCAL VOC 2012 official *test* set. It achieves 70.3% mean IOU performance.

4.1.2 Improvements after conference version of this work

After the conference version of this work [38], we have pursued three main improvements of our model, which we discuss below: (1) different learning policy during training, (2) atrous spatial pyramid pooling, and (3) employment of deeper networks and multi-scale processing.

Learning rate policy: We have explored different learning rate policies when training DeepLab-LargeFOV. Similar to [86], we also found that employing a "poly" learning rate policy (the learning rate is multiplied by $(1-\frac{iter}{max\_iter})^{power}$) is more effective than "step" learning rate (reduce the learning rate at a fixed step size). As shown in Tab. 2, employing "poly" (with $power = 0.9$) and using the same batch size and same training iterations yields 1.17% better performance than employing "step" policy. Fixing the batch size and increasing the training iteration to 10K improves the performance to 64.90% (1.48% gain); however, the total training time increases due to more training iterations. We then reduce the batch size to 10 and found that comparable performance is still maintained (64.90% *vs*. 64.71%). In the end, we employ batch size = 10 and 20K iterations in order to maintain similar training time as previous "step" policy. Surprisingly, this gives us the performance of 65.88% (3.63% improvement over "step") on *val*, and 67.7% on *test*, compared to 65.1% of the original "step" setting for DeepLab-LargeFOV before CRF. We employ the "poly" learning rate policy for all experiments reported in the rest of the paper.

Atrous Spatial Pyramid Pooling: We have experimented with the proposed Atrous Spatial Pyramid Pooling (ASPP) scheme, described in Sec. 3.1. As shown in Fig. 7, ASPP for VGG-16 employs several parallel fc6-fc7-fc8 branches. They all use 3×3 kernels but different atrous rates r in the 'fc6' in order to capture objects of different size. In Tab. 3, we report results with several settings: (1) Our baseline LargeFOV model, having a single branch with $r = 12$, (2) ASPP-S, with four branches and smaller atrous rates (r = {2, 4, 8, 12}), and (3) ASPP-L, with four branches and larger rates (r = {6, 12, 18, 24}). For each variant we report results before and after CRF. As shown in the table, ASPP-S yields 1.22% improvement over the baseline LargeFOV before CRF. However, after CRF both LargeFOV and ASPP-S perform similarly. On the other hand, ASPP-L yields consistent improvements over the baseline LargeFOV both before and after CRF. We evaluate on *test* the proposed ASPP-L + CRF model, attaining 72.6%. We visualize the effect of the different schemes in Fig. 8.

Deeper Networks and Multiscale Processing: We have experimented building DeepLab around the recently pro-

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

表1：通过调整卷积核大小和'fc6'层的空洞采样率r来展示视野范围的影响。我们列出了模型参数数量、训练速度（图像/秒）以及CRF处理前后在*val*数据集上的平均IOU。DeepLab-LargeFOV（卷积核大小3×3，r =为12）实现了最佳平衡。

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

表2：PASCAL VOC 2012 *val* 数据集结果（%）（CRF处理前）随不同学习超参数变化的情况。在训练DeepLab-LargeFOV时，采用“poly”学习策略比“step”更有效。

1, 449 (*val*)、1, 456 (*test*)张像素级标注图像分别用于训练、验证和测试。该数据集通过[85]提供的额外标注进行了增强，最终得到10, 582 (*trainaug*)张训练图像。性能评估采用像素交并比（IOU）作为指标，并计算21个类别的平均值。

4.1.1 Results from our conference version

我们采用在Imagenet上预训练的VGG-16网络，并根据第3.1节所述的方法将其调整为语义分割任务。我们使用20张图像的小批量数据，初始学习率为0.001（最终分类器层为0.01），每2000次迭代将学习率乘以0.1。我们使用的动量为0.9，权重衰减为0.0005。

在DCNN于*trainaug*上完成微调后，我们参照[22]的方法对CRF参数进行交叉验证。我们采用默认值$w_2 = 3$和$\sigma_\gamma = 3$，并通过在*val*的100张图像上进行交叉验证来寻找w_1、σ_α和σ_β的最佳参数值。我们采用由粗到精的搜索策略：参数的初始搜索范围为$w_1 \in$ [3:6]、$\sigma_\alpha \in$ [30:10:100]和$\sigma_\beta \in$ [3:6]（MATLAB表示法），随后围绕第一轮得到的最佳值缩小搜索步长。我们使用10次平均场迭代进行计算。

视野与CRF：在表1中，我们报告了使用不同视野大小的DeepLab模型变体的实验，这些视野大小通过调整'fc6'层中的卷积核大小和空洞采样率r获得，如第3.1节所述。我们首先直接改编VGG-16网络，使用原始的7×7卷积核大小和r =4的空洞采样率（因为我们在最后两个最大池化层中未使用步长）。该模型在CRF后达到67.64%的性能，但速度相对较慢（训练时每秒处理1.44张图像）。通过将卷积核大小减小到4×4，我们将模型速度提升至每秒2.9张图像。我们尝试了两种具有较小（r =4）和较大（r =8）视野大小的网络变体；后者表现更优。最后，我们采用3×3的卷积核和更大的空洞采样率（r =12），同时通过保留'fc6'和'fc7'层中4096个滤波器中的1024个随机子集使网络更精简。最终模型DeepLab-CRF-LargeFOV达到了直接改编VGG-16（7×7卷积核，r =4空洞采样率）的性能水平。与此同时，DeepLab-LargeFOV的速度提高了3.36倍，且参数数量显著减少（从134.3M降至20.5M）。

CRF显著提升了所有模型变体的性能，使平均IOU实现了3-5%的绝对增长。

测试集评估：我们在PASCAL VOC 2012官方*test*集上评估了我们的DeepLab-CRF-LargeFOV模型，其取得了70.3%的平均IOU性能。

4.1.2 Improvements after conference version of this work

在本工作的会议版本[38]之后，我们对模型进行了三项主要改进，具体讨论如下：(1) 训练期间采用不同的学习策略，(2) 使用空洞空间金字塔池化，以及(3) 采用更深层网络与多尺度处理。

学习率策略：在训练DeepLab-LargeFOV时，我们探索了不同的学习率策略。与[86]类似，我们也发现采用“poly”学习率策略（学习率乘以$(1-\frac{iter}{max_iter})^{power}$）比“step”学习率策略（以固定步长降低学习率）更有效。如表2所示，采用“poly”策略（设置$power$ =为0.9）并使用相同的批次大小和训练迭代次数，性能比“step”策略提高了1.17%。固定批次大小并将训练迭代次数增加到10K，性能提升至64.90%（增益1.48%）；然而，由于训练迭代次数增加，总训练时间也随之增长。随后，我们将批次大小减少到10，发现仍能保持相当的性能（64.90% *vs* 64.71%）。最终，我们采用批次大小=为10和20K迭代次数，以保持与之前“step”策略相似的训练时间。令人惊讶的是，这在*val*上实现了65.88%的性能（比“step”策略提升3.63%），在*test*上达到67.7%，而CRF之前的DeepLab-LargeFOV原始“step”设置仅为65.1%。在本文后续报告的所有实验中，我们均采用了“poly”学习率策略。

空洞空间金字塔池化：我们尝试了第3.1节中提出的空洞空间金字塔池化（ASPP）方案。如图7所示，基于VGG-16的ASPP采用多个并行的fc6-fc7-fc8分支。所有分支均使用3×3卷积核，但在'fc6'层采用不同的空洞率r以捕捉不同尺寸的目标。表3展示了多种配置的结果：(1) 基线LargeFOV模型（单分支，空洞率r =12），(2) ASPP-S（四分支，较小空洞率r = {2,4,8,12}），(3) ASPP-L（四分支，较大空洞率r = {6,12,18,24}）。每个变体均汇报了CRF处理前后的结果。如表所示，在CRF处理前，ASPP-S较基线LargeFOV提升1.22%；但经过CRF后，两者性能趋于接近。而ASPP-L在CRF处理前后均持续优于基线LargeFOV。我们评估了*test*提出的ASPP-L+ CRF模型，获得72.6%的精度。图8直观展示了不同方案的效果对比。

