条带池化层: 重新思考场景分析中的空间池化

Qibin Hou[1] Li Zhang[2] Ming-Ming Cheng[3] Jiashi Feng[1]
[1]National University of Singapore [2]University of Oxford [3]CS, Nankai University

Abstract

空间池化已经被证明在像素级预测任务中获取长远上下文信息时非常有效，例如场景解析。在本文中，我们在常规形状 $N \times N$ 的空间池化的基础上，通过引入一种新的称为条带池化层 (Strip Pooling) 的池化策略来重新思考空间池化的工作方式。它采用一种长而窄的池化核，即 $1 \times N$ 或 $N \times 1$。基于条带池化层，我们进一步研究了空间池化结构的设计：1) 引入一个新的条带池化模块，使主干网络能够有效地建模长远依赖，2) 设计以多种空间池化为核心的新型网络结构模块，3) 系统地比较了和传统空间池化技术的性能。这两种新的基于池化的设计都是轻量级的，可以在现有的场景解析网络中作为一个高效的即插即用模块。在流行的基准 (如 ADE20K 和 Cityscape) 上的广泛实验表明，我们的简单方法实现了最佳结果。代码已开源在https://github.com/Andrew-Qibin/SPNet。

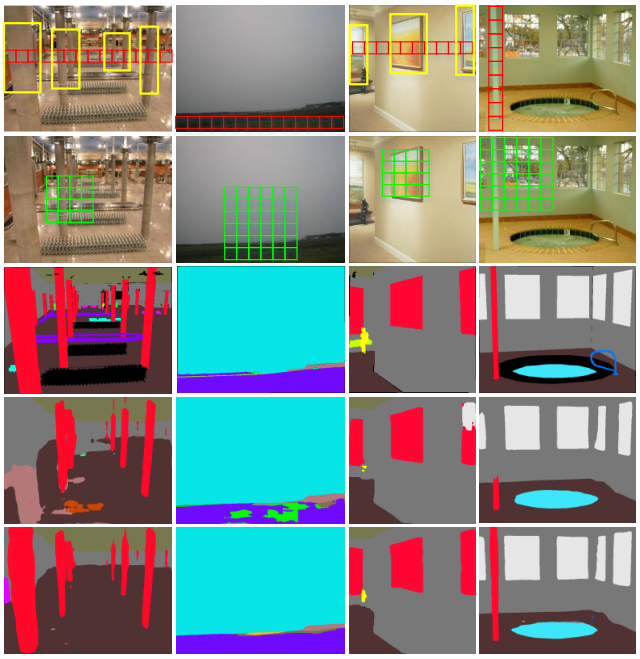
1. Introduction

场景解析也称为语义分割，其目的是为图像中的每个像素分配一个语义标签。作为最基本的任务之一，它已被应用于各种计算机视觉和图形应用程序 [10]，如自动驾驶 [47]，医学诊断 [46]，图像/视频编辑 [41, 27]，显著目标检测 [3]，以及航空图像分析 [38]。近年来，基于全卷积网络 (FCN) 的方法 [37, 5] 在场景解析方面取得了显著的进展，它们具有捕获高级语义的能力。然而，这些方法大多是堆叠局部卷积和池化操作，由于有效视野有限 [65, 23]，因而难以很好地处理各种不同类别的复杂场景。

一种提高 CNN 长远依赖关系建模能力的方法是采用自注意力或 non-local 模块 [51, 23, 7, 45, 21, 53, 66, 62, 61, 28]。然而，在每个空间位置计算大型关联矩阵会消耗巨大的内存。其他用于长远上下文建模的方法包括：空洞卷积 [5, 8, 6, 57]，目的是在不引入额外参数的情况下扩大 CNN 感受野；或全局/金字塔池化 [26, 65, 19, 5, 8, 54] 捕获图像的全局信息。然而，这些方法 (包括空洞卷积和池化) 的一个共同限制是，它们都在方形窗口内探测输入特征图。这限制了它们捕捉现实场景中广泛存在的各向异性上下文的灵活性。例如，在某些情况下，目标物可能具有长条带状结构 (例如图1b 中的草地) 或离散地分布 (例如图1a 中的柱子)。使用大的方形池化窗口并不能很好地解决这个问题，因为它不可避免地会合并来自不相关区域的噪声信息 [19]。

在本文中，为了更有效地捕获长远依赖关系，我们利用空间池化来扩大 CNN 的感受野和收集有效的上下文信息，并提出了条带池化的概念。作为全局池化的替代方案，条带池化有两个优点。其一，它沿着一个空间维度部署池化核，从而能够捕获孤立区域的长远关系，如图1a 和1c 的顶部所示。其二，它在另一空间维度上保持一个狭窄的核形状，有利于捕获局部上下文，防止不相关区域干扰标签预测。集成这种长而窄的池化核使场景解析网络能够同时聚合全局上下文和局部上下文。这与传统的从固定的方形区域收集上下文的空间池化有本质上的不同。

