Rethinking BiSeNet For Real-time Semantic Segmentation

摘要

BiSeNet[28, 27]已被证明是一种流行的用于实时分割的双流网络。然而，其增加一个额外的路径来编码空间信息的原理是很耗时的，而且由于缺乏特定任务的设计，从预训练的任务（如图像分类）中借用的骨干可能对图像分割是低效的。为了处理这些问题，我们提出了一种新的、高效的结构，名为短期密集串联网络（STDC网络），通过去除结构冗余。具体来说，我们逐步减少特征图的维度，并使用它们的聚合来表示图像，这构成了STDC网络的基本模块。在解码器中，我们通过将空间信息的学习以单流方式整合到低层，提出了一个细节聚合模块。最后，低层特征和深层特征被融合以预测最终的分割结果。在Cityscapes和CamVid数据集上进行的大量实验表明，我们的方法在分割精度和推理速度之间实现了很好的权衡。在Cityscapes上，我们在NVIDIA GTX 1080Ti上以250.4 FPS的速度在测试集上实现了71.9%的mIoU，比最新的方法快45.2%；在更高分辨率的图像上进行推断时，实现了76.8%的mIoU和97.0 FPS。代码可在https://github.com/ MichaelFan01/STDC-Seg。

1. 介绍

语义分割是计算机视觉中的一个经典和基本课题，其目的是在图像中分配像素级标签。深度学习的繁荣通过各种突破大大促进了语义分割的性能[18, 27, 22, 4]，在许多应用中，如自动驾驶、视频监控、机器人传感等，都有快速增长的需求。这些应用促使研究人员探索有效和高效的分割网络，特别是在移动领域。为了满足这些需求，许多研究人员提出设计低延迟、高效率的CNN模型，并具有令人满意的分割精度。这些实时语义分割方法在各种基准上都取得了很好的性能。对于实时推理，一些工作，例如DFANet[18]和BiSeNetV1[28]选择了轻量级骨干，并研究了特征融合或聚合模块的方式来弥补准确率的下降。然而，由于特定任务设计的缺陷，这些从图像分类任务中借来的轻量级骨干网对于图像分割问题可能并不完美。除了选择轻量级骨干，限制输入图像的大小也是另一个常用的方法来提高推理速度。较小的输入分辨率似乎是有效的，但它很容易忽略边界和小物体周围的详细外观。为了解决这个问题，如图2(a)所示，BiSeNet[28, 27]采用了多路径框架来结合低级细节和高级语义。然而，增加一个额外的路径来获得低层次的特征是非常耗时的，而且辅助路径总是缺乏低层次的信息指导。为此，我们提出了一种新型的手工网络，目的是为了提高推理速度、可解释的结构以及与现有方法相比具有竞争力的性能。首先，我们设计了一种新的结构，称为短期密集串联模块（STDC模块），以获得具有少数参数的变体可扩展感受野。然后，STDC模块被无缝集成到U-net架构中，形成STDC网络，这大大促进了网络在语义分割任务中的表现。具体来说，如图3所示，我们将多个连续层的响应图串联起来，每个层都以不同的尺度和各自的领域对输入图像/特征进行编码，从而形成多尺度的特征表示。为了加快速度，各层的过滤器大小逐渐减少，分割性能的损失可以忽略不计。STDC网络的详细结构可以在表2中找到。在解码阶段，如图2(b)所示，我们没有使用额外的耗时路径，而是采用了细节指导来指导低层的空间细节学习。我们首先利用细节聚合模块来生成细节的基础事实。然后，采用二元交叉熵损失和骰子损失来优化细节信息的学习任务，这被认为是侧面信息学习的一种类型。需要注意的是，这种侧面信息在推理时是不需要的。最后，来自低层的空间细节和来自深层的语义信息被融合，以预测语义分割的结果。我们方法的整个架构如图4所示。我们的主要贡献可以概括为以下几点。- 我们设计了一个短时密集串联模块（STDC模块）来提取具有可扩展的感受野和多尺度信息的深度特征。该模块以可承受的计算成本提升了我们的STDC网络的性能。- 我们提出了细节聚合模块来学习解码器，从而更精确地保留低层的空间细节，而不需要推理时间的额外计算成本。- 我们进行了广泛的实验来展示我们方法的有效性。实验结果表明，STDC网络在ImageNet、Cityscapes和CamVid上取得了新的技术成果。具体来说，我们的STDC1-Seg50在Cityscapes测试集上实现了71.9%的mIoU，在一块NVIDIA GTX 1080Ti卡上的速度为250.4 FPS。在同样的实验设置下，我们的STDC2-Seg75在97.0FPS的速度下实现了76.8%的mIoU。

