Deep Dual-resolution Networks for Real-time and Accurate Semantic Segmentation of Road Scenes

用于实时和准确的道路场景语义分割的深度双分辨率网络

arxiv：2021 Cited：97 IEEE



摘要

语义分割是自动驾驶汽车理解周围场景的一项关键技术。当代模型的吸引人的性能通常是以繁重的计算和冗长的推理时间为代价的，这对自动驾驶来说是不可容忍的。使用轻量级架构（编码器-解码器或双路径）或在低分辨率图像上进行推理，最近的方法实现了非常快的场景解析，甚至在单个1080Ti GPU上以超过100 FPS运行。然而，这些实时方法与基于扩张骨架的模型在性能上仍有很大差距。为了解决这个问题，我们提出了一个专门为实时语义分割设计的高效骨干网络系列。所提出的深度双分辨率网络（DDRNets）由两个深度分支组成，在这两个分支之间进行多次双边融合。此外，我们设计了一个新的上下文信息提取器，名为深度聚合金字塔池模块（DAPPM），以扩大有效的感受野，并基于低分辨率特征图融合多尺度上下文。我们的方法在Cityscapes和CamVid数据集上实现了准确性和速度之间的新的最先进权衡。特别是，在单个2080Ti GPU上，DDRNet-23-slim在Cityscapes测试集上以102 FPS产生77.4%的mIoU，在CamVid测试集上以230 FPS产生74.7%的mIoU。在广泛使用测试增强的情况下，我们的方法优于大多数最先进的模型，并且需要更少的计算。代码和训练好的模型可以在线获得。

代码链接：<https://github.com/chenjun2hao/DDRNet.pytorch>

介绍

图像分割是一项基本任务，输入图像的每个像素都应该被分配到相应的标签[1]-[3]。它在许多实际应用中起着至关重要的作用，如医学图像分割、自主车辆和机器人的导航[4], [5]。随着深度学习技术的兴起，卷积神经网络被应用于图像分割，并大大超过了基于手工制作的特征的传统方法。自从完全卷积网络（FCN）[6]被提出来处理语义分割问题后，一系列新颖的网络被提出。DeepLab[7]消除了ResNet中的一些下采样以保持高分辨率，并利用具有大扩张的卷积[8]来扩大感受野。从那时起，基于扩张卷积的骨干与上下文提取模块已经成为广泛用于各种方法的标准布局，包括DeepLabV2[9]，DeepLabV3[10]，PSPNet[11]，以及DenseASPP[12]。由于语义分割是一种密集预测的任务，神经网络需要输出大感受野的高分辨率特征图才能产生令人满意的结果，这在计算上是很昂贵的。这个问题对于自动驾驶的场景解析尤其关键，因为自动驾驶需要在非常大的图像上执行，以覆盖广阔的视野。因此，上述方法在推理阶段非常耗时，不能直接部署在实际的自动驾驶车辆上。由于利用多尺度测试来提高准确性，它们甚至不能在一秒钟内处理一幅图像。随着移动设备部署需求的不断增加，实时分割算法[13]-[17]正得到越来越多的关注。DFANet[18]采用了深度多尺度特征聚合和轻量级的深度可分离卷积，在100FPS的情况下实现了71.3%的测试mIoU。与编码器-解码器范式不同，作者在[19]中提出了一个由空间路径和背景路径组成的新型双边网络。特别是，空间路径利用三个相对宽的3×3卷积层来捕捉空间细节，而上下文路径是一个紧凑的预训练骨干，用于提取上下文信息。包括[20]在内的这类双边方法在当时取得了比encoderdecoder结构更高的推理速度。最近，一些旨在对道路场景进行语义分割的有竞争力的实时方法被提出。这些方法可以分为两类。一类是利用GPU高效的骨干网，特别是ResNet-18[21]-[23]。另一种是开发从头开始训练的复杂的轻量级编码器，其中的BiSeNetV2[24]在实时性能方面达到了一个新的高峰，在Cityscapes上以156 FPS实现了72.6%的测试mIoU。然而，除了[23]使用了额外的训练数据，这些最近的作品并没有显示出获得更高质量结果的潜力。其中一些作品由于刻意设计的架构和调整的超参数而缺乏可扩展性。此外，鉴于更强大的骨干网的繁荣，ResNet-18的优势不大。在本文中，我们提出了具有深度高分辨率表示的双分辨率网络，用于高分辨率图像的实时语义分割，特别是道路驾驶图像。我们的DDR网络从一个主干开始，然后分成两个具有不同分辨率的平行深度分支。一个深层分支生成相对高分辨率的特征图，另一个通过多次降采样操作提取丰富的语义信息。两个分支之间架起了多个双边连接，以实现高效的信息融合。此外，我们提出了一个名为DAPPM的新模块，它输入低分辨率的特征图，提取多尺度的上下文信息，并以级联的方式将它们合并。在对语义分割数据集进行训练之前，双分辨率网络在ImageNet上按照常见范式进行训练。根据对三个流行基准（即Cityscapes、CamVid和COCOStuff）的大量实验结果，DDRNets在分割精度和推理速度之间达到了很好的平衡。与其他实时算法相比，我们的方法在Cityscapes和CamVid上都达到了新的最先进的精确度，而不需要注意机制和额外的铃声或口哨。通过标准的测试增强，DDRNet可以与最先进的模型相媲美，并且需要更少的计算资源。我们还报告了统计学上的相关性能，并进行了消融实验来分析架构改进和标准训练技巧的效果。主要贡献总结如下：  
- 提出了一个具有深度双分辨率分支和多个双边融合的新型双边网络系列，作为高效的骨干网，用于实时语义分割。