更深的网络与多尺度处理：我们尝试围绕最近提出的

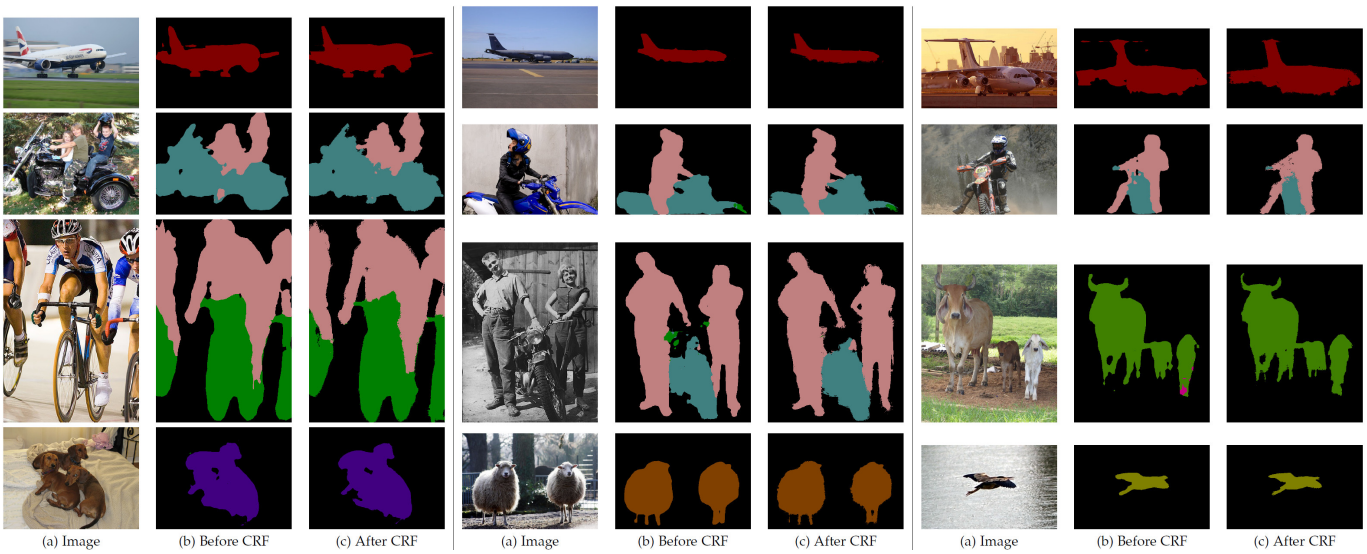
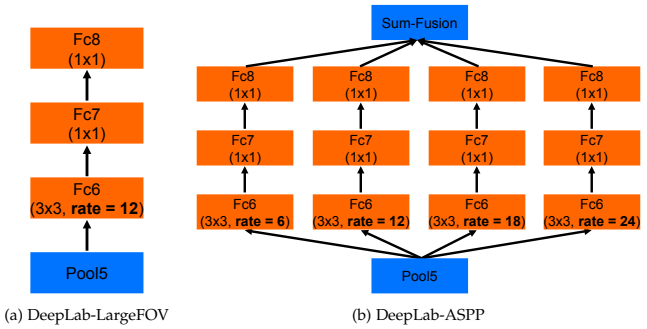
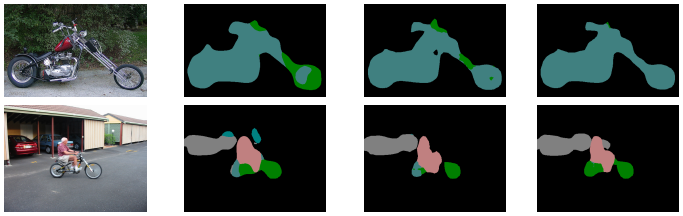


Fig. 6: PASCAL VOC 2012 *val* results. Input image and our DeepLab results before/after CRF.



(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.

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

TABLE 3: Effect of ASPP on PASCAL VOC 2012 *val* set performance (mean IOU) for VGG-16 based DeepLab model. **LargeFOV**: single branch, $r = 12$. **ASPP-S**: four branches, r = {2, 4, 8, 12}. **ASPP-L**: four branches, r = {6, 12, 18, 24}.

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

TABLE 4: Employing ResNet-101 for DeepLab on PASCAL VOC 2012 *val* set. **MSC**: Employing mutli-scale inputs with max fusion. **COCO**: Models pretrained on MS-COCO. **Aug**: Data augmentation by randomly rescaling inputs.

posed residual net ResNet-101 [11] instead of VGG-16. Similar to what we did for VGG-16 net, we re-purpose ResNet-101 by atrous convolution, as described in Sec. 3.1. On top of that, we adopt several other features, following recent work of [17], [18], [39], [40], [58], [59], [62]: (1) Multi-scale inputs: We separately feed to the DCNN images at scale = {0.5, 0.75, 1}, fusing their score maps by taking the maximum response across scales for each position separately [17]. (2) Models pretrained on MS-COCO [87]. (3) Data augmentation by randomly scaling the input images (from 0.5 to 1.5) during training. In Tab. 4, we evaluate how each of these factors, along with LargeFOV and atrous spatial pyramid pooling (ASPP), affects *val* set performance. Adopting ResNet-101 instead of VGG-16 significantly improves DeepLab performance (*e.g.*, our simplest ResNet-101 based model attains 68.72%, compared to 65.76% of our DeepLab-LargeFOV VGG-16 based variant, both before CRF). Multiscale fusion [17] brings extra 2.55% improvement, while pretraining the model on MS-COCO gives another 2.01% gain. Data augmentation during training is effective (about 1.6% improvement). Employing LargeFOV (adding an atrous convolutional layer on top of ResNet, with 3×3 kernel and rate = 12) is beneficial (about 0.6% improvement). Further 0.8% improvement is achieved by atrous spatial pyramid pooling (ASPP). Post-processing our best model by dense CRF yields

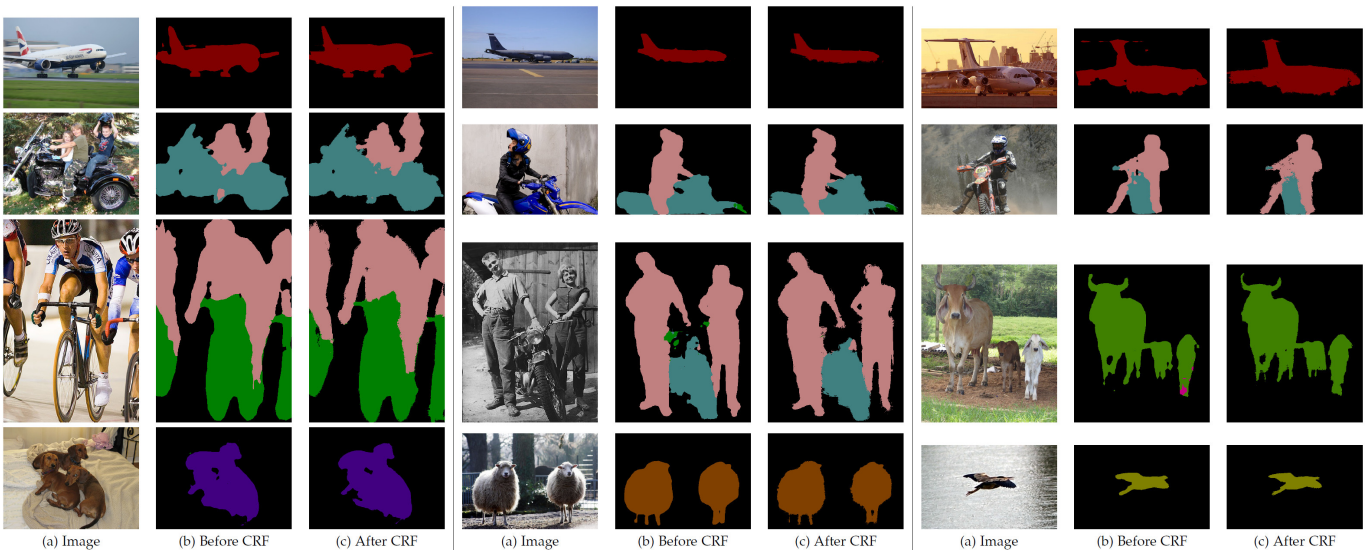


图6：PASCAL VO C 2012 *val* 结果。输入图像及我们的DeepLab结果 b CRF 前/后。

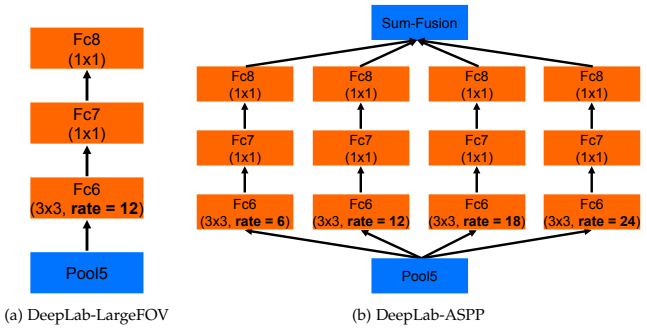


图7：DeepLab-ASPP采用多个具有不同速率的滤波器，以在多尺度上捕捉目标与上下文信息。

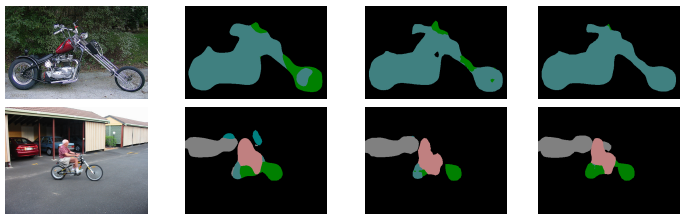


图8：采用ASPP与基线LargeFOV模型的定性分割结果对比。使用多尺度*large*视野的ASPP-L模型，能成功捕捉多尺度下的物体及图像上下文信息。

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

表3：ASPP对基于VGG-16的DeepLab模型在PASCAL VOC 2012 *val*数据集上性能（平均IOU）的影响。LargeFOV：单分支，r =12。ASPP-S：四分支，r= {2、4、8、12}。ASPP-L：四分支，r = {6、12、18、24}。

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

表4：在PASCAL VOC 2012 *val* 数据集上使用ResNet-101进行DeepLab。MSC：采用多尺度输入与最大融合策略。COCO：基于MS-COCO预训练的模型。Aug：通过随机缩放输入实现数据增强。

提出的残差网络ResNet-101[11]而非VGG-16。

类似于我们对VGG-16网络所做的操作，我们通过空洞卷积重新调整ResNet-101的用途，如第3.1节所述。在此基础上，我们借鉴了近期研究[17]、[18]、[39]、[40]、[58]、[59]、[62]的若干特性：(1) 多尺度输入：我们将尺度为= {0.5、0.75、1}的图像分别输入DCNN，通过逐位置取跨尺度最大响应值融合其得分图[17]。(2) 使用在MS-COCO[87]上预训练的模型。(3) 训练期间通过随机缩放输入图像（从0.5到1.5）进行数据增强。在表4中，我们评估了这些因素与LargeFOV及空洞空间金字塔池化（ASPP）对验证集性能的影响。采用ResNet-101替代VGG-16显著提升了DeepLab性能（*e.g*例如，我们最简单的基于ResNet-101的模型达到68.72%，而基于VGG-16的DeepLab-LargeFOV变体为65.76%，两者均未使用CRF）。多尺度融合[17]带来额外2.55%的提升，在MS-COCO上预训练模型再带来2.01%增益。训练期间的数据增强效果显著（约提升1.6%）。采用LargeFOV（在ResNet顶部添加3×3卷积核、膨胀率=12的空洞卷积层）具有积极作用（约提升0.6%）。通过空洞空间金字塔池化（ASPP）可进一步获得0.8%提升。使用密集CRF对我们的最佳模型进行后处理最终实现

performance of 77.69%.