在条带池化操作的基础上，我们提出了两个基于池化的场景解析网络模块。首先，我们设计了一个 Strip Pooling Module (SPM) 来有效地扩大主干网络的感受野。更具体地说，SPM 由两条路径组成，



(a) (b) (c) (d)

图 1. 条带池化和空间池化在场景解析中如何不同工作的展示。从上到下：条带池化；传统的空间池化；真实标注；我们的结果 (使用传统的空间池化)；我们的结果 (使用条带池化)。第一行所示，与传统的空间池化 (绿色网格) 相比，条带池化具有带形核 (红色网格)，因此可以捕获离散分布的区域 (黄色边界框) 之间的长远依赖关系。

它们侧重于沿着水平或垂直空间维度编码长远距离上下文。对于池化后特征图中的每个空间位置，它对其全局水平和垂直信息进行编码，然后使用这些编码来平衡自己的权重以进行特征细化。此外，我们提出了一个新的附加残差结构块，称为 Mixed Pooling module(MPM)，以进一步在高级别语义上建模长远依赖。它通过利用不同核形状的池化操作来收集丰富的上下文信息，从而处理具有复杂场景的图像。为了证明所提出的基于池化的模块的有效性，我们提出了 SPNet，它将两个模块合并到 ResNet [20] 主干网络中。实验表明，我们的 SPNet 在流行的场景解析基准测试中建立了新佳结果。

本文工作贡献如下：(1) 我们研究了空间池化的传统设计，提出了条带池化的概念，它继承了全局平均池化的优点，在收集长远依赖的同时关注局部细节。(2) 我们设计了一个基于条带池化的 Strip Pooling Module 和一个 Mixed Pooling Module。这两个模块都是轻量级的，可以作为有效的附加模块插入到任何主干网络中，以生成高质量的分割预测。(3) 我们提出了将上述两个基于池化的模块集成到一个单一体系结构中的 SPNet，它在基线上取得了显著的改进，并在广泛使用的场景解析基准数据集上实现了最好的结果。

2. 相关工作

目前最先进的场景解析 (或语义分割) 方法主要利用卷积神经网络 (CNN)。然而，通过叠加局部卷积或池化操作，CNN 的感受野增长缓慢，因此阻碍了它们考虑足够有用的上下文信息。早期的场景解析上下文关系建模技术包括条件随机场 (CRFs) [25, 49, 1, 67]。它们大多在离散标签空间中建模，计算成本昂贵，因此，尽管已经集成到 CNN 中，但现在在生成最先进的场景解析结果方面不太成功。

对于连续特征空间学习，之前工作通过在多比率和多视野下卷积或池化探测特征的方式，使用多尺度特征聚合 [37, 5, 33, 18, 42, 31, 32, 2, 44, 4, 48, 17] 来融合上下文信息。DeepLab [5, 6] 及其后续工作 [8, 54, 39] 采用空洞卷积并融合不同的膨胀比特征，以增加网络的感受野。此外，聚合 non-local 上下文 [36, 58, 29, 15, 7, 45, 21, 53, 66, 23, 14] 对于场景解析也是有效的。

另一个改善感受野的研究方向是空间金字塔池化 [65, 19]。通过在每个金字塔层采用一组具有唯一内核大小的并行池化操作，网络能够捕获大范围的上下文信息。在几个场景解析基准测试中，它显示出了良好的前景。然而，它利用上下文信息的能力是有限的，因为只应用正方形核形状。而且，空间金字塔池化只是在主干网络的顶部进行了模块化，使得它不能灵活或直接应用于特征学习的网络结构块中。相比之下，我们提出的 strip pooling 模块和 mixed pooling 模块采用大小为 $1 \times N$ 或 $N \times 1$ 的池化内核，这两种池化内核都可以插入并堆叠到现有网络中。这种差异使得网络能够在每个所提出的结构块中利用丰富的上下文关系。在我们的实验中，所提出的模块已经被证明比空间金字塔池化更强大且适应性更强。

3. 方法

在本节中，我们首先给出了条带池化的概念，然后介绍了基于条带池化的两种模型设计，以说明它是如何改进场景解析网络的。最后，描述了条带池化增强的场景解析网络的整体结构。

