Pyramid Scene Parsing Network

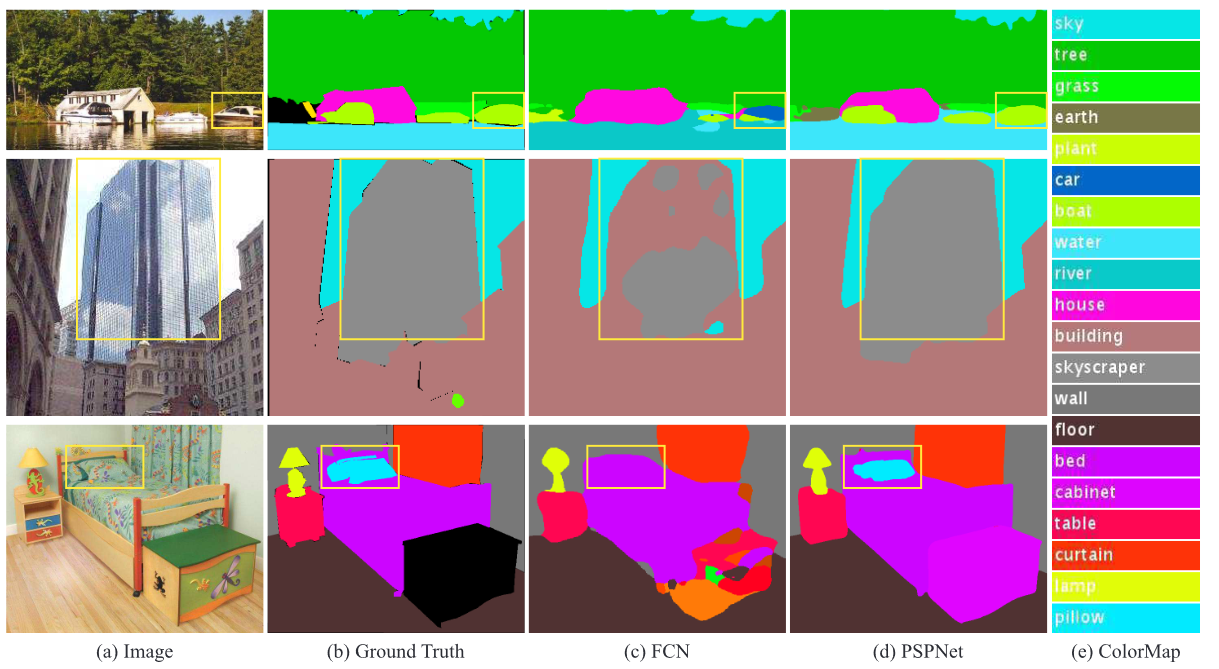
摘要

场景解析对于不受限制的开放词汇和多样化的场景来说是一个挑战。在本文中，我们通过我们的金字塔集合模块和提议的金字塔场景解析网络（PSPNet），利用基于不同区域的全局上下文信息的能力。我们的全局先验表示能够有效地在场景解析任务中产生高质量的结果，而PSPNet为像素级预测提供了一个卓越的框架。所提出的方法在各种数据集上取得了最先进的性能。它在2016年ImageNet场景解析挑战赛、PASCAL VOC 2012基准测试和Cityscapes基准测试中名列第一。单个PSPNet在PASCAL VOC 2012上获得了85.4%的mIoU准确性，在Cityscapes上获得了80.2%的准确性。

代码链接：https://github.com/hszhao/PSPNet

介绍

基于语义分割的场景解析是计算机视觉的一个基本课题。其目的是为图像中的每个像素分配一个类别标签。场景解析提供了对场景的完整理解。它预测了每个元素的标签、位置以及形状。这个话题对于自动驾驶、机器人感应等潜在应用具有广泛的意义。场景解析的难度与场景和标签的种类密切相关。最早的场景解析任务[23]是对LMO数据集[22]上的2,688张图像进行33个场景的分类。最近的PASCAL VOC语义分割和PASCAL语境数据集[8, 29]包括更多具有类似语境的标签，如椅子和沙发，马和牛等。新的ADE20K数据集[43]是最具挑战性的数据集，它拥有大量无限制的开放词汇和更多的场景类别。图1中显示了一些有代表性的图像。要为这些数据集开发一个有效的算法，需要克服一些困难。目前最先进的场景解析框架大多是基于完全卷积网络（FCN）[26]。基于深度卷积神经网络（CNN）的方法促进了动态物体的理解，但考虑到多样化的场景和不受限制的词汇，仍然面临挑战。一个例子是在图2的第一行