Qualitative results: We provide qualitative visual comparisons of DeepLab's results (our best model variant) before and after CRF in Fig. 6. The visualization results obtained by DeepLab before CRF already yields excellent segmentation results, while employing the CRF further improves the performance by removing false positives and refining object boundaries.

Test set results: We have submitted the result of our final best model to the official server, obtaining *test* set performance of 79.7%, as shown in Tab. 5. The model substantially outperforms previous DeepLab variants (*e.g.*, DeepLab-LargeFOV with VGG-16 net) and is currently the top performing method on the PASCAL VOC 2012 segmentation leaderboard.

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

TABLE 5: Performance on PASCAL VOC 2012 *test* set. We have added some results from recent arXiv papers on top of the official leadearboard results.

VGG-16 *vs.* ResNet-101: We have observed that DeepLab based on ResNet-101 [11] delivers better segmentation results along object boundaries than employing VGG-16 [4], as visualized in Fig. 9. We think the identity mapping [94] of ResNet-101 has similar effect as hyper-column features [21], which exploits the features from the intermediate layers to better localize boundaries. We further quantize this effect in Fig. 10 within the "trimap" [22], [31] (a narrow band along object boundaries). As shown in the figure, employing ResNet-101 before CRF has almost the same accuracy along object boundaries as employing VGG-16 in conjunction with a CRF. Post-processing the ResNet-101 result with a CRF further improves the segmentation result.

4.2 PASCAL-Context

Dataset: The PASCAL-Context dataset [35] provides detailed semantic labels for the whole scene, including both object (*e.g.*, person) and stuff (*e.g.*, sky). Following [35], the proposed models are evaluated on the most frequent 59 classes along with one background category. The training set and validation set contain 4998 and 5105 images.

Evaluation: We report the evaluation results in Tab. 6. Our VGG-16 based LargeFOV variant yields 37.6% before and 39.6% after CRF. Repurposing the ResNet-101 [11] for

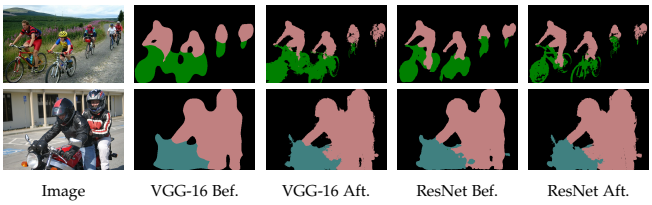


Fig. 9: DeepLab results based on VGG-16 net or ResNet-101 before and after CRF. The CRF is critical for accurate prediction along object boundaries with VGG-16, whereas ResNet-101 has acceptable performance even before CRF.

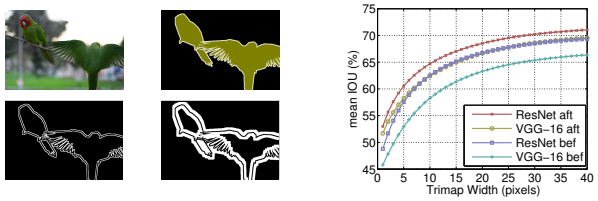


Fig. 10: (a) Trimap examples (top-left: image. top-right: ground-truth. bottom-left: trimap of 2 pixels. bottom-right: trimap of 10 pixels). (b) Pixel mean IOU as a function of the band width around the object boundaries when employing VGG-16 or ResNet-101 before and after CRF.

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

TABLE 6: Comparison with other state-of-art methods on PASCAL-Context dataset.

DeepLab improves 2% over the VGG-16 LargeFOV. Similar to [17], employing multi-scale inputs and max-pooling to merge the results improves the performance to 41.4%. Pretraining the model on MS-COCO brings extra 1.5% improvement. Employing atrous spatial pyramid pooling is more effective than LargeFOV. After further employing dense CRF as post processing, our final model yields 45.7%, outperforming the current state-of-art method [40] by 2.4% without using their non-linear pairwise term. Our final model is slightly better than the concurrent work [93] by 1.2%, which also employs atrous convolution to repurpose

性能为77.69%。

定性结果：我们在图6中提供了DeepLab（我们的最佳模型变体）在CRF前后结果的定性视觉比较。DeepLab在CRF之前获得的可视化结果已经产生了出色的分割效果，而采用CRF则通过减少误报和细化物体边界进一步提升了性能。

测试集结果：我们已将最终最佳模型的结果提交至官方服务器，在*test*集上获得了79.7%的性能表现，如表5所示。该模型显著超越了之前的DeepLab变体（*e.g*，如采用VGG-16网络的DeepLab-LargeFOV），目前是PASCAL VOC 2012分割排行榜上性能最优的方法。

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

表5：在PASCAL VOC 2012 *test* 数据集上的性能表现。我们在官方排行榜结果的基础上，补充了近期arXiv论文中的一些结果。

VGG-16 *vs*。ResNet-101：我们观察到，基于ResNet-101 [11] 的DeepLab在物体边界处能提供比使用VGG-16 [4] 更好的分割结果，如图9所示。我们认为ResNet-101的恒等映射 [94] 与超列特征 [21] 具有相似的效果，后者利用中间层的特征来更好地定位边界。我们在图10的“trimap” [22], [31]（物体边界附近的窄带区域）中进一步量化了这种效果。如图所示，在CRF之前使用ResNet-101，其在物体边界处的准确度与使用VGG-16并结合CRF几乎相同。对ResNet-101的结果进行CRF后处理，可进一步提升分割效果。

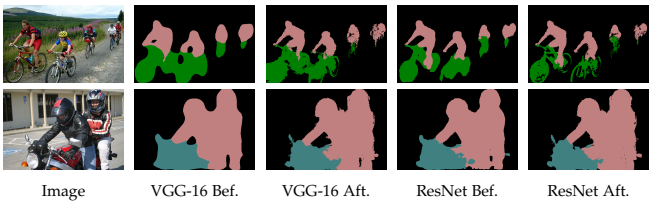
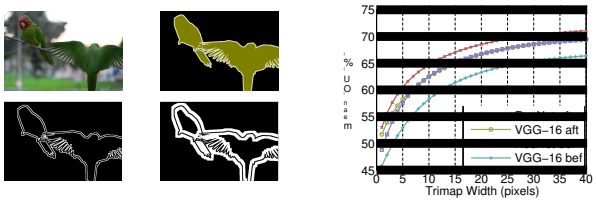


图9：基于VGG-16网络或ResNet-101网络在CRF处理前后的DeepLab结果。CRF对于VGG-16沿物体边界的精确预测至关重要，而ResNet-101即使在CRF处理前也已具备可接受的性能。



(a) (b)

图10：(a) Trimap示例（左上：图像。右上：真实标注。左下：2像素宽度的trimap。右下：10像素宽度的trimap）。(b) 在使用CRF前后采用VGG-16或ResNet-101时，像素平均IOU随物体边界周围带宽变化的情况。

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

表6：在PASCAL-Context数据集上与其他先进方法的比较。

4.2 PASCAL-上下文

数据集：PASCAL-Context数据集[35]为整个场景提供了详细的语义标注，包括物体（*e.g*，如人物）和背景物（*e.g*，如天空）。依照[35]的方法，所提出的模型在59个最常见类别及一个背景类别上进行评估。训练集和验证集分别包含4998张和5105张图像。

评估：我们在表6中报告了评估结果。我们基于VGG-16的LargeFOV变体在CRF处理前达到37.6%，处理后达到39.6%。将ResNet-101 [11]重新用于DeepLab相较于VGG-16 LargeFOV提升了2%。与[17]类似，采用多尺度输入和最大池化融合结果后，性能提升至41.4%。在MS-COCO上预训练模型额外带来1.5%的改进。采用空洞空间金字塔池化比LargeFOV更有效。进一步使用密集CRF进行后处理后，我们的最终模型达到45.7%，在不使用非线性配对项的情况下超越了当前最佳方法[40] 2.4%。我们的最终模型比同样采用空洞卷积的同期工作[93]略优1.2%。

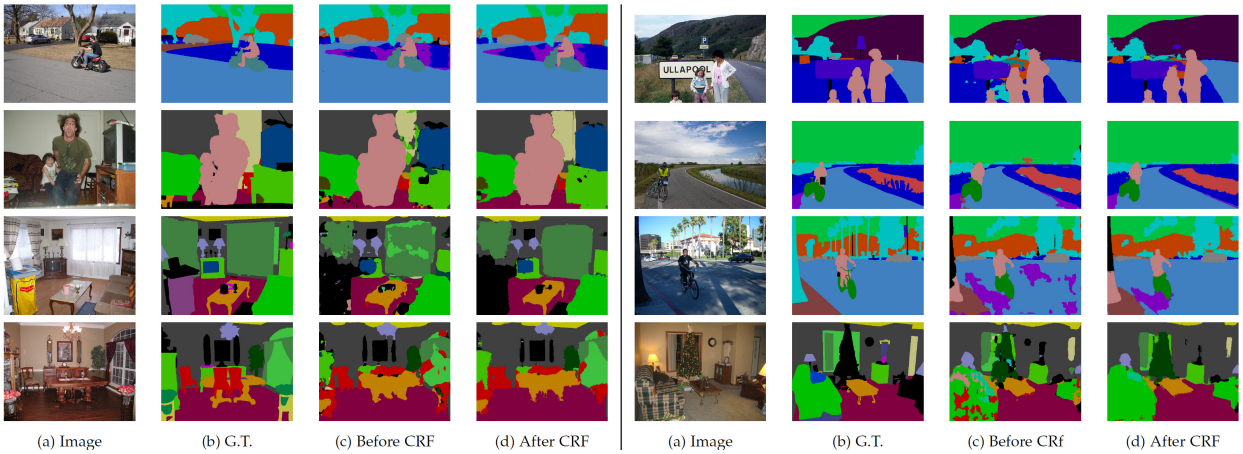


Fig. 11: PASCAL-Context results. Input image, ground-truth, and our DeepLab results before/after CRF.