3.1. 条带池化

在描述条带池化的表达之前，我们首先简单回顾一下平均池化操作。

标准空间平均池化：设 $\mathrm{x} \in \mathbb{R}^{H \times W}$ 是一个二维输入张量，其中 H 和 W 分别为空间高度和宽度。在平均池化层中，池化的空间范围 $(h \times w)$ 是必需的。考虑一个简单的例子，H 可被 h 整除，W 可被 w 整除。那么池化后输出的 y 也是一个高度为 $H_o = \frac{H}{h}$ 和宽度为 $W_o = \frac{W}{w}$ 的二维张量。形式上，平均池化操作可以写为

$$y_{i_o,j_o} = \frac{1}{h \times w} \sum_{0 \le i < h} \sum_{0 \le j < w} x_{i_o \times h + i, j_o \times w + j}, \qquad (1)$$

其中 $0 \le i_o < H_o$，$0 \le j_o < W_o$。在式1中，y 的每个空间位置对应一个大小为 $h \times w$ 的池化窗口。上述池化操作已成功应用于之前的工作 [65, 19] 中，用于收集长远上下文。但是，在处理形状不规则的物体时，如图1所示，不可避免地会合并很多不相关的区域。

条带池化：为了缓解上述问题，我们在这里提出了“条带池化”的概念。它使用一个条带形状的池化窗口沿水平或垂直维度进行池化，如图1的第一行所示。数学上，给定二维张量 $\mathrm{x} \in \mathbb{R}^{H \times W}$，在条带池化中，需要池化 $(H, 1)$ 或 $(1, W)$ 的空间范围。与二维平均池化不同，条带池化将每行或每列中的所有特征值平均。因此，水平条带池化后的输出 $\mathrm{y}^h \in \mathbb{R}^H$ 可以写成

$$y_i^h = \frac{1}{W} \sum_{0 \le j < W} x_{i,j}. \qquad (2)$$

类似地，垂直条带池化后的输出 $\mathrm{y}^v \in \mathbb{R}^W$ 可以写成

$$y_j^v = \frac{1}{H} \sum_{0 \le i < H} x_{i,j}. \qquad (3)$$

考虑水平和垂直的条带池化层，由于核形状长而窄，很容易在离散分布的区域之间建立长远依赖关系，并使用带状编码区域。同时，由于它在其他维度上的核形状较窄，它也注重捕捉局部细节。这些特性使得条带池化不同于传统的依赖于方形核的空间池化。以下，我们将描述如何利用条带池化 (式2和式3) 来改进场景解析网络。

3.2. Strip Pooling Module

在以前的工作 [8, 16] 中已经证明，扩大主干网络的感受野有利于场景解析。在这一小节中，我们将介绍一种有效的方法，通过利用条带池化来帮助主干网络捕获长远上下文。特别地，我们提出了一个新颖的条带池化模块 (SPM)，它利用水平和垂直的条带池化操作从不同的空间维度收集长远上下文。图2描述了我们提出的 SPM。设 $\mathrm{x} \in \mathbb{R}^{C \times H \times W}$ 为输入张量，其中 C 表示通道数。我们首先将 x 输入两个并行路径，每条路径都包含一个水平或垂直的条带池化层，然后是一个核大小为 3 的一维卷积层，用于调制当前位置及其邻近特征。这样就得到 $\mathrm{y}^h \in \mathbb{R}^{C \times H}$ 和 $\mathrm{y}^v \in \mathbb{R}^{C \times W}$。为了获得包含更多有用全局先验的输出 $\mathrm{z} \in \mathbb{R}^{C \times H \times W}$，我们首先将 y^h 和 y^w 组合在一起，如下所示，得到 $\mathrm{y} \in \mathbb{R}^{C \times H \times W}$：

$$y_{c,i,j} = y_{c,i}^h + y_{c,j}^v. \qquad (4)$$

然后，输出 z 由下式计算

$$\mathrm{z} = \mathrm{Scale}(\mathrm{x},\ \sigma(f(\mathrm{y}))), \qquad (5)$$

其中 $\mathrm{Scale}(\cdot, \cdot)$ 指元素乘法，σ 为 sigmoid 函数，f 是一个 1×1 卷积。需要注意的是，将两个条带池化层提取的特征进行组合有多种方法，如计算提取的两个一维特征向量之间的内积。然而，考虑到效率和使 SPM 轻量化，我们采用了上述操作，我们发现这些操作仍然很有效。

在上述过程中，允许输出张量中的每个位置与输入张量中的各种位置建立关系。例如，在图2中，输出张量中以黑色框为边界的正方形连接到与其具有相同水平或垂直坐标的所有位置 (由红色框和紫色框包围)。因此，通过多次重复上述聚合过程，就