- 设计了一个新的模块，通过结合特征聚合和金字塔池来获取丰富的上下文信息。当在低分辨率特征图上执行时，它导致推理时间的增加很少。

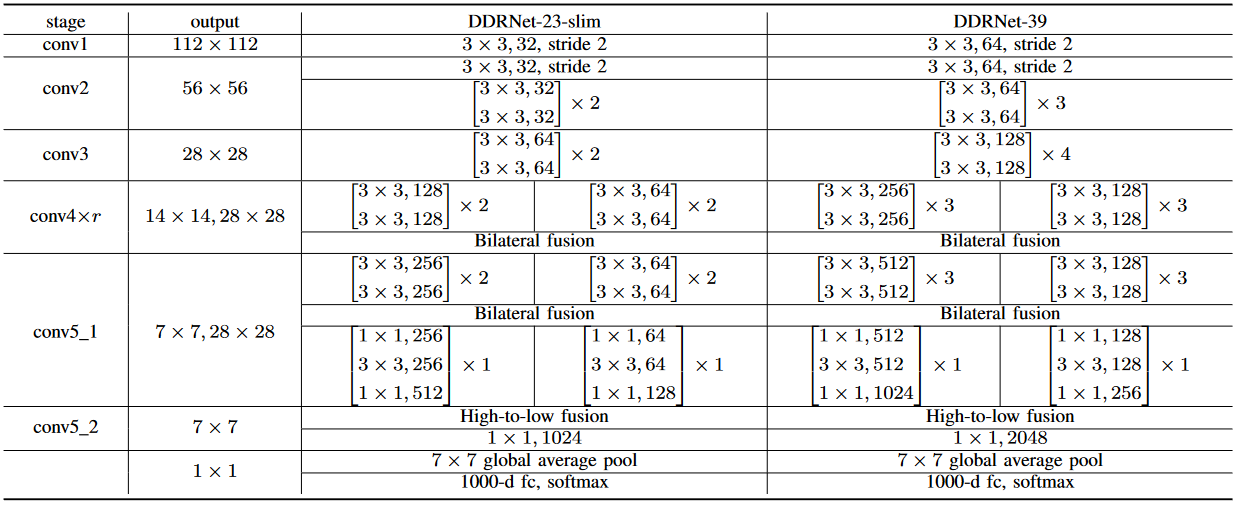
- 我们的方法在2080Ti上实现了准确度和速度之间的新的最先进的权衡，在Cityscapes测试集上以102 FPS的速度实现了77.4%的mIoU，在CamVid测试集上以230 FPS实现了74.7%的mIoU。据我们所知，我们是第一个在几乎实时（22 FPS）的情况下在Cityscapes上使用精细注释达到80.4%的mIoU。