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

TABLE 7: Comparison with other state-of-art methods on PASCAL-Person-Part dataset.

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 8: Test set results on the Cityscapes dataset, comparing our DeepLab system with other state-of-art methods.

the residual net of [11] for semantic segmentation.

Qualitative results: We visualize the segmentation results of our best model with and without CRF as post processing in Fig. 11. DeepLab before CRF can already predict most of the object/stuff with high accuracy. Employing CRF, our model is able to further remove isolated false positives and improve the prediction along object/stuff boundaries.

4.3 PASCAL-Person-Part

Dataset: We further perform experiments on semantic part segmentation [98], [99], using the extra PASCAL VOC 2010 annotations by [36]. We focus on the *person* part for the dataset, which contains more training data and large variation in object scale and human pose. Specifically, the dataset contains detailed part annotations for every person, *e.g.* eyes, nose. We merge the annotations to be Head, Torso, Upper/Lower Arms and Upper/Lower Legs, resulting in six person part classes and one background class. We only use those images containing persons for training (1716 images) and validation (1817 images).

Evaluation: The human part segmentation results on PASCAL-Person-Part is reported in Tab. 7. [17] has already conducted experiments on this dataset with re-purposed VGG-16 net for DeepLab, attaining 56.39% (with multi-scale inputs). Therefore, in this part, we mainly focus on the effect of repurposing ResNet-101 for DeepLab. With ResNet-101, DeepLab alone yields 58.9%, significantly outperforming DeepLab-LargeFOV (VGG-16 net) and DeepLab-Attention (VGG-16 net) by about 7% and 2.5%, respectively. Incorporating multi-scale inputs and fusion by max-pooling further improves performance to 63.1%. Additionally pretraining the model on MS-COCO yields another 1.3% improvement. However, we do not observe any improvement when adopting either LargeFOV or ASPP on this dataset. Employing the dense CRF to post process our final output substantially outperforms the concurrent work [97] by 4.78%.

Qualitative results: We visualize the results in Fig. 12.

4.4 Cityscapes

Dataset: Cityscapes [37] is a recently released large-scale dataset, which contains high quality pixel-level annotations of 5000 images collected in street scenes from 50 different cities. Following the evaluation protocol [37], 19 semantic labels (belonging to 7 super categories: ground, construction, object, nature, sky, human, and vehicle) are used for evaluation (the void label is not considered for evaluation). The training, validation, and test sets contain 2975, 500, and 1525 images respectively.

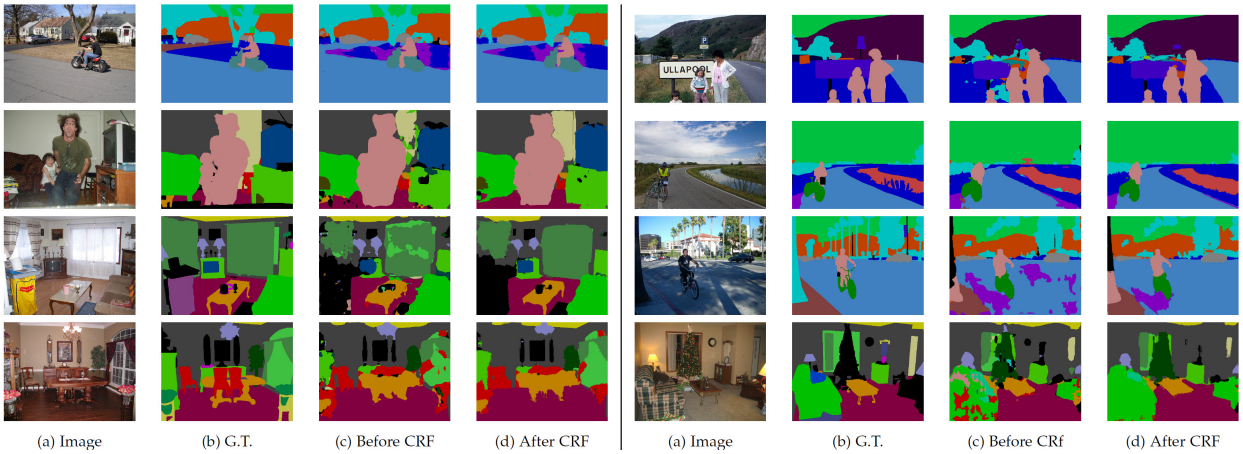


图11：PASCAL-Context数据集结果。输入图像、真实标注以及我们DeepLab模型在CRF处理前/后的结果。

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

表7：在PASCAL人体部位数据集上与其他先进方法的比较。

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

表8：Cityscapes数据集上的测试集结果，将我们的DeepLab系统与其他最先进方法进行比较。

[11]中用于语义分割的残差网络。

定性结果：我们在图11中可视化展示了我们最佳模型在使用和不使用CRF作为后处理时的分割结果。在CRF处理前，DeepLab已能高精度预测大部分物体/背景。通过采用CRF，我们的模型能够进一步消除孤立的误检，并提升物体/背景边界处的预测精度。

4.3 PASCAL-人体部位

数据集：我们进一步在语义部分分割[98]、[99]上进行了实验，使用了[36]提供的额外PASCAL VOC 2010标注。我们专注于该数据集的*person*部分，这部分包含更多的训练数据，且在物体尺度和人体姿态上具有更大的变化。具体来说，该数据集为每个人提供了详细的部分标注，*e.g.*。眼睛、鼻子。我们将标注合并为头部、躯干、上/下臂和上/下腿，从而得到六个身体部分类别和一个背景类别。我们仅使用包含人物的图像进行训练（1716张图像）和验证（1817张图像）。

评估：在PASCAL-Person-Part数据集上的人体部位分割结果如表7所示。[17]已在此数据集上使用为DeepLab改造的VGG-16网络进行过实验，取得了56.39%的准确率（采用多尺度输入）。因此，在本部分中，我们主要关注为DeepLab改造ResNet-101的效果。使用ResNet-101时，仅使用DeepLab就达到了58.9%的准确率，显著优于DeepLab-LargeFOV（VGG-16网络）和DeepLab-Attention（VGG-16网络），分别领先约7%和2.5%。通过最大池化融合多尺度输入后，性能进一步提升至63.1%。在MS-COCO数据集上进行额外预训练又带来了1.3%的提升。然而，在该数据集上采用LargeFOV或ASPP并未观察到任何改进。使用密集CRF对我们的最终输出进行后处理，结果比同期研究[97]大幅领先4.78%。

定性结果：我们在图12中可视化结果。

4.4 城市场景

数据集：Cityscapes [37] 是近期发布的大规模数据集，包含从50个不同城市街景中采集的5000张图像的高质量像素级标注。按照评估协议[37]，采用19个语义标签（属于7个超类别：地面、建筑、物体、自然、天空、人、车辆）进行评估（无效标签不计入评估）。训练集、验证集和测试集分别包含2975张、500张和1525张图像。

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

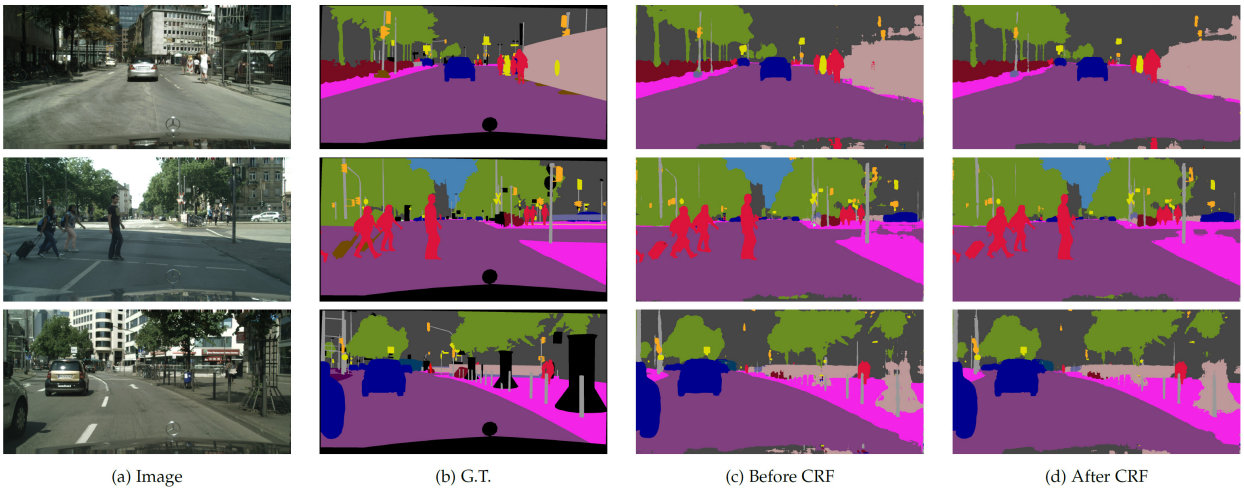


Fig. 13: Cityscapes results. Input image, ground-truth, and our DeepLab results before/after CRF.

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

TABLE 9: Val set results on Cityscapes dataset. **Full**: model trained with full resolution images.

Test set results of pre-release: We have participated in benchmarking the Cityscapes dataset pre-release. As shown in the top of Tab. 8, our model attained third place, with performance of 63.1% and 64.8% (with training on additional coarsely annotated images).