，一艘船被误认为是一辆汽车。这些错误是由于物体的相似外观造成的。但是，当查看图像时，考虑到之前的背景，即场景被描述为河边的船屋，应该会产生正确的预测。为了实现准确的场景感知，知识图谱依赖于场景背景的先验信息。我们发现，目前基于FCN的模型的主要问题是缺乏合适的策略来利用全局场景类别线索。对于典型的复杂场景的理解，以前为了得到一个全局的图像级别的特征，空间金字塔集合[18]被广泛采用，其中空间统计为整个场景的解释提供了一个良好的描述符。空间金字塔集合网络[12]进一步增强了这种能力。与这些方法不同，为了纳入合适的全局特征，我们提出金字塔场景解析网络（PSPNet）。除了传统的稀释FCN[3, 40]用于像素预测外，我们将像素级的特征扩展到特别设计的全局金字塔集合的特征。局部和全局的线索共同使最终的预测结果更加可靠。我们还提出了一种具有深度监督损失的优化策略。我们给出了所有的实施细节，这是我们在本文中取得良好性能的关键，并公开了代码和训练好的模型。我们的方法在所有可用的数据集上实现了最先进的性能。它是2016年ImageNet场景解析挑战赛的冠军[43]，并在PASCAL VOC 2012语义分割基准中获得第一名[8]，在城市场景Cityscapes数据中获得第一名[6]。他们表明PSPNet为像素级预测任务提供了一个很好的方向，这甚至可能有利于基于CNN的立体匹配、光流、深度估计等后续工作。我们的主要贡献有三方面。

- 我们提出了一个金字塔式的场景解析网络，在基于FCN的像素预测框架中嵌入了困难的景物背景特征。

- 我们为深度ResNet[13]开发了一个基于深度监督损失的有效优化策略。

- 我们为最先进的场景解析和语义分割建立了一个实用的系统，其中包括所有关键的实施细节。

3. 金字塔场景解析网络

我们首先观察和分析了将FCN方法用于场景解析时的代表性失败案例。它们促使我们提出了金字塔集合模块作为有效的全局环境先验。我们的金字塔场景解析网络（PSPNet）如图3所示，用于提高复杂场景解析中开放词汇的对象和物品识别的性能。

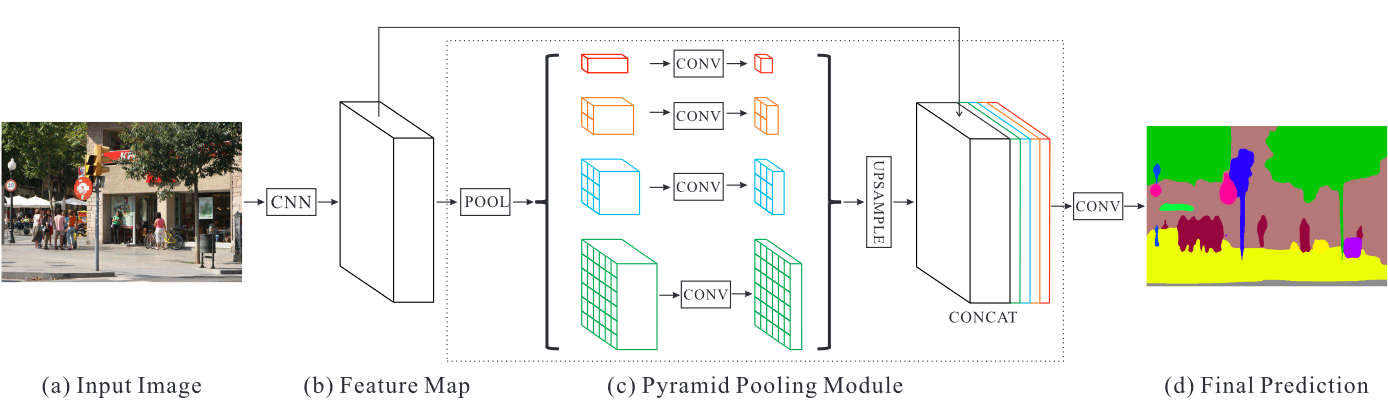


图3. 我们提出的PSPNet的概述。给定一个输入图像（a），我们首先使用CNN得到最后一个卷积层的特征图（b），然后应用金字塔解析模块来收获不同的子区域表征，接着是上采样和串联层，形成最终的特征表征，在（c）中同时携带局部和全局信息。最后，该表征被送入卷积层以获得最终的每像素预测（d）。