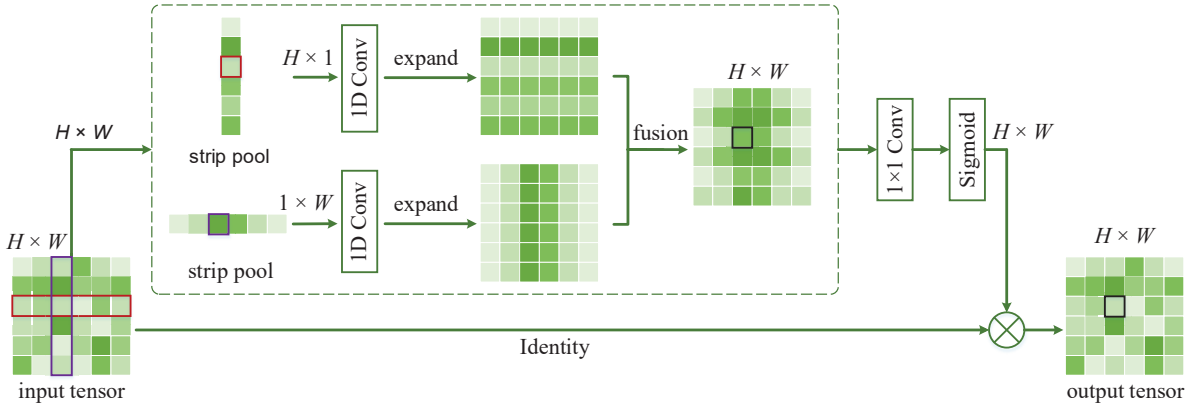


图 2. Strip Pooling (SP) 模块示意图。

有可能在整个场景中构建长远依赖关系。此外，得益于元素乘法运算，SPM 还可以被视为一种注意力机制，可直接应用于任何预训练的主干网络中，而无需从头开始训练。

与全局平均池化相比，条带池化考虑的是长而窄的范围，而不是整张特征图，避免了在相距较远的位置之间建立太多不必要的连接。与需要大量的计算来建立每一对位置之间的关系的基于注意力的模型 [16, 19] 相比，我们的 SPM 是轻量级的，可以轻松地嵌入到任何结构块中，以提高捕获长远空间依赖关系和利用通道间依赖关系的能力。我们将提供更多关于我们的方法相对于现有的基于注意力的方法的性能分析。

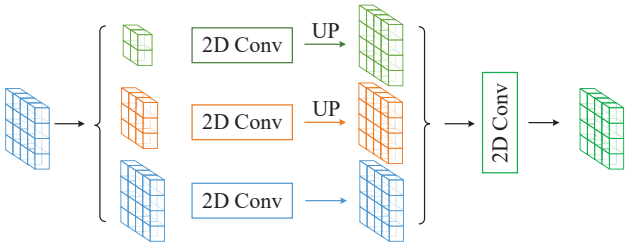
3.3. 混合池化模块

结果表明，金字塔池化模块 (PPM) 是增强场景解析网络 [65] 的有效方法。然而，PPM 严重依赖于标准的空间池化操作 (尽管在不同的金字塔级别使用不同的池化内核)，这使得它仍然像第3.1小节中分析的那样受到影响。考虑到标准空间池化和提出的条带池化的优点，我们改进了 PPM，并设计了混合池化模块 (MPM)，其通过各种池化操作聚合不同类型的上下文信息以使特征表示更具甄别性。

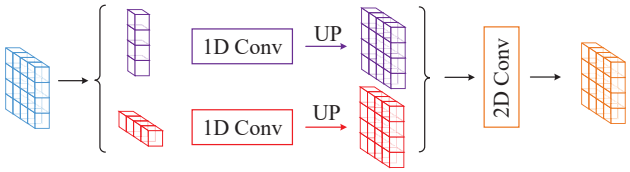
该模型由两个子模块组成，可同时捕获不同位置间的短程和长远依赖关系，这两个子模块对场景解析网络来说都是必不可少的。对于长远依赖，不像之前使用全局平均池化层的工作 [60, 65, 8]，我们提出通过使用水平和垂直的条带池化操作来收集此类线索。一个简单示意图由图3(b) 所示。正如第3.2小节分析的，条带池化使整个场景中离散分布的区域之间建立连接，并使带状结构的编码区域成为可能。但是，对于语义区域分布紧密的情况，空间池化对于捕获局部上下文信息也是必要的。考虑到这一点，如图3(a) 所示，我们采用轻量级的金字塔池化子模块进行短距离依赖收集。它有两个空间池化层，后接用于多尺度特征提取的卷积层和用于原始空间信息保存的二维卷积层组成。每张池化后的特征图大小分别为 20×20 和 12×12。所有三条子路径通过求和合并。