3.方法

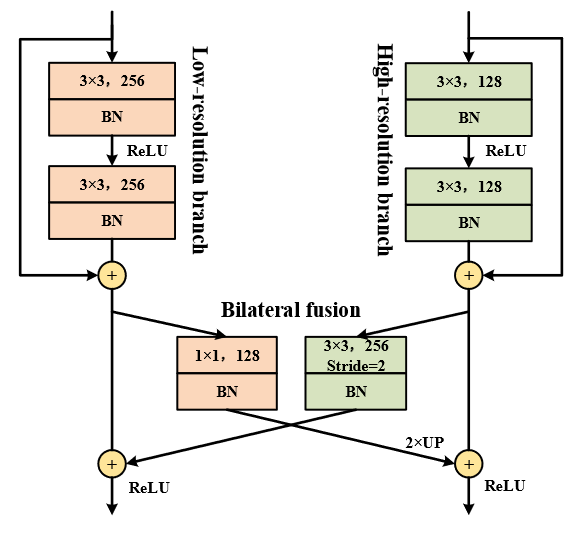
在这一节中，将描述整个pipeline，它由两个主要部分组成：深度双分辨率网络和深度聚合金字塔池模块。

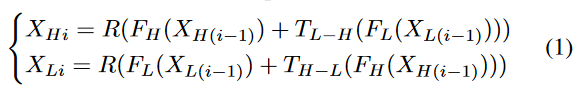
1. 深度双分辨率网络

为方便起见，我们可以在广泛使用的分类骨干网（如ResNets）上增加一个额外的高分辨率分支。为了实现分辨率和推理速度之间的权衡，我们让高分辨率分支创建分辨率为输入图像分辨率1/8的特征图。因此，高分辨率分支被附加到conv3阶段的末端。请注意，高分辨率分支不包含任何下采样操作，并且与低分辨率分支有一对一的对应关系，以形成深度的高分辨率表示。然后，在不同阶段可以进行多个双边特征融合，以充分融合空间信息和语义信息。DDRNet-23-slim和DDRNet-39的详细架构见表一。