3. 提出的方法

BiSeNetV1[28]利用轻量级骨干网，如ResNet18和空间路径作为编码网络，形成双流分割架构。然而，由于结构冗余，分类骨干网和双流架构可能是低效的。在本节中，我们首先介绍我们提出的STDC网络的细节。然后，我们介绍我们的单流方法的整个架构，并提供详细的指导。

3.1. 编码网络的设计

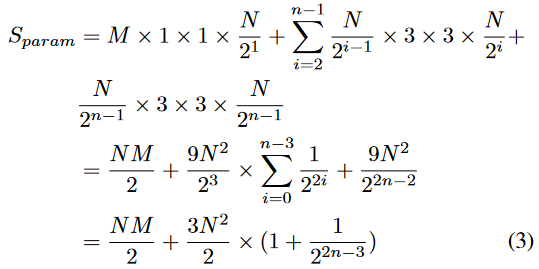
3.1.1 短时密集串联模块 我们提出的网络的关键部分是短时密集串联模块（STDC模块）。图3（b）和（c）说明了STDC模块的布局。具体来说，每个模块被分成几个块。我们用ConvXi表示第i块的操作。因此，第i块的输出计算如下：



其中xi-1和xi分别为第i块的输入和输出。ConvX包括一个卷积层、一个批量归一化层和ReLU激活层，ki是卷积层的核大小。在STDC模块中，第一个块的核大小为1，其余的简单设置为3。 给定STDC模块输出的通道数N，第i个块中的卷积层的滤波器数量为N/2i，除了最后一个卷积层的滤波器，其数量与前一个卷积层的相同。在图像分类任务中，常见的做法是在较高的层中使用更多的通道。但在语义分割任务中，我们关注的是可扩展的感受野和多尺度信息。低层需要足够的通道来编码具有小接收域的更精细的信息，而具有大接收域的高层则更注重高层信息的归纳，在低层设置相同的通道可能会造成信息冗余。Down-sample只发生在Block2中。为了丰富特征信息，我们通过跳过路径将x1到xn的特征图连接起来，作为STDC模块的输出。在串联之前，STDC模块中不同区块的响应图通过3×3的平均池化操作下采样到相同的空间大小，如图3（c）所示。在我们的设定中，STDC模块的最终输出是：



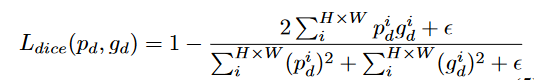
其中，xoutput表示STDC模块的输出，F是我们方法中的融合操作，而x1, x2, ..., xn是所有n个区块的特征图。出于对效率的考虑，我们采用串联法作为我们的融合操作。在我们的方法中，我们使用4个区块的STDC模块。表1列出了STDC模块中各区块的感受野，因此xoutput收集了所有区块的多尺度信息，我们声称我们的STDC模块有两个优点：（1）我们通过以几何级数的方式逐渐减少，精心调整了区块的过滤器大小，导致计算复杂性的显著降低。(2）STDC模块的最终输出是由所有块串联而成的，它保留了可扩展的各自领域和多尺度信息。给定输入通道尺寸M和输出通道尺寸N，STDC模块的参数数为：



如公式3所示，STDC模块的参数数量由预定的输入和输出通道尺寸主导，而块的数量对参数大小的影响很小。特别是，如果n达到最大限度，STDC模块的参数数几乎保持不变，它只由M和N定义。