在上述两个子模块的基础上，我们提出将它们嵌套到具有瓶颈结构的残差块 [20] 中，以进行参数约简和模块设计。具体来说，在每个子模块之前，首先使用 1×1 卷积层进行通道缩减。两个子模块的输出被连接在一起，然后送到另一个 1×1 卷积层，用于通道扩展，正如在 [20] 中所做的那样。注意所有的卷积层，除了用于通道缩减和扩展的层，内核大小为 3×3 或 3(对于 1D 卷积层)。

值得一提的是，与空间金字塔池化模块 [65, 8] 不同，所提出的 MPM 是一种模块化设计。其优点是可以方便地按顺序使用它来扩展长远依赖集合子模块。我们发现，在相同的主干网络下，我们的网络只有两个 MPM(大约是原始 PPM[65] 的 1/3 个参数)，性能甚至比 PSPNet 更好。在我们的实验部分，我们将提供更多的结果和分析。



(a)



(b)

图 3. (a) 短程依赖聚合子模块。(b) 长远依赖聚合子模块。受 [34, 35] 的启发，在每个子模块的融合操作后加入卷积层，以减少降采样操作带来的混叠效应。

3.4. 整体架构

在提出的 SPM 和 MPM 的基础上，我们在本小节中介绍了一个总体架构，称为 SPNet。我们采用经典的残差网络 [20] 作为我们的主干网络。遵循 [5, 65, 16]，我们用空洞策略改进了原始的 ResNet，最终的特征图大小设置为输入图像的 1/8。SPM 被添加在每个阶段的最后一个结构块和最后一个阶段的所有结构块的 3×3 卷积层之后。SPM 中的所有卷积层共享与输入张量相同的通道数。

对于 MPM，由于其模块化设计，我们直接将其构建在主干网络上。由于主干网络的输出有 2048 个通道，我们首先连接一个 1×1 卷积层到主干网络，以将输出通道从 2048 减少到 1024，然后添加两个 MPM。在每个 MPM 中，遵循 [20]，所有具有内核大小 3×3 或 3 的卷积层有 256 个通道 (即下采样率为 1/4)。在最后添加一个卷积层来预测分割图。

4. 实验

我们在流行的场景解析数据集上评估了所提出的 SPM 和 MPM，包括 ADE20K [68]，Cityscapes [11]，以及 Pascal Context [40]。此外，我们还根据 ADE20K 数据集中的做法 [65] 对 strip pooling 的影响进行了综合消融分析。

Settings	#Params	SPM	mIoU	Pixel Acc
Base FCN	27.7 M	✗	37.63	77.60%
Base FCN + PPM [65]	+21.0 M	✗	41.68	80.04%
Base FCN + 1 MPM	+4.4 M	✗	40.50	79.60%
Base FCN + 2 MPM	+8.8 M	✗	41.92	80.03%
Base FCN + 2 MPM	+11.9 M	✓	44.03	80.65%

表 1. 混合池化模块 (MPM) 的消融分析。'SPM' 指条带池化模块。可以看到，当使用更多 MPM 时，可取得了更好的结果。所有结果均基于 ResNet-50 主干网络和单模型测试。最好的结果以粗体高亮。

4.1. 实验设置

我们的网络是基于两个公共的工具包 [64, 59] 及 Pytorch [43] 实现的。我们使用 4 GPU 来训练所有的实验。在训练过程中，Cityscapes 的批大小设置为 8，其他数据集的批大小设置为 16。遵循之前的工作 [5, 65, 60]，我们采用 'poly' 学习率策略 (即以 1 为底乘以 $(1-\frac{iter}{max\_iter})^{power}$)。ADE20K 与 Cityscapes 数据集的基础学习率设为 0.004，Pascal Context 设为 0.001。指数设为 0.9。训练轮数设为：ADE20K (120), Cityscapes (180), 以及 Pascal Context (100)。动量与权重衰减率分别设为 0.9 和 0.0001。我们在训练中采用了 [60, 65] 中的同步批归一化。

对于数据增强，与 [65, 60] 类似，我们随机翻转和缩放 0.5 倍至 2 倍输入图像，最终裁剪图像至固定的尺寸，对于 Cityscapes 是 768×768，对于其它数据集是 480×480。默认情况下，我们报告结果的标准评估指标为平均交并比 (mIoU)。对于没有真实标注的测试数据集，我们从官方评估服务器获得结果。在所有的实验中，我们使用交叉熵损失来优化所有的模型。遵循 [65]，我们利用了一个辅助损失 (连接到第四阶段的最后残差块)，损失权重设置为 0.4。我们还报告了多模型的结果，以公平地比较我们的方法与其他方法，即从多个图像尺度 $\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$ 平均分割概率图，正如 [32, 65, 60] 中的做法。