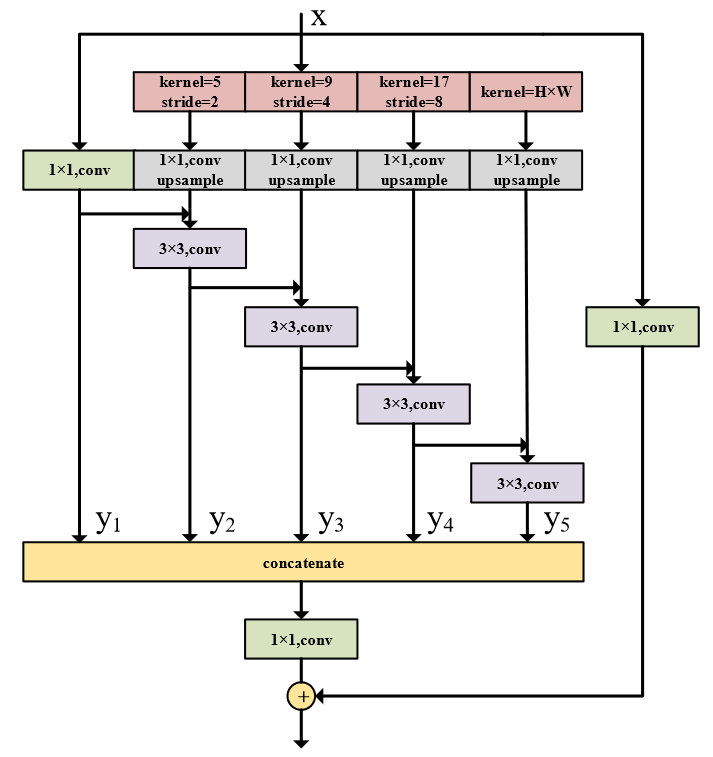


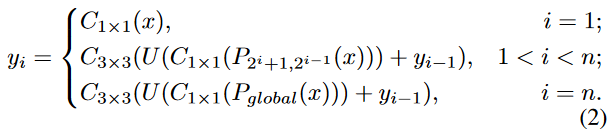
我们修改了原始ResNet的输入干，用两个连续的3×3卷积层代替了一个7×7卷积层。剩余的基本块被利用来构建主干和随后的两个分支。为了扩大输出维度，在每个分支的末端增加一个瓶颈块。双边融合包括将高分辨率分支融合到低分辨率分支（高-低融合）和将低分辨率融合到高分辨率分支（低-高融合）。对于高-低融合，高分辨率特征图在进行点状求和之前，通过一连串的3×3卷积进行降采样，其跨度为2。对于低到高的分辨率，低分辨率的特征图首先被1×1的卷积压缩，然后用双线性插值上升采样。图3显示了双边融合的实现方式。



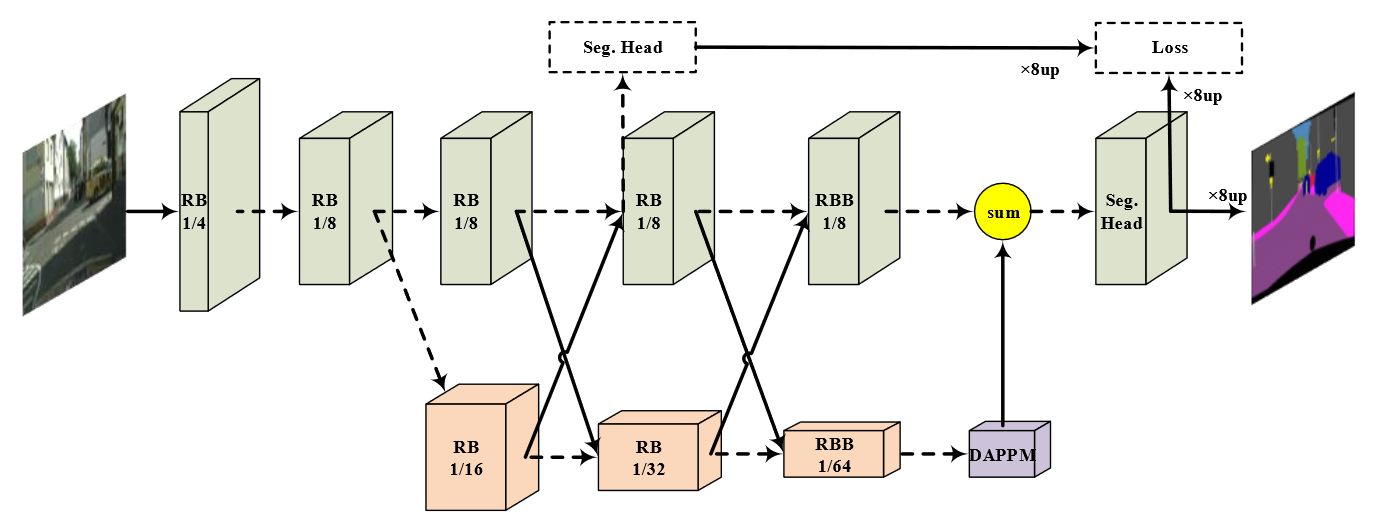
第i个高分辨率特征图XHi和低分辨率特征图XLi可以写成：。  
  
其中FH和FL对应于高分辨率和低分辨率的残余基本块序列，TL-H和TH-L指的是低到高和高到低的变换器，R表示ReLU函数。我们完全构建了四个不同深度和宽度的双分辨率网络。DDRNet-23的宽度是DDRNet-23-slim的两倍，DDRNet-39 1.5×也是DDRNet-39的一个加宽版本。  
B. 深度聚合金字塔池模块