3.1. 重要观察

新的ADE20K数据集[43]包含150个东西/物体类别标签（如墙、天空和树）和1038个图像级场景描述符（如机场航站楼、卧室和街道）。所以大量的标签和庞大的场景分布就出现了。检查[43]中提供的FCN基线的预测结果，我们总结了复杂场景解析的几个共同问题。

**不匹配的关系**

语境关系是普遍存在的，特别是对于复杂的场景理解来说，是非常重要的。存在着共同出现的视觉模式。例如，一架飞机可能在跑道上或在天空中飞行，而不是在道路上。对于图2中第一行的例子，FCN根据黄色框中的船的外观预测它是一辆 "汽车"。但是常识告诉我们，汽车很少会在河上行驶。缺乏收集上下文信息的能力，就会增加错误分类的机会。

**混淆类别**

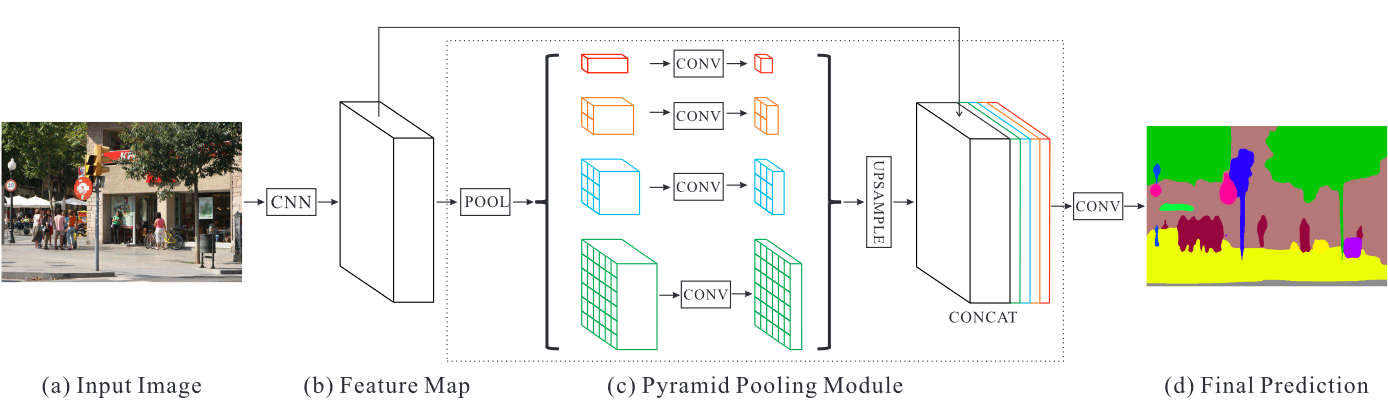
ADE20K数据集[43]中有许多类标签对在分类中是混乱的。例如，田地和大地；山和丘陵；墙、房子、建筑和摩天大楼。它们具有相似的外观。如[43]所述，对整个数据集进行标注的专家注释者仍会出现17.60%的像素错误。在图2的第二行，FCN预测方框中的物体是摩天大楼的一部分和建筑的一部分。这些结果应该被排除，这样整个物体要么是摩天大楼，要么是建筑物，而不是两者。这个问题可以通过利用类别之间的关系来补救。