Settings	SPM	mIoU	Pixel Acc
Base FCN	✗	37.63	77.60%
Base FCN + 2 MPM (SRD only)	✗	40.50	79.34%
Base FCN + 2 MPM (LRD only)	✗	41.14	79.64%
Base FCN + 2 MPM (SRD + LRD)	✗	41.92	80.03%
Base FCN + 2 MPM (SRD + LRD)	✓	44.03	80.65%

表 2. 混合池化模块 (MPM) 消融分析。'SPM' 指 strip pooling 模块。'SRD' 和 'LRD' 分别表示短程依赖聚合子模块和长远依赖聚合子模块。可以看出，为了得到更好的分割结果，同时收集短程和长远依赖关系是必要的。所有结果均为单模型测试。

4.2. ADE20K

ADE20K 数据集 [68] 是最具挑战性的基准测试之一，它包含 150 个类和各种场景，有 1038 个图像级标签。我们遵循官方协议划分整个数据集。像以前的大多数工作一样，我们使用像素精度 (Pixel Acc.) 和平均交并比值 (mIoU) 来评估。我们还遵循 [32, 65] 采用多模型测试，并使用平均结果进行评估。在消融实验中，我们采用 [65] 中的做法，以 ResNet-50 作为我们的主干网络。在与之前的工作相比时，我们使用的是 ResNet-101。

4.2.1 消融实验

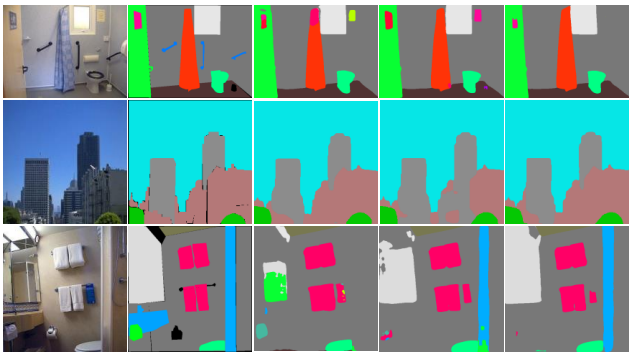
MPM **的数量**: 正如第3.3小节所述，MPM 是基于残差块的瓶颈结构构建的，因此可以很容易地重复多次扩大条带池化的作用。在这里，我们研究了需要多少 MPM 来平衡所提出方法的性能和运行时间成本。如表1所示，我们列出了基于 ResNet-50 主干网络使用不同数目的 MPM 时的结果。可以看到，当不使用 MPM(基础 FCN) 时，我们在 mIoU 方面的结果是 37.63%。当使用 1 个 MPM 时，达到 40.50%，即 3.0% 的提升。更进一步，当使用两个 MPM 时，约有 4.3% 的提升。然而，添加更多 MPM 带来的性能提升微乎其微。这可能的原因是感受野已经足够大了。因此，考虑到运行时间成本，我们将 MPM 的数量默认设置为 2。

为了说明提出的 MPM 相对于 PPM 的优势，我们还在表1中显示了 PSPNet 的结果和参数数量。可以明显看到，在 'Base FCN + 2 MPM' 的设定下，我们的方法已经优于 PSPNet，且比其少了 12M 的参数。该现象表明我们的模块化 MPM 设计比 PPM 更有效。

条带池化在 MPM **中的作用：** 第3.3小节已描述了 MPM 包含两个子模块用于分别捕捉短程依赖和长远依赖。在这里，我们探究了提出的条带池化的重要性。相关结果于表2展示。显然，条带池化捕捉长远依赖 (41.14%) 要比只捕捉短程依赖 (40.5%) 有效得多，但综合二者只能提升至 (41.92%)。为进一步验证条带池化在 MPM 中如何工作，我们可视化了一些 MPM 在不同位置处的特征图，如图 Figure 5所示，以及一些在不同设定的 MPM 下的分割结果，如图4所示。显然，所提出的条带池化可以更有效地捕捉长远依赖关系。例如，图5中最上面一行的长远依赖聚合模块 (LRD) 输出的特征图可以准确定位天空的位置。然而，全局平均池化就不能做到这一点，因为它将整个特征图编码为单个值。