Val set results: After the initial release, we further explored the validation set in Tab. 9. The images of Cityscapes have resolution 2048×1024, making it a challenging problem to train deeper networks with limited GPU memory. During benchmarking the pre-release of the dataset, we downsampled the images by 2. However, we have found that it is beneficial to process the images in their original resolution. With the same training protocol, using images of original resolution significantly brings 1.9% and 1.8% improvements before and after CRF, respectively. In order to perform inference on this dataset with high resolution images, we split each image into overlapped regions, similar to [37]. We have also replaced the VGG-16 net with ResNet-101. We do not exploit multi-scale inputs due to the limited GPU memories at hand. Instead, we only explore (1) deeper networks (*i.e.*, ResNet-101), (2) data augmentation, (3) LargeFOV or ASPP, and (4) CRF as post processing on this dataset. We first find that employing ResNet-101 alone is better than using VGG-16 net. Employing LargeFOV brings 2.6% improvement and using ASPP further improves results by 1.2%. Adopting data augmentation and CRF as post processing brings another 0.6% and 0.4%, respectively.

图12：PASCAL人体部位分割结果。输入图像、真实标注以及我们DeepLab模型在CRF处理前后的结果。

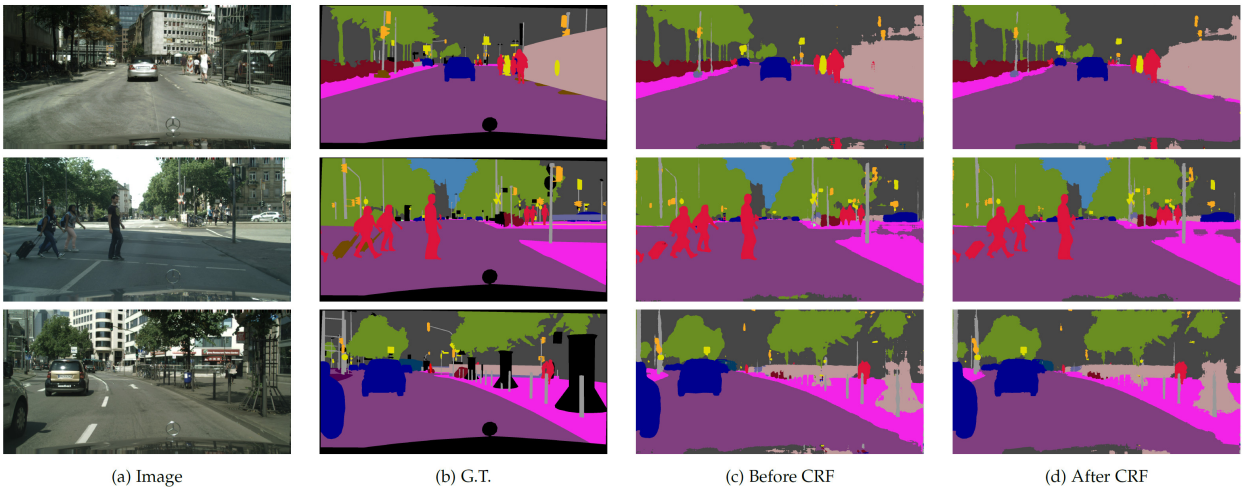


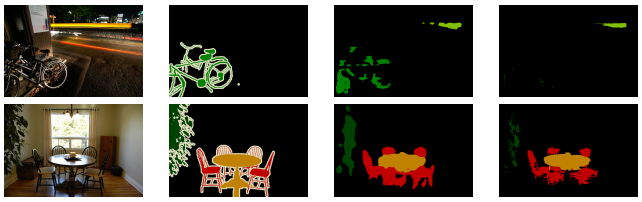
图13：Cityscapes结果。输入图像、真实标注以及我们DeepLab在CRF处理前/后的结果。

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

表9：Cityscapes数据集上的验证集结果。完整版：使用全分辨率图像训练的模型。

预发布版本测试集结果：我们参与了Cityscapes数据集的预发布基准测试。如表8顶部所示，我们的模型取得了第三名的成绩，在使用额外粗标注图像训练的情况下，性能分别达到63.1%和64.8%。

验证集结果：在初始发布后，我们进一步扩我们在表9中探讨了验证集。Cityscapes数据集的图像分辨率为2048×1024，这使得在有限的GPU内存下训练更深层的网络成为一个具有挑战性的问题。在数据集预发布版本的基准测试期间，我们将图像下采样了2倍。然而，我们发现以原始分辨率处理图像是有益的。采用相同的训练协议，使用原始分辨率的图像在CRF处理前后分别显著带来了1.9%和1.8%的提升。为了在这个数据集上对高分辨率图像进行推理，我们将每张图像分割成重叠的区域，类似于[37]的方法。我们还用ResNet-101替换了VGG-16网络。由于手头GPU内存有限，我们没有利用多尺度输入。相反，我们仅在此数据集上探索了（1）更深层的网络（*i.e*即ResNet-101）、（2）数据增强、（3）LargeFOV或ASPP，以及（4）作为后处理的CRF。我们首先发现，单独使用ResNet-101就比使用VGG-16网络效果更好。采用LargeFOV带来了2.6%的提升，而使用ASPP则进一步将结果提高了1.2%。采用数据增强和CRF作为后处理分别带来了额外的0.6%和0.4%的提升。



(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.

Current test result: We have uploaded our best model to the evaluation server, obtaining performance of 70.4%. Note that our model is only trained on the train set.

Qualitative results: We visualize the results in Fig. 13.

4.5 Failure Modes

We further qualitatively analyze some failure modes of our best model variant on PASCAL VOC 2012 *val* set. As shown in Fig. 14, our proposed model fails to capture the delicate boundaries of objects, such as bicycle and chair. The details could not even be recovered by the CRF post processing since the unary term is not confident enough. We hypothesize the encoder-decoder structure of [100], [102] may alleviate the problem by exploiting the high resolution feature maps in the decoder path. How to efficiently incorporate the method is left as a future work.

5 Conclusion

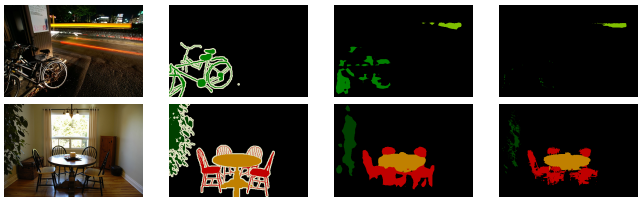
Our proposed "DeepLab" system re-purposes networks trained on image classification to the task of semantic segmentation by applying the 'atrous convolution' with upsampled filters for dense feature extraction. We further extend it to atrous spatial pyramid pooling, which encodes objects as well as image context at multiple scales. To produce semantically accurate predictions and detailed segmentation maps along object boundaries, we also combine ideas from deep convolutional neural networks and fully-connected conditional random fields. Our experimental results show that the proposed method significantly advances the state-of-art in several challenging datasets, including PASCAL VOC 2012 semantic image segmentation benchmark, PASCAL-Context, PASCAL-Person-Part, and Cityscapes datasets.

Acknowledgments

This work was partly supported by the ARO 62250-CS, FP7-RECONFIG, FP7-MOBOT, and H2020-ISUPPORT EU projects. We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. IEEE*, 1998.
[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2013.
[3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv:1312.6229*, 2013.
[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv:1409.4842*, 2014.
[6] G. Papandreou, I. Kokkinos, and P.-A. Savalle, "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection," in *CVPR*, 2015.
[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[8] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *CVPR*, 2014.
[9] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.
[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385*, 2015.
[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," *arXiv:1512.02325*, 2015.
[13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.
[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
[15] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets: Time-Frequency Methods and Phase Space*, 1989, pp. 289–297.
[16] A. Giusti, D. Ciresan, J. Masci, L. Gambardella, and J. Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks," in *ICIP*, 2013.
[17] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *CVPR*, 2016.
[18] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," in *ICLR*, 2016.
[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV*, 2014.
[21] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *CVPR*, 2015.
[22] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *NIPS*, 2011.
[23] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," in *SIGGRAPH*, 2004.
[24] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *IJCV*, 2009.
[25] A. Lucchi, Y. Li, X. Boix, K. Smith, and P. Fua, "Are spatial and global constraints really necessary for segmentation?" in *ICCV*, 2011.
[26] X. He, R. S. Zemel, and M. Carreira-Perpindn, "Multiscale conditional random fields for image labeling," in *CVPR*, 2004.
[27] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr, "Associative hierarchical crfs for object class image segmentation," in *ICCV*, 2009.
[28] V. Lempitsky, A. Vedaldi, and A. Zisserman, "Pylon model for semantic segmentation," in *NIPS*, 2011.
[29] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov, "Fast approximate energy minimization with label costs," *IJCV*, 2012.
[30] J. M. Gonfaus, X. Boix, J. Van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez, "Harmony potentials for joint classification and segmentation," in *CVPR*, 2010.
[31] P. Kohli, P. H. Torr *et al.*, "Robust higher order potentials for enforcing label consistency," *IJCV*, vol. 82, no. 3, pp. 302–324, 2009.
[32] L.-C. Chen, G. Papandreou, and A. Yuille, "Learning a dictionary of shape epitomes with applications to image labeling," in *ICCV*, 2013.