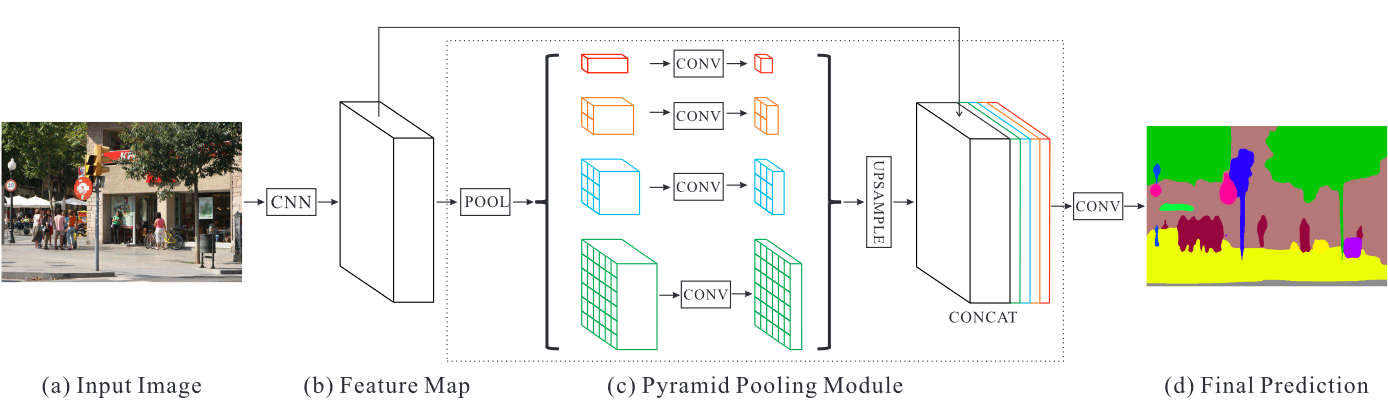
**不显眼的类别**

场景包含任意大小的物体/物品。一些小尺寸的东西，如路灯和招牌，很难被发现，而它们可能是非常重要的。相反，大的物体或东西可能会超出FCN的接收范围，从而导致不连续的预测。如图2的第三行所示，枕头和床单的外观相似。忽略全局场景类别可能无法解析枕头。为了提高对明显小的或大的物体的性能，我们应该非常注意包含不明显类别的不同子区域。总结这些观察结果，许多错误部分或完全与上下文关系和不同感受区的全局信息有关。因此，一个具有合适的全局场景级先验的深度网络可以大大改善场景解析的性能。

3.2. 金字塔集合模块

通过上述分析，在下文中，我们将介绍金字塔集合模块，根据经验，它被证明是一个有效的全局上下文先验。在一个深度神经网络中，感受野的大小可以大致表明我们对上下文信息的使用程度。虽然理论上ResNet[13]的感受野已经大于输入图像，但Zhou等人[42]表明CNN的经验感受野比理论上的感受野小得多，特别是在高层。这使得许多网络不能充分地纳入重要的全局景物先验。我们通过提出一个有效的全局先验表示来解决这个问题。全局平均池是一个很好的基线模型，作为全局背景先验，它通常用于图像分类任务[34，13]。在[24]中，它被成功应用于语义分割。但是对于ADE20K[43]中的复杂场景图像，这种策略不足以涵盖必要的信息。这些场景图像中的像素被注释了许多东西和物体。直接融合它们形成一个单一的矢量可能会失去空间关系并导致歧义。在这方面，全局背景信息和子区域背景有助于区分不同的类别。一个更强大的表征可以是来自不同子区域的信息与这些感受野的融合。在场景/图像分类的经典工作[18, 12]中也得出了类似的结论。在[12]中，由金字塔集合产生的不同层次的特征图最终被扁平化并串联起来，送入一个全连接层进行分类。这种全局先验的设计是为了消除CNN在图像分类中的固定尺寸约束。为了进一步减少不同子区域之间的上下文信息损失，我们提出了一个分层的全局先验，包含不同尺度的信息，并在不同子区域之间变化。我们称之为金字塔集合模块，用于在深度神经网络的最后一层特征图上构建全局场景先验，如图3的（c）部分所示。