3.1.2 网络结构   
我们在图3（a）中展示了我们的网络结构。它由6个阶段组成，除了输入层和预测层。一般来说，阶段1∼5分别对输入的空间分辨率进行下采样，跨度为2，阶段6通过一个ConvX、一个全局平均池层和两个全连接层输出预测对数。阶段1和阶段2通常被视为外观特征提取的低级层。为了追求效率，我们只在第一阶段和第二阶段各使用一个卷积块，根据我们的经验，这已经足够了。在我们的网络中，第三、四、五阶段的STDC模块的数量是经过仔细调整的。在这些阶段中，每个阶段中的第一个STDC模块以2的步长对空间分辨率进行下采样，每个阶段中的后续STDC模块保持空间分辨率不变。我们把阶段的输出通道数表示为Nl，其中l是阶段的索引。在实践中，我们根据经验将N6设定为1024，并仔细调整其余阶段的通道数，直到在精度和效率之间达到一个良好的平衡点。由于我们的网络主要由短期密集串联模块组成，我们称我们的网络为STDC网络。表2显示了我们的STDC网络的详细结构。  
3.2. 解码器的设计   
3.2.1 分割结构 我们使用预先训练好的 STDC 网络作为编码器的主干， 并采用 BiSeNet[28]的上下文路径来编码上下文信息。如图4(a)所示，我们使用阶段3、4、5分别产生下采样率为1/8、1/16、1/32的特征图。然后，我们使用全局平均池来提供具有大感受野的全局背景信息。U型结构被用来对全局特征进行上采样，并在我们的编码阶段将它们与后两个阶段（第四和第五阶段）的对应特征结合起来。按照BiSeNet[28]的做法，我们使用注意精化模块来精化每两个阶段的组合特征。对于最终的语义分割预测，我们采用BiSeNet[28]中的特征融合模块来融合编码器中第三阶段的1/8下采样特征和解码器的对应特征。我们声称，这两个阶段的特征处于不同的特征表示水平。来自编码骨干的特征保留了丰富的细节信息，而来自解码器的特征由于来自全局池化层的输入而包含了上下文信息。具体来说，Seg Head包括一个3×3的Conv-BN-ReLU算子，然后是1×1的卷积，得到输出维度N，它被设定为类的数量。我们采用交叉损失和在线硬例挖掘来优化语义分割学习任务。3.2.2 低层特征的详细指导 我们在图5(b)中可视化了BiSeNet的空间路径的特征。与相同下采样率的骨干低层（阶段3）相比，空间路径可以编码更多的空间细节，如边界、角落。基于这一观察，我们提出了一个细节指导模块，引导低层以单流方式学习空间信息。我们将细节预测建模为一个二元分割任务。如图4(c)所示，我们首先通过拉普拉斯算子从分割的基础实况中生成细节图的基础实况。如图4(a)所示，我们在第三阶段插入细节头以生成细节特征图。然后，我们用细节地面实况作为细节特征图的指导，引导低层学习空间细节特征。如图5(d)所示，与图5(c)所示的结果相比，有细节指导的特征图可以编码更多的空间细节。最后，学习到的细节特征与来自解码器深层块的上下文特征相融合，进行分割预测。细节真实生成。我们通过细节聚合模块从语义分割的基础事实中生成二进制细节基础事实，如图4(c)中的蓝色虚线框所示。这一操作可以通过名为拉普拉斯核的二维卷积核和可训练的1×1卷积来进行。我们使用图4(e)所示的拉普拉斯算子来产生具有不同步长的软细部特征图，以获得多尺度的细节信息。然后，我们对细节特征图进行上采样，使其达到原始尺寸，并与可训练的1×1卷积相融合，进行动态重构。最后，我们采用0.1的阈值，将预测的细节转换为最终的二进制细节地面实况，包括边界和角落信息。细节损失：由于细节像素的数量比非细节像素少得多，细节预测是一个类平衡问题。由于加权交叉熵总是导致粗略的结果，按照[7]，我们采用二元交叉熵和骰子损失来共同优化细节学习。骰子损失衡量预测地图和地面实况之间的重叠。同时，它对前景/背景像素的数量不敏感，这意味着它可以缓解类不平衡的问题。因此，对于高度为H、宽度为W的预测细节图，细节损失Ldetail的计算方法如下。  


其中pd∈RH×W表示预测的细节，gd∈RH×W表示相应的细节地面实况。Lbce表示二进制交叉熵损失，而Ldice表示骰子损失，具体如下。



其中i表示第i个像素，ǫ是一个拉普拉斯平滑项，以避免零分。在本文中，我们设定ǫ=1。

如图4(b)所示，我们使用细节头来制作细节图，引导浅层的空间信息编码。细节头包括一个3×3的Conv-BN-ReLU算子，然后是一个1×1的卷积，得到输出的细节图。在实验中，Detail Head被证明能有效地增强特征表示。请注意，这个分支在推理阶段被丢弃了。因此，这个侧面信息可以很容易地提高分割任务的准确性，而无需在推理中付出任何代价。

5. 结论 在本文中，我们重新审视了经典的分割架构BiSeNet[28, 27]的结构优化。一般来说，BiSeNet的分类骨架和额外的空间路径大大阻碍了推理效率。因此，我们提出了一个新颖的短时密集串联模块，以提取具有可扩展的感受野和多尺度信息的深度特征。基于该模块，我们设计了STDC网络，并在图像分类中实现了具有竞争力的准确性和高FPS。以STDC网络为骨干，我们的细节引导的STDC-Seg在实时语义分割中实现了最先进的速度和准确性的权衡。大量的实验和可视化结果表明了我们提出的STDC-Seg网络的有效性。在未来，我们将通过以下方向扩展我们的方法：