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

图14：失败模式。输入图像、真实标注以及我们DeepLab在CRF处理前/后的结果。

当前测试结果：我们已将最佳模型上传至评估服务器，获得了70.4%的性能表现。请注意，我们的模型仅在训练集上进行过训练。

定性结果：我们在图13中可视化结果。

4.5 失效模式

我们进一步定性分析了我们最佳模型变体在PASCAL VOC 2012 *val*数据集上的一些失败模式。如图14所示，我们提出的模型未能捕捉到物体的精细边界，例如自行车和椅子。由于一元项置信度不足，这些细节甚至无法通过CRF后处理恢复。我们推测[100]、[102]所采用的编码器-解码器结构，可能通过利用解码器路径中的高分辨率特征图来缓解此问题。如何高效地整合该方法将作为未来的研究工作。

5 结论

我们提出的“DeepLab”系统通过采用上采样滤波器的“空洞卷积”进行密集特征提取，将训练用于图像分类的网络重新应用于语义分割任务。我们进一步将其扩展为空洞空间金字塔池化，该结构能够在多尺度下编码对象及图像上下文信息。为了生成语义准确的预测结果并沿物体边界输出精细的分割图，我们还融合了深度卷积神经网络与全连接条件随机场的相关思想。实验结果表明，该方法在多个具有挑战性的数据集上显著推进了现有技术水平，包括PASCAL VOC 2012语义图像分割基准、PASCAL-Context、PASCAL-Person-Part以及Cityscapes数据集。

致谢

本工作部分得到了ARO 62250-CS、FP7-RECONFIG、FP7-MOBOT以及H2020-ISUPPORT欧盟项目的支持。我们衷心感谢英伟达公司为本研究捐赠GPU设备所提供的支持。

参考文献

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “基于梯度的学习应用于文档识别”，发表于 *Proc. IEEE*, 1998年。[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “使用深度卷积神经网络进行ImageNet分类”，发表于 *NIPS*, 2013年。[3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat：使用卷积网络进行集成识别、定位与检测”，*arXiv:1312.6229*, 2013年。

[4] K. Simonyan 与 A. Zisserman, “用于大规模图像识别的极深卷积网络”，发表于 *ICLR*, 2015年。[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke 与 A. Rabinovich, “更深入的卷积”，*arXiv:1409.4842*, 2014年。[6] G. Papandreou, I. Kokkinos 与 P.-A. Savalle, “深度学习中的局部与全局形变建模：缩影卷积、多示例学习与滑动窗口检测”，发表于 *CVPR*, 2015年。[7] R. Girshick, J. Donahue, T. Darrell 与 J. Malik, “用于精确目标检测与语义分割的丰富特征层次结构”，发表于 *CVPR*, 2014年。[8] D. Erhan, C. Szegedy, A. Toshev 与 D. Anguelov, “使用深度神经网络的可扩展目标检测”，发表于 *CVPR*, 2014年。[9] R. Girshick, “快速 R-CNN”，发表于 *ICCV*, 2015年。[10] S. Ren, K. He, R. Girshick 与 J. Sun, “更快的 R-CNN：利用区域提议网络实现实时目标检测”，发表于 *NIPS*, 2015年。[11] K. He, X. Zhang, S. Ren 与 J. Sun, “用于图像识别的深度残差学习”，*arXiv:1512.03385*, 2015年。[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy 与 S. Reed, “SSD：单次多框检测器”，*arXiv:1512.02325*, 2015年。[13] M. D. Zeiler 与 R. Fergus, “卷积网络的可视化与理解”，发表于 *ECCV*, 2014年。[14] J. Long, E. Shelhamer 与 T. Darrell, “用于语义分割的全卷积网络”，发表于 *CVPR*, 2015年。[15] M. Holschneider, R. Kronland-Martinet, J. Morlet 与 P. Tchamitchian, “一种借助小波变换进行信号分析的实时算法”，发表于 *Wavelets: Time-Frequency Methods and Phase Space*, 1989年, 第289–297页。[16] A. Giusti, D. Ciresan, J. Masci, L. Gambardella 与 J. Schmidhuber, “使用深度最大池化卷积神经网络进行快速图像扫描”，发表于 *ICIP*, 2013年。[17] L.-C. Chen, Y. Yang, J. Wang, W. Xu 与 A. L. Yuille, “关注尺度：尺度感知的语义图像分割”，发表于 *CVPR*, 2016年。[18] I. Kokkinos, “利用深度学习推进边界检测的边界”，发表于 *ICLR*, 2016年。[19] S. Lazebnik, C. Schmid 与 J. Ponce, “超越特征袋：用于自然场景类别识别的空间金字塔匹配”，发表于 *CVPR*, 2006年。[20] K. He, X. Zhang, S. Ren 与 J. Sun, “深度卷积网络中用于视觉识别的空间金字塔池化”，发表于 *ECCV*, 2014年。[21] B. Hariharan, P. Arbeláez, R. Girshick 与 J. Malik, “用于目标分割和细粒度定位的超列”，发表于 *CVPR*, 2015年。[22] P. Krähenb ühl 与 V. Koltun, “高斯边势全连接条件随机场中的高效推理”，发表于 *NIPS*, 2011年。[23] C. Rother, V. Kolmogorov 与 A. Blake, “GrabCut：使用迭代图割进行交互式前景提取”，发表于 *SIGGRAPH*, 2004年。[24] J. Shotton, J. Winn, C. Rother 与 A. Criminisi, “用于图像理解的TextonBoost：通过联合建模纹理、布局和上下文进行多类目标识别与分割”，*IJCV*, 2009年。[25] A. Lucchi, Y. Li, X. Boix, K. Smith 与 P. Fua, “空间和全局约束对于分割真的必要吗？”，发表于 *ICCV*, 2011年。[26] X. He, R. S. Zemel 与 M. Carreira-Perpindn, “用于图像标注的多尺度条件随机场”，发表于 *CVPR*, 2004年。[27] L. Ladicky, C. Russell, P. Kohli 与 P. H. Torr, “用于目标类别图像分割的关联层次化条件随机场”，发表于 *ICCV*, 2009年。[28] V. Lempitsky, A. Vedaldi 与 A. Zisserman, “用于语义分割的塔架模型”，发表于 *NIPS*, 2011年。[29] A. Delong, A. Osokin, H. N. Isack 与 Y. Boykov, “具有标签成本的快速近似能量最小化”，*IJCV*, 2012年。[30] J. M. Gonfaus, X. Boix, J. Van de Weijer, A. D. Bagdanov, J. Serrat 与 J. Gonzalez, “用于联合分类与分割的和谐势”，发表于 *CVPR*, 2010年。[31] P. Kohli, P. H. Torr *et al.*, “用于强制标签一致性的鲁棒高阶势”，*IJCV*, 第82卷, 第3期, 第302–324页, 2009年。[32] L.-C. Chen, G. Papandreou 与 A. Yuille, “学习形状缩影字典及其在图像标注中的应用”，发表于 *ICCV*, 2013年。

[33] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Towards unified depth and semantic prediction from a single image," in *CVPR*, 2015.

[34] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserma, "The pascal visual object classes challenge a retrospective," *IJCV*, 2014.

[35] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.

[36] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *CVPR*, 2014.

[37] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[39] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *PAMI*, 2013.

[40] G. Lin, C. Shen, I. Reid *et al.*, "Efficient piecewise training of deep structured models for semantic segmentation," *arXiv:1504.01013*, 2015.

[41] 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:1408.5093*, 2014.

[42] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757, 2010.

[43] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *CVPR*, 2008.

[44] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *ICCV*, 2009.

[45] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *ECCV*, 2012.

[46] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *PAMI*, vol. 34, no. 7, pp. 1312–1328, 2012.

[47] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014.

[48] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *IJCV*, 2013.

[49] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *ECCV*, 2014.

[50] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *CVPR*, 2015.

[51] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," *arXiv:1412.1283*, 2014.

[52] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *arXiv:1411.4734*, 2014.

[53] M. Cogswell, X. Lin, S. Purushwalkam, and D. Batra, "Combining the best of graphical models and convnets for semantic segmentation," *arXiv:1412.4313*, 2014.

[54] D. Geiger and F. Girosi, "Parallel and deterministic algorithms from mrfs: Surface reconstruction," *PAMI*, vol. 13, no. 5, pp. 401–412, 1991.

[55] D. Geiger and A. Yuille, "A common framework for image segmentation," *IJCV*, vol. 6, no. 3, pp. 227–243, 1991.

[56] I. Kokkinos, R. Deriche, O. Faugeras, and P. Maragos, "Computational analysis and learning for a biologically motivated model of boundary detection," *Neurocomputing*, vol. 71, no. 10, pp. 1798–1812, 2008.

[57] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," *arXiv:1412.0623*, 2014.

[58] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a dcnn for semantic image segmentation," in *ICCV*, 2015.

[59] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *ICCV*, 2015.

[60] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *ICCV*, 2015.

[61] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.

[62] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *ICCV*, 2015.

[63] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," in *CVPR*, 2016.

[64] L.-C. Chen, A. Schwing, A. Yuille, and R. Urtasun, "Learning deep structured models," in *ICML*, 2015.

[65] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," *arXiv:1503.02351*, 2015.

[66] S. Chandra and I. Kokkinos, "Fast, exact and multi-scale inference for semantic image segmentation with deep Gaussian CRFs," *arXiv:1603.08358*, 2016.

[67] E. S. L. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," in *SIGGRAPH*, 2011.

[68] G. Bertasius, J. Shi, and L. Torresani, "High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision," in *ICCV*, 2015.

[69] P. O. Pinheiro and R. Collobert, "Weakly supervised semantic segmentation with convolutional networks," *arXiv:1411.6228*, 2014.

[70] D. Pathak, P. Krähenbühl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," 2015.

[71] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *NIPS*, 2015.

[72] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *ICCV*, 2011.

[73] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan, "Proposal-free network for instance-level object segmentation," *arXiv preprint arXiv:1509.02636*, 2015.

[74] J. E. Fowler, "The redundant discrete wavelet transform and additive noise," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 629–632, 2005.

[75] P. P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial," *Proceedings of the IEEE*, vol. 78, no. 1, pp. 56–93, 1990.

[76] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[77] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *arXiv:1605.06409*, 2016.

[78] J. Dai, K. He, Y. Li, S. Ren, and J. Sun, "Instance-sensitive fully convolutional networks," *arXiv:1603.08678*, 2016.

[79] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv:1511.05960*, 2015.

[80] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, "Optical flow with semantic segmentation and localized layers," *arXiv:1603.03911*, 2016.