SPM **的有效性：** 我们从经验上发现，尽管 SPM 轻量，但也没有必要将它添加到主干网络的每个结构块中。在本实验中，我们考虑了四个场景，它们列在表3中。我们以 base FCN 和 2 个 MPM 为基线。我们首先在每个阶段的最后一个结构块中添加一个 SPM；结果 mIoU 得分是 42.61%。其次，我们尝试在最后阶段将 SPM 添加到所有结构块中，发现性能略有下降到 42.30%。接下来，当我们将 SPM 添加到上述两个位置时，mIoU 得分可以达到 44.03%。但是，当我们试图将 SPM 添加到主干网络的所有结构块中时，几乎没有任何性能增益。对于上述结果，默认情况下，我们将 SPM 添加到每个阶段的最后一个结构块以及最后一个阶段的所有结构块中。此外，当我们仅将 base FCN 作为我们的基线并添加所提出的 SPM 时，mIoU 分数从 37.63% 增加到 41.66%，提高了近 4%。以上结果表明，在主干网络中加入 SPM 对场景解析网络有一定的帮助。

条带池化 v.s. **全局平均池化：** 为了证明所提出的条带池化相对于全局平均池化的优势，我们尝试将所



(a) Image (b) GT (c) 2 SRD (d) 2 LRD (e) 2 MPM

图 4. MP 模块 (MPM) 的不同设置之间的视觉比较。“2 SRD” 意味着我们使用 2 个 MPM，只包含短程依赖聚合模块，而 “2 LRD” 意味着我们使用 2 个 MPM，只包含 C 长远依赖聚合模块。

Settings	SPM Pos.	#MPM	mIoU	Pixel Acc.
Base FCN	-	2	41.92	80.03%
Base FCN + SPM	L	2	42.61	80.38%
Base FCN + SPM	A	2	42.30	80.22%
Base FCN + SE [22]	A + L	2	41.34	80.05%
Base FCN + SPM	A + L	0	41.66	79.69%
Base FCN + SPM	A + L	2	44.03	80.65%

表 3. 条带池化模块 (SPM) 的消融分析。L：每个阶段的最后一个结构块。A：最后阶段的所有结构块。可以看出，SPM 可以大大提高 base FCN 的性能，从 37.63 提高到 41.66。

Settings	MS + Flip	mIoU (%)	Pixel Acc. (%)
SPNet-50		44.03	80.65
SPNet-50	✓	45.03	81.32
SPNet-101		44.52	81.37
SPNet-101	✓	45.60	82.09

表 4. 使用不同主干网络时的消融实验。

提出的 SPM 中的条带池化操作更改为全局平均池化。以 base FCN 和 2 个 MPM 为基准，当我们在 base FCN 中添加 SPM 时，性能从 41.92% 增加到 44.03%。但是，当我们将条带池化更改为全局平均池化 [22] 后，性能从 41.92% 降至 41.34%，甚至比表3中所呈现的基线还低。这可能是由于直接融合特征图来构造一维向量，导致丢失过多的空间信息，从而产生以往工作 [65] 中指出的模糊性。

Method	Backbone	mIoU	Pixel Acc.	Score
RefineNet [32]	ResNet-152	40.70	-	-
PSPNet [65]	ResNet-101	43.29	81.39	62.34
PSPNet [65]	ResNet-269	44.94	81.69	63.32
SAC [63]	ResNet-101	44.30	81.86	63.08
EncNet [60]	ResNet-101	44.65	81.69	63.17
DSSPN [30]	ResNet-101	43.68	81.13	62.41
UperNet [52]	ResNet-101	42.66	81.01	61.84
PSANet [66]	ResNet-101	43.77	81.51	62.64
CCNet [23]	ResNet-101	45.22	-	-
APNB [69]	ResNet-101	45.24	-	-
APCNet [19]	ResNet-101	45.38	-	-
SPNet (Ours)	ResNet-50	45.03	81.32	63.18
SPNet (Ours)	ResNet-101	45.60	82.09	63.85

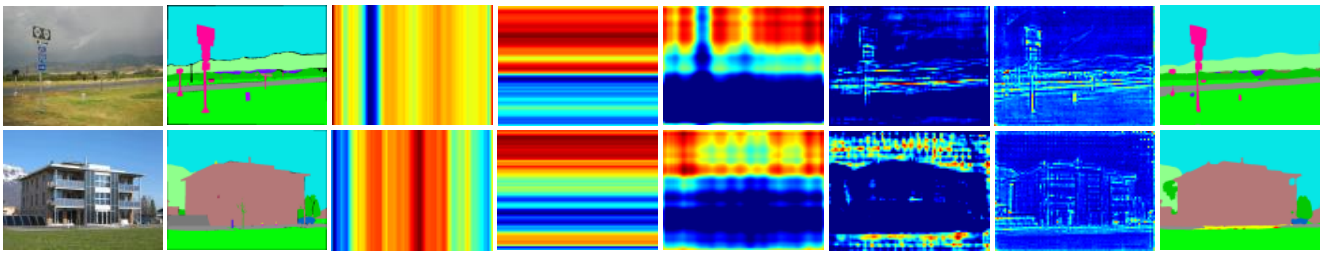
表 5. ADE20K [68] 验证集上与目前最先进的方法进行比较。我们在该基准测试上报告 mIoU 和 Pixel Acc。最好的结果以粗体高亮。