(i)骨干将在更多的任务中得到验证，例如物体检测。

(ii)我们将更深入地探索空间边界在语义分割任务中的利用

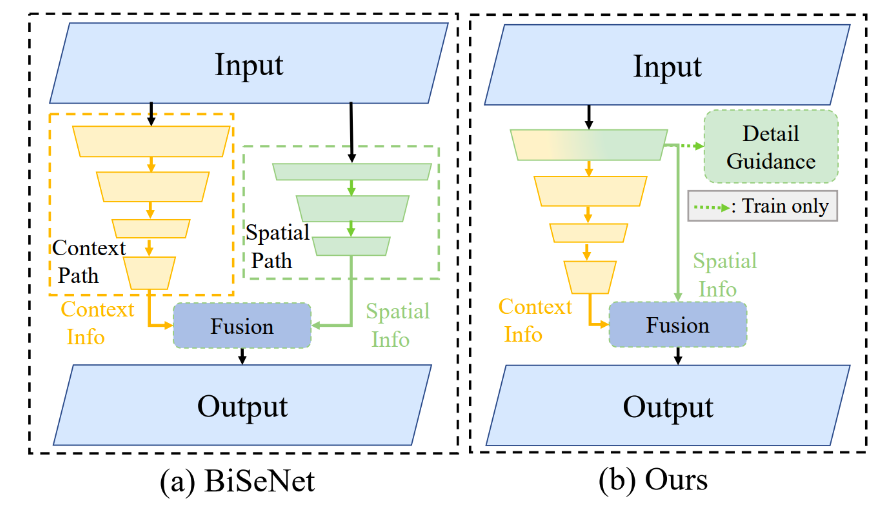


图2. BiSeNet[28]和我们提出的方法的架构说明。(a)展示了双边分割网络(BiSeNet[28])，它使用一个额外的空间路径来编码空间信息。(b)展示了我们提出的方法，它使用一个细节指导模块来编码低级特征中的空间信息，而不需要额外的耗时路径。

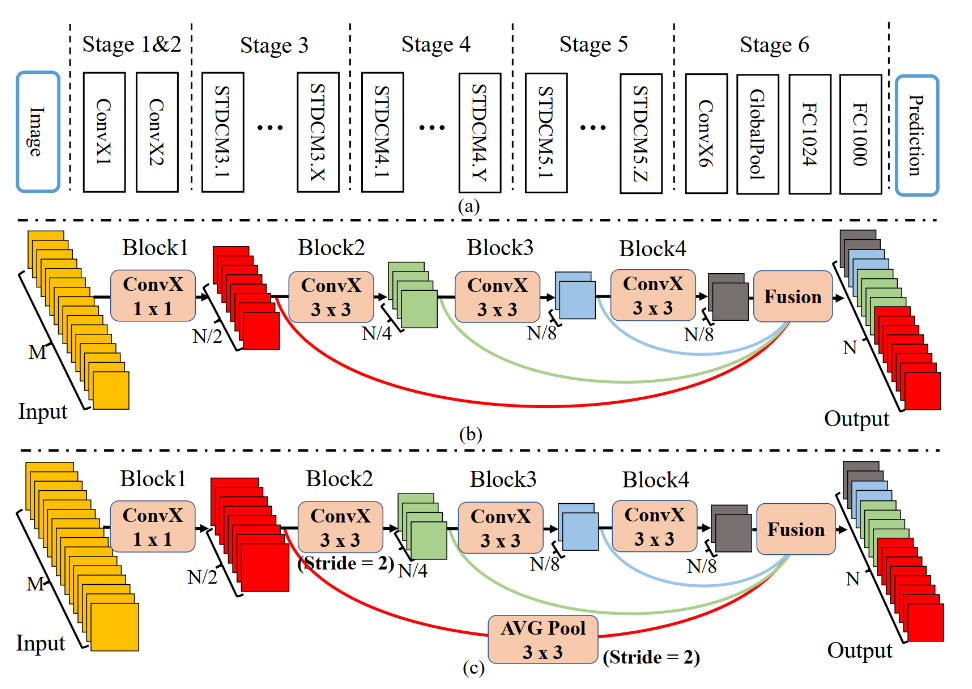


图3. (a) 一般STDC网络结构。ConvX操作是指Conv-BN-ReLU。(b) 我们网络中使用的短时密集串联模块（STDC模块）。M表示输入通道的尺寸，N表示输出通道的尺寸。每个区块是一个具有不同内核大小的ConvX操作。(c) stride=2的STDC模块。