[81] Z. Wu, C. Shen, and A. van den Hengel, "High-performance semantic segmentation using very deep fully convolutional networks," *arXiv:1604.04339*, 2016.

[82] M. J. Shensa, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *Signal Processing, IEEE Transactions on*, vol. 40, no. 10, pp. 2464–2482, 1992.

[83] M. Abadi, A. Agarwal *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, 2016.

[84] A. Adams, J. Baek, and M. A. Davis, "Fast high-dimensional filtering using the permutohedral lattice," in *Eurographics*, 2010.

[85] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *ICCV*, 2011.

[86] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv:1506.04579*, 2015.

[87] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.

[88] R. Vemulapalli, O. Tuzel, M.-Y. Liu, and R. Chellappa, "Gaussian conditional random field network for semantic segmentation," in *CVPR*, 2016.

[89] Z. Yan, H. Zhang, Y. Jia, T. Breuel, and Y. Yu, "Combining the best of convolutional layers and recurrent layers: A hybrid network for semantic segmentation," *arXiv:1603.04871*, 2016.

[90] G. Ghiasi and C. C. Fowlkes, "Laplacian reconstruction and refinement for semantic segmentation," *arXiv:1605.02264*, 2016.

[33] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, “面向从单张图像进行统一深度与语义预测的研究”, 发表于 *CVPR*, 2015.[34] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserma, “PASCAL视觉对象分类挑战回顾”, *IJCV*, 2014.[35] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, “真实场景中上下文对目标检测与语义分割的作用”, 发表于 *CVPR*, 2014.[36] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, “检测所能检测的：使用整体模型和身体部件进行目标检测与表示”, 发表于 *CVPR*, 2014.[37] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “用于语义城市场景理解的城市景观数据集”, 发表于 *CVPR*, 2016.[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “使用深度卷积网络和全连接条件随机场进行语义图像分割”, 发表于 *ICLR*, 2015.[39] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “学习用于场景标注的层次化特征”, *PAMI*, 2013.[40] G. Lin, C. Shen, I. Reid *et al.*, “用于语义分割的深度结构化模型的高效分段训练”, *arXiv:1504.01013*, 2015.[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe：用于快速特征嵌入的卷积架构”, *arXiv:1408.5093*, 2014.[42] Z. Tu and X. Bai, “自动上下文及其在高级视觉任务和3D脑图像分割中的应用”, *IEEE Trans. Pattern Anal. Mach. Intell.*, 卷 32, 期 10, 页 1744–1757, 2010.[43] J. Shotton, M. Johnson, and R. Cipolla, “用于图像分类与分割的语义纹理森林”, 发表于 *CVPR*, 2008.[44] B. Fulkerson, A. Vedaldi, and S. Soatto, “使用超像素邻域的类别分割与目标定位”, 发表于 *ICCV*, 2009.[45] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, “使用二阶池化的语义分割”, 发表于 *ECCV*, 2012.[46] J. Carreira and C. Sminchisescu, “CPMC：使用约束参数化最小割的自动目标分割”, *PAMI*, 卷 34, 期 7, 页 1312–1328, 2012.[47] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, “多尺度组合分组”, 发表于 *CVPR*, 2014.[48] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, “用于目标识别的选择性搜索”, *IJCV*, 2013.[49] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, “同步检测与分割”, 发表于 *ECCV*, 2014.[50] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, “使用放大特征的前馈语义分割”, 发表于 *CVPR*, 2015.[51] J. Dai, K. He, and J. Sun, “用于联合目标与背景分割的卷积特征掩蔽”, *arXiv:1412.1283*, 2014.[52] D. Eigen and R. Fergus, “使用通用的多尺度卷积架构预测深度、表面法线和语义标签”, *arXiv:1411.4734*, 2014.[53] M. Cogswell, X. Lin, S. Purushwalkam, and D. Batra, “结合图模型与卷积网络的优势进行语义分割”, *arXiv:1412.4313*, 2014.[54] D. Geiger and F. Girosi, “源自马尔可夫随机场的并行与确定性算法：表面重建”, *PAMI*, 卷 13, 期 5, 页 401–412, 1991.[55] D. Geiger and A. Yuille, “图像分割的通用框架”, *IJCV*, 卷 6, 期 3, 页 227–243, 1991.[56] I. Kokkinos, R. Deriche, O. Faugeras, and P. Maragos, “一种生物启发的边界检测模型的计算分析与学习”, *Neurocomputing*, 卷 71, 期 10, 页 1798–1812, 2008.[57] S. Bell, P. Upchurch, N. Snavely, and K. Bala, “使用材料上下文数据库在真实环境中进行材料识别”, *arXiv:1412.0623*, 2014.[58] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, “用于语义图像分割的深度卷积神经网络的弱监督与半监督学习”, 发表于 *ICCV*, 2015.[59] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, “作为循环神经网络的条件随机场”, 发表于 *ICCV*, 2015.[60] J. Dai, K. He, and J. Sun, “Boxsup：利用边界框监督卷积网络进行语义分割”, 发表于 *ICCV*, 2015.

[61] H. Noh, S. Hong, 和 B. Han, “用于语义分割的去卷积网络学习”, 发表于 *ICCV*, 2015年。[62] Z. Liu, X. Li, P. Luo, C. C. Loy, 和 X. Tang, “通过深度解析网络进行语义图像分割”, 发表于 *ICCV*, 2015年。[63] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, 和 A. L. Yuille, “使用CNN和判别性训练域变换进行任务特定边缘检测的语义图像分割”, 发表于 *CVPR*, 2016年。[64] L.-C. Chen, A. Schwing, A. Yuille, 和 R. Urtasun, “学习深度结构化模型”, 发表于 *ICML*, 2015年。[65] A. G. Schwing 和 R. Urtasun, “全连接深度结构化网络”, *arXiv:1503.02351*, 2015年。[66] S. Chandra 和 I. Kokkinos, “使用深度高斯CRF进行语义图像分割的快速、精确和多尺度推理”, *arXiv:1603.08358*, 2016年。[67] E. S. L. Gastal 和 M. M. Oliveira, “用于边缘感知图像和视频处理的域变换”, 发表于 *SIGGRAPH*, 2011年。[68] G. Bertasius, J. Shi, 和 L. Torresani, “高为低与低为高：从深度对象特征进行高效边界检测及其在高级视觉中的应用”, 发表于 *ICCV*, 2015年。[69] P. O. Pinheiro 和 R. Collobert, “使用卷积网络进行弱监督语义分割”, *arXiv:1411.6228*, 2014年。[70] D. Pathak, P. Krähenbühl, 和 T. Darrell, “用于弱监督分割的约束卷积神经网络”, 2015年。[71] S. Hong, H. Noh, 和 B. Han, “用于半监督语义分割的解耦深度神经网络”, 发表于 *NIPS*, 2015年。[72] A. Vezhnevets, V. Ferrari, 和 J. M. Buhmann, “使用多图像模型进行弱监督语义分割”, 发表于 *ICCV*, 2011年。[73] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, 和 S. Yan, “用于实例级对象分割的无提议网络”, *arXiv preprint arXiv:1509.02636*, 2015年。[74] J. E. Fowler, “冗余离散小波变换与加性噪声”, *IEEE Signal Processing Letters*, 第12卷, 第9期, 第629–632页, 2005年。[75] P. P. Vaidyanathan, “多速率数字滤波器、滤波器组、多相网络及应用：教程”, *Proceedings of the IEEE*, 第78卷, 第1期, 第56–93页, 1990年。[76] F. Yu 和 V. Koltun, “通过空洞卷积进行多尺度上下文聚合”, 发表于 *ICLR*, 2016年。[77] J. Dai, Y. Li, K. He, 和 J. Sun, “R-FCN：通过基于区域的全卷积网络进行目标检测”, *arXiv:1605.06409*, 2016年。[78] J. Dai, K. He, Y. Li, S. Ren, 和 J. Sun, “实例敏感的全卷积网络”, *arXiv:1603.08678*, 2016年。[79] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, 和 R. Nevatia, “ABC-CNN：一种基于注意力的卷积神经网络用于视觉问答”, *arXiv:1511.05960*, 2015年。[80] L. Sevilla-Lara, D. Sun, V. Jampani, 和 M. J. Black, “结合语义分割和局部化层的光流”, *arXiv:1603.03911*, 2016年。[81] Z. Wu, C. Shen, 和 A. van den Hengel, “使用非常深的全卷积网络进行高性能语义分割”, *arXiv:1604.04339*, 2016年。[82] M. J. Shensa, “离散小波变换：融合à trous和Mallat算法”, *Signal Processing, IEEE Transactions on*, 第40卷, 第10期, 第2464–2482页, 1992年。[83] M. Abadi, A. Agarwal *et al.*, “TensorFlow：异构分布式系统上的大规模机器学习”, *arXiv:1603.04467*, 2016年。[84] A. Adams, J. Baek, 和 M. A. Davis, “使用置换面体晶格进行快速高维滤波”, 发表于 *Eurographics*, 2010年。[85] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, 和 J. Malik, “来自逆向检测器的语义轮廓”, 发表于 *ICCV*, 2011年。[86] W. Liu, A. Rabinovich, 和 A. C. Berg, “ParseNet：看得更广以见得更好”, *arXiv:1506.04579*, 2015年。[87] T.-Y. Lin *et al.*, “Microsoft COCO：上下文中的常见对象”, 发表于 *ECCV*, 2014年。[88] R. Vemulapalli, O. Tuzel, M.-Y. Liu, 和 R. Chellappa, “用于语义分割的高斯条件随机场网络”, 发表于 *CVPR*, 2016年。[89] Z. Yan, H. Zhang, Y. Jia, T. Breuel, 和 Y. Yu, “结合卷积层和循环层的优点：一种用于语义分割的混合网络”, *arXiv:1603.04871*, 2016年。[90] G. Ghiasi 和 C. C. Fowlkes, “用于语义分割的拉普拉斯重建与细化”, *arXiv:1605.02264*, 2016年。

[91] A. Arnab, S. Jayasumana, S. Zheng, and P. Torr, "Higher order potentials in end-to-end trainable conditional random fields," *arXiv:1511.08119*, 2015.