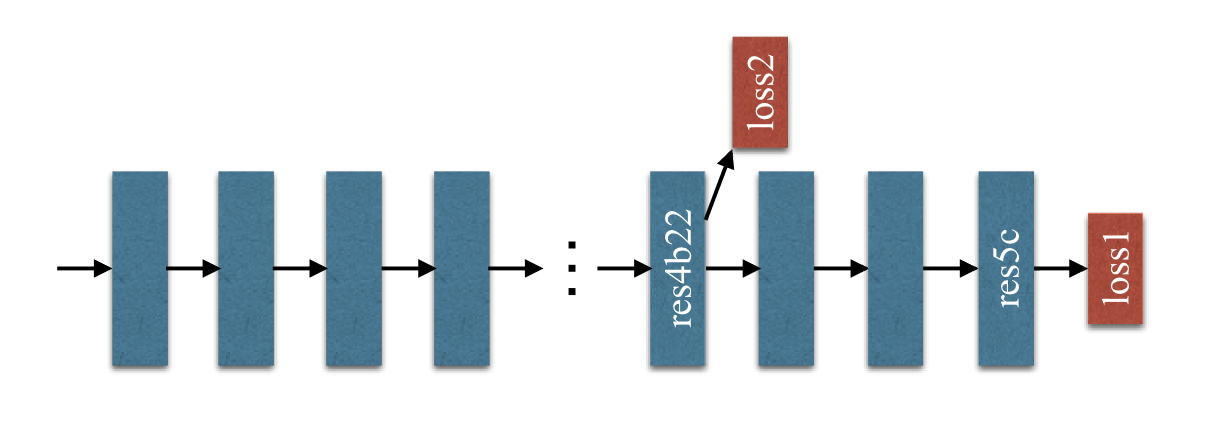
金字塔集合模块融合了四个不同的金字塔尺度下的特征。红色显示的最粗级别是全局池化，以产生一个单仓输出。接下来的金字塔级别将特征图分离成不同的子区域，形成不同位置的集合表示。金字塔集合模块中不同级别的输出包含不同大小的特征图。为了保持全局特征的权重，我们在每个金字塔层之后使用1×1卷积层，如果金字塔层的大小为N，则将上下文表示的维度减少到原始维度的1/N。然后，我们直接对低维度的特征图进行上采样，通过双线性插值得到与原始特征图相同大小的特征。最后，不同级别的特征图被串联起来作为最终的金字塔集合全局特征。需要注意的是，金字塔的层数和每层的大小都可以修改。它们与送入金字塔汇集层的特征图的大小有关。该结构通过采用不同大小的池化核在几步内抽象出不同的子区域。因此，多级核应该在表示上保持合理的差距。我们的金字塔池化模块是一个四级的模块，仓的大小分别为1×1、2×2、3×3和6×6。对于最大和平均之间的池化操作类型，我们在第5.2节中进行了大量的实验来显示其区别。  
  
3.3. 网络结构



通过金字塔集合模块，我们提出了金字塔场景解析网络（PSPNet），如图3所示。给定图3(a)中的输入图像，我们使用预训练的ResNet[13]模型和扩张网络策略[3, 40]来提取特征图。最终的特征图大小为输入图像的1/8，如图3（b）所示。在地图之上，我们使用（c）所示的金字塔集合模块来收集上下文信息。使用我们的4级金字塔，池化核覆盖了整个图像、一半的图像和小部分的图像。它们被融合为全局先验。然后我们在(c)的最后部分将先验与原始特征图连接起来。接着是卷积层，生成（d）中的最终预测图。为了解释我们的结构，PSPNet为像素级场景解析提供了一个有效的全局上下文先验。金字塔汇集模块可以收集各级信息，比全局汇集更具代表性[24]。在计算成本（池化和双线性插值）方面，我们的PSPNet与原来的扩张FCN网络相比并没有增加多少。在端到端学习中，全局金字塔池化模块和局部FCN特征可以同时优化。

4. 基于ResNet的FCN的深度监督

深度预训练的网络会带来良好的性能[17, 33, 13]。然而，增加网络的深度可能会引入额外的优化困难，如[32, 19]中的图像分类。ResNet通过在每个块中的跳过连接解决了这个问题。深度ResNet的后层主要是在前层的基础上学习残差。

  
相反，我们建议通过监督产生最初的结果，并有额外的损失，之后用最终的损失学习残差。因此，深度网络的优化被分解成两个，每个都更容易解决。图4展示了我们的深度监督ResNet101[13]模型的一个例子。除了使用softmax loss训练最终分类器的主分支外，另一个分类器在第四阶段后被应用，即res4b22残差块。与中继反向传播[32]将后向辅助损失阻断到几个浅层不同，我们让这两个损失函数通过所有的前层。辅助损失有助于优化学习过程，而主分支损失则承担最大责任。我们增加权重来平衡辅助损失。在测试阶段，我们放弃这个辅助分支，只使用经过优化的主分支进行最终预测。这种基于ResNet的FCN深度监督训练策略在不同的实验环境下都很有用，并且可以与预先训练好的ResNet模型一起使用。这体现了这种学习策略的通用性。更多的细节将在第5.2节提供。

6. 结束语

我们已经提出了一个有效的金字塔场景解析网络，用于复杂场景的理解。全局金字塔集合功能提供了额外的上下文信息。我们还为基于ResNet的FCN网络提供了一个深度监督的优化策略。我们希望公开的实施细节能够帮助社区采用这些有用的策略进行场景解析和语义分割，并推动相关技术的发展。