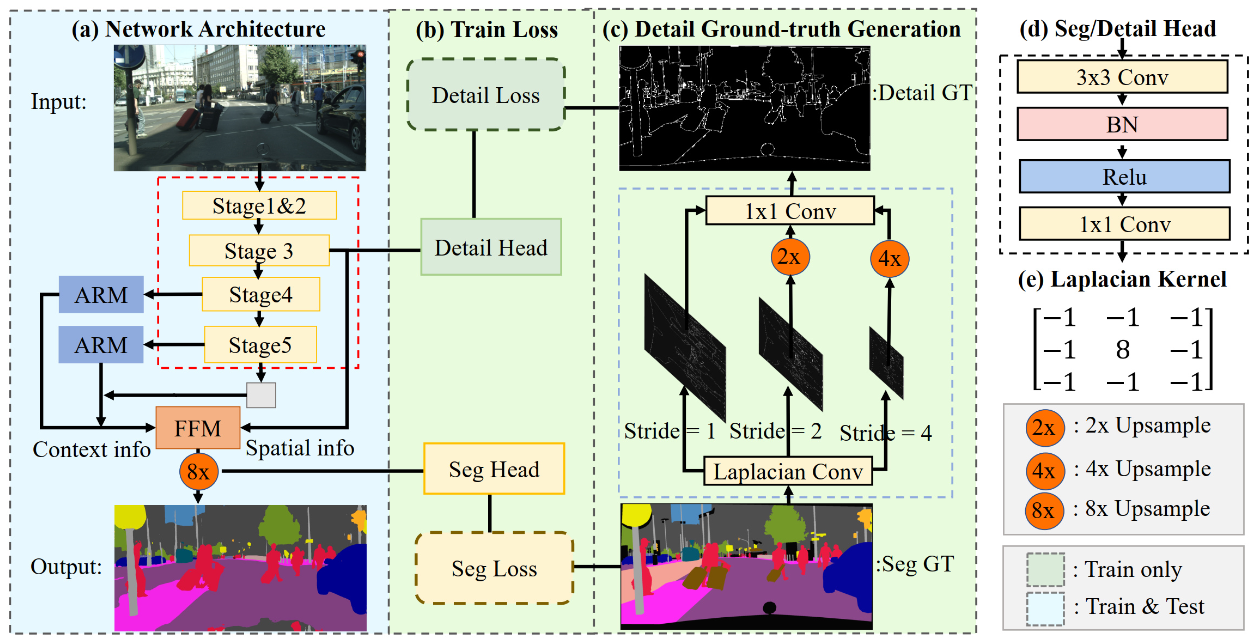


图4. STDC分割网络的概述。ARM表示注意精化模块，FFM表示[28]中的特征融合模块。红色虚线框内的操作是我们的STDC网络。蓝色虚线框内的操作是细节聚合模块。

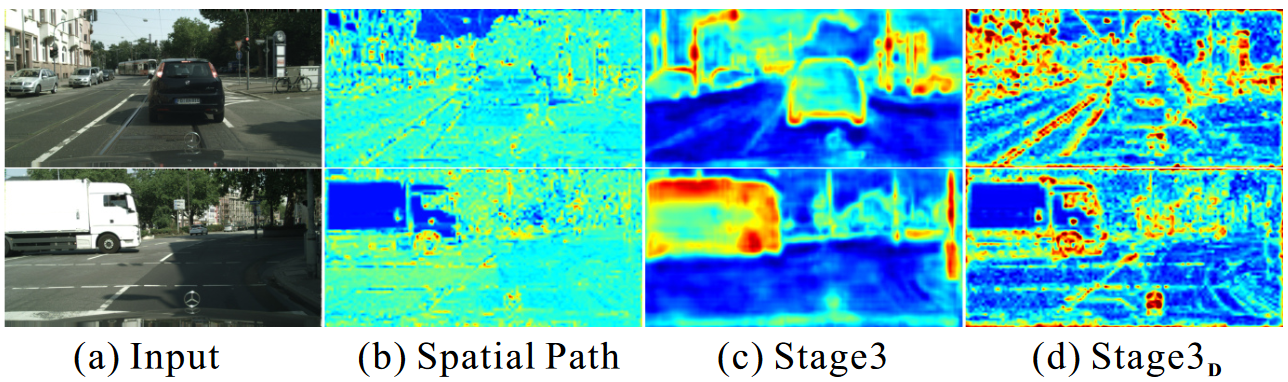


图5. 在没有和有详细指导的情况下，对空间路径和第三阶段的特征进行视觉解释。带下标D的一列表示有细节指导的结果。可视化显示，空间路径可以比骨干网的低层编码更多的空间细节，如边界、角落，而我们的细节指导模块可以做同样的事情，没有额外的计算成本。