[92] F. Shen and G. Zeng, "Fast semantic image segmentation with high order context and guided filtering," *arXiv:1605.04068*, 2016.

[93] Z. Wu, C. Shen, and A. van den Hengel, "Bridging category-level and instance-level semantic image segmentation," *arXiv:1605.06885*, 2016.

[94] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *arXiv:1603.05027*, 2016.

[95] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille, "Zoom better to see clearer: Huamn part segmentation with auto zoom net," *arXiv:1511.06881*, 2015.

[96] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," *arXiv:1511.04510*, 2015.

[97] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," *arXiv:1603.07063*, 2016.

[98] J. Wang and A. Yuille, "Semantic part segmentation using compositional model combining shape and appearance," in *CVPR*, 2015.

[99] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Joint object and part segmentation using deep learned potentials," in *ICCV*, 2015.

[100] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv:1511.00561*, 2015.

[101] J. Uhrig, M. Cordts, U. Franke, and T. Brox, "Pixel-level encoding and depth layering for instance-level semantic labeling," *arXiv:1604.05096*, 2016.

[102] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

Liang-Chieh Chen received his B.Sc. from National Chiao Tung University, Taiwan, his M.S. from the University of Michigan- Ann Arbor, and his Ph.D. from the University of California- Los Angeles. He is currently working at Google. His research interests include semantic image segmentation, probabilistic graphical models, and machine learning.

George Papandreou (S'03–M'09–SM'14) holds a Diploma (2003) and a Ph.D. (2009) in Electrical Engineering and Computer Science, both from the National Technical University of Athens (NTUA), Greece. He is currently a Research Scientist at Google, following appointments as Research Assistant Professor at the Toyota Technological Institute at Chicago (2013-2014) and Postdoctoral Research Scholar at the University of California, Los Angeles (2009-2013).

His research interests are in computer vision and machine learning, with a current emphasis on deep learning. He regularly serves as a reviewer and program committee member to the main journals and conferences in computer vision, image processing, and machine learning. He has been a co-organizer of the NIPS 2012, 2013, and 2014 Workshops on Perturbations, Optimization, and Statistics and co-editor of a book on the same topic (MIT Press, 2016).

Iasonas Kokkinos (S'02–M'06) obtained the Diploma of Engineering in 2001 and the Ph.D. Degree in 2006 from the School of Electrical and Computer Engineering of the National Technical University of Athens in Greece, and the Habilitation Degree in 2013 from Universit Paris-Est. In 2006 he joined the University of California at Los Angeles as a postdoctoral scholar, and in 2008 joined as faculty the Department of Applied Mathematics of Ecole Centrale Paris (CentraleSupelec), working an associate professor in the Center for Visual Computing of CentraleSupelec and affiliate researcher at INRIA-Saclay. In 2016 he joined University College London and Facebook Artificial Intelligence Research. His currently research activity is on deep learning for computer vision, focusing in particular on structured prediction for deep learning, shape modeling, and multi-task learning architectures. He has been awarded a young researcher grant by the French National Research Agency, has served as associate editor for the Image and Vision Computing and Computer Vision and Image Understanding Journals, serves regularly as a reviewer and area chair for all major computer vision conferences and journals.

Kevin Murphy was born in Ireland, grew up in England, went to graduate school in the USA (MEng from U. Penn, PhD from UC Berkeley, Postdoc at MIT), and then became a professor at the Computer Science and Statistics Departments at the University of British Columbia in Vancouver, Canada in 2004. After getting tenure, Kevin went to Google in Mountain View, California for his sabbatical. In 2011, he converted to a full-time research scientist at Google. Kevin has published over 50 papers in refereed conferences and journals related to machine learning and graphical models. He has recently published an 1100-page textbook called "Machine Learning: a Probabilistic Perspective" (MIT Press, 2012).

Alan L. Yuille (F'09) received the BA degree in math- ematics from the University of Cambridge in 1976. His PhD on theoretical physics, supervised by Prof. S.W. Hawking, was approved in 1981. He was a research scientist in the Artificial Intelligence Laboratory at MIT and the Division of Applied Sciences at Harvard University from 1982 to 1988. He served as an assistant and associate professor at Harvard until 1996. He was a senior research scientist at the Smith-Kettlewell Eye Research Institute from 1996 to 2002. He joined the University of California, Los Angeles, as a full professor with a joint appointment in statistics and psychology in 2002, and computer science in 2007. He was appointed a Bloomberg Distinguished Professor at Johns Hopkins University in January 2016. He holds a joint appointment between the Departments of Cognitive science and Computer Science. His research interests include computational models of vision, mathematical models of cognition, and artificial intelligence and neural network

[91] A. Arnab, S. Jayasumana, S. Zheng, and P. Torr, “端到端可训练条件随机场中的高阶势函数，” *arXiv:1511.08119*, 2015.[92] F. Shen and G. Zeng, “利用高阶上下文和引导滤波的快速语义图像分割，” *arXiv:1605.04068*, 2016.[93] Z. Wu, C. Shen, and A. van den Hengel, “桥接类别级与实例级语义图像分割，” *arXiv:1605.06885*, 2016.[94] K. He, X. Zhang, S. Ren, and J. Sun, “深度残差网络中的恒等映射，” *arXiv:1603.05027*, 2016.[95] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille, “放大以看得更清：基于自动缩放网络的人体部位分割，” *arXiv:1511.06881*, 2015.[96] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, “基于局部-全局长短期记忆的语义对象解析，” *arXiv:1511.04510*, 2015.[97] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, “基于图长短期记忆的语义对象解析，” *arXiv:1603.07063*, 2016.[98] J. Wang and A. Yuille, “使用结合形状与外观的组合模型进行语义部件分割，” 发表于 *CVPR*, 2015.[99] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, “使用深度学习势函数进行联合对象与部件分割，” 发表于 *ICCV*, 2015.[100] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet：一种用于图像分割的深度卷积编码器-解码器架构，” *arXiv:1511.00561*, 2015.[101] J. Uhrig, M. Cordts, U. Franke, and T. Brox, “用于实例级语义标注的像素级编码与深度分层，” *arXiv:1604.05096*, 2016.[102] O. Ronneberger, P. Fischer, and T. Brox, “U-Net：用于生物医学图像分割的卷积网络，” 发表于 *MICCAI*, 2015.

Iasonas Kokkinos (S’02–M’06) 于2001年获得工程学文凭，并于2006年在希腊雅典国家技术大学电气与计算机工程学院取得博士学位，2013年在巴黎东大学获得特许任教资格。2006年，他加入加州大学洛杉矶分校担任博士后研究员，2008年加入巴黎中央理工学院（中央理工-高等电力学院）应用数学系任教，担任副教授。

CentraleSupelec视觉计算中心研究员，同时兼任INRIA-Saclay附属研究员。2016年他加入伦敦大学学院与Facebook人工智能研究院。他当前的研究聚焦于计算机视觉领域的深度学习，特别关注深度学习中的结构化预测、形状建模以及多任务学习架构。他曾获得法国国家研究署颁发的青年学者研究基金，担任《图像与视觉计算》和《计算机视觉与图像理解》期刊的副主编，并长期为所有主流计算机视觉会议及期刊担任审稿人和领域主席。

凯文·墨菲出生于爱尔兰，在英格兰长大，于美国攻读研究生（获宾夕法尼亚大学工程硕士、加州大学伯克利分校博士学位，并在麻省理工学院从事博士后研究），随后于2004年成为加拿大温哥华不列颠哥伦比亚大学计算机科学与统计系的教授。获得终身教职后，凯文前往加利福尼亚州山景城的谷歌进行学术休假。2011年，他转为谷歌的全职研究科学家。凯文已在同行评审的学术会议上发表了50多篇论文——他涉足机器学习与图模型相关的会议和期刊。他最近出版了一本1100页的教科书，名为《机器学习：概率视角》（麻省理工学院出版社，2012年）。

Liang-Chieh Chen 在台湾国立交通大学获得学士学位，在密歇根大学安娜堡分校获得硕士学位，并在加州大学洛杉矶分校获得博士学位。他目前任职于谷歌。他的研究方向包括语义图像分割、概率图模型以及机器学习。

乔治·帕潘德里欧（S’03–M’09–SM’14）持有希腊雅典国家技术大学（NTUA）颁发的电气工程与计算机科学专业文凭（2003年）和博士学位（2009年）。他目前是谷歌的研究科学家，此前曾担任芝加哥丰田技术研究所的研究助理教授（2013-2014年）和加州大学洛杉矶分校的博士后研究员（2009-2013年）。

他的研究兴趣在于计算机视觉以及机器学习，目前侧重于深度学习。他定期担任计算机视觉、图像处理和机器学习领域主要期刊和会议的审稿人及程序委员会委员。他曾共同组织2012、2013及2014年神经信息处理系统会议关于扰动、优化与统计的研讨会，并合编了同主题著作（麻省理工学院出版社，2016年）。

艾伦·L·于勒（F’09）于1976年获得剑桥大学数学学士学位。其理论物理学博士学位由S.W.霍金教授指导，于1981年获批。1982年至1988年间，他先后在麻省理工学院人工智能实验室和哈佛大学应用科学部担任研究科学家。此后至1996年，他在哈佛大学担任助理教授和副教授。1996年至2002年，他加入加州大学洛杉矶分校，于2002年受聘为统计学与心理学联合聘任的正教授，并于2007年增聘计算机科学职位。2016年1月，他被任命为约翰斯·霍普金斯大学的彭博特聘教授。他在认知科学系与计算机科学系拥有联合聘任职位。其研究兴趣包括视觉计算模型、认知数学模型、人工智能与神经网络。