这里，我们提出了一个新颖的模块，以进一步从低分辨率的特征图中提取上下文信息。图5显示了DAPPM的内部结构。



以1/64图像分辨率的特征图为输入，进行指数级的大池化核，生成1/128、1/256、1/512图像分辨率的特征图。输入的特征图和由全局平均池产生的图像级信息也被利用了。我们认为，通过单一的3×3或1×1卷积来混合所有的多尺度背景信息是不够的。受Res2Net[39]的启发，我们首先对特征图进行上采样，然后使用更多的3×3卷积来融合不同尺度的上下文信息，以分层-残留的方式。考虑到一个输入x，每个尺度yi可以写成。  
  
其中C1×1为1×1卷积，C3×3为3×3卷积，U表示上采样操作，Pj,k表示核大小为j、步长为k的池层，Pglobal表示全局平均池。最后，所有的特征图被串联起来，并使用1×1卷积进行压缩。此外，为了便于优化，还增加了一个1×1的投影快捷方式。与SwiftNet[21]中的SPP类似，DAPPM是通过BN-ReLU-Conv序列实现的。在DAPPM内部，由较大的池化核提取的语境与较深的信息流相结合，通过将不同深度的池化核与不同大小的池化核相结合，形成多尺度性质。表二显示，DAPPM能够提供比PPM更丰富的上下文。尽管DAPPM包含更多的卷积层和更复杂的融合策略，但由于输入分辨率仅为图像分辨率的1/64，所以几乎不影响推理速度。例如，对于1024×1024的图像，特征图的最大分辨率为16×16。  
C. 语义分割的整体架构

我们的方法概述见图4。



为了适应语义分割的任务，我们对双分辨率网络做了一些改变。首先，低分辨率分支的RBB中3×3卷积的跨度被设置为2，以进一步降采样。然后，在低分辨率分支的输出中加入DAPPM，从1/64图像分辨率的高层特征图中提取丰富的上下文信息。此外，最后的高-低融合被由双线性插值和求和融合实现的低-高融合所取代。最后，我们设计了一个简单的分割头，包括一个3×3卷积层和一个1×1卷积层。分割头的计算负荷可以通过改变3×3卷积层的输出尺寸来调整。我们对DDRNet-23-slim设定为64，对DDRNet-23设定为128，而对DDRNet39设定为256。请注意，除了分割头和DAPPM模块，所有的模块都在ImageNet上进行了预训练。

D. 深度监督

训练阶段的额外监督可以缓解深度卷积神经网络（DCNNs）的优化。在PSPNet中，加入了一个辅助损失来监督ResNet-101的res4 22块的输出，根据实验结果，相应的权重被设置为0.4[11]。BiSeNetV2[24]提出了一种助推器训练策略，即在语义分支的每个阶段的末尾添加额外的分割头。然而，它需要大量的实验来找到平衡每个损失的最佳权重，并导致训练记忆的不可忽略的增加。为了获得更好的结果，SFNet[23]利用了一个类似的策略，名为级联深度监督学习。在本文中，我们只采用简单的额外监督，以便与大多数方法进行公平的比较。如图4所示，我们添加了辅助损失，并按照PSPNet的方法设置权重为0.4。在测试阶段，辅助分割头被丢弃了。最终的损失是交叉熵损失的加权和，可以表示为。  
  
其中Lf、Ln、La分别代表最终损失、正常损失、辅助损失，α表示辅助损失的权重，本文中为0.4。

V. 结论

在本文中，我们致力于道路场景的实时和准确的语义分割，并提出了一个简单的解决方案，没有使用额外的铃声或口哨。特别是，我们提出了新的深度双分辨率网络作为实时语义分割的有效骨干。并设计了一个新的模块，用于从低分辨率的特征图中提取多尺度的背景信息。据我们所知，我们是第一个在实时语义分割中引入深度高分辨率表示的人，我们的简单策略在三个流行的基准上超过了所有以前的实时模型。DDRNets主要由剩余的基本块和瓶颈块组成，通过扩展模型的宽度和深度，提供了一个广泛的速度和准确性的权衡。由于我们方法的简单和高效，它可以被看作是统一实时和高准确度语义分割的一个强有力的基线。进一步的研究将集中在改进基线和将骨干转移到其他下游任务上。