更多实验分析： 在这一部分中，我们展示了不同的实验设置对性能的影响，包括主干网络的深度和使用翻转的多尺度测试。如表4所示，使用翻转的多尺度测试可以在很大程度上改善两个主干网络的结果。此外，使用更深的主干网络也有利于性能提升 (ResNet-50：45.03% → ResNet-101：45.60%)。

可视化： 在图6中我们展示了几个在所提出方法的不同设定下的视觉示例。显然，不论添加至 base FCN 的是 MPM 还是 SPM，均可以显著提升分割结果。当 MPM 与 SPM 同时添加时，分割图的质量可以被进一步增强。

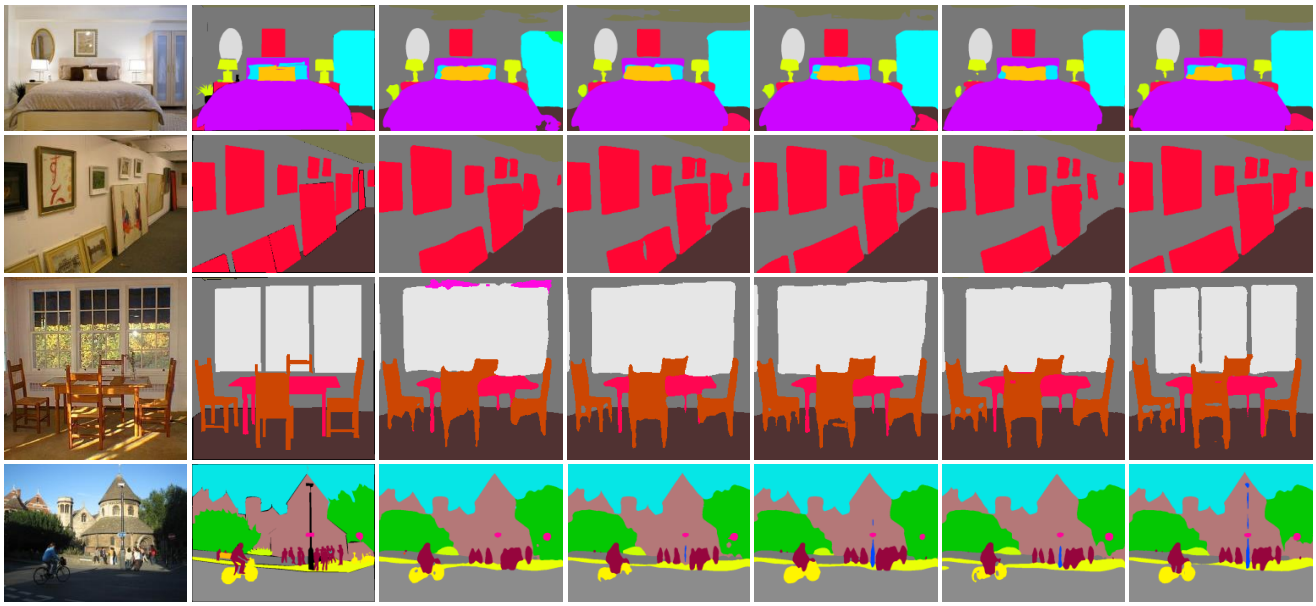
4.2.2 与最先进的方法比较

这里,我们将我们的方法与目前最先进方法进行比较。实验结果于表5展示。可以看到，以 ResNet-50 为主干网络的我们的方法 mIoU 评分达到 45.03%，像素精度达到 81.32%，已经优于以往的大多数方法。当以 ResNet-101 为主干网络时，我们得到了最先进的 mIoU 于像素精度结果。



(a) Image (b) GT (c) After VSP (d) After HSP (e) After LRD (f) After SRD (g) After MPM (h) Results

图 5. 在提出的 MP 模块的不同位置可视化选定的特征图。VSP：垂直条带池化；HSP：水平条带池化；SRD：短程依赖聚合子模块 (图3a)；LRD：长远依赖聚合子模块 (图3b)；MPM：混合池化模块。



(a) Image (b) GT (c) Base FCN (d) 1 MPM only (e) 2 MPM only (f) SPM only (g) SPNet

图 6. 该方法在不同模型设置下的可视化结果。

4.3. Cityscapes

Cityscapes [11] 是另一个流行的场景解析数据集，包含 19 个类别。它包含了从 50 个城市在不同季节下收集的 5K 高质量像素级标注图像，所有这些图像都是 1024×2048 的尺寸。根据以往工作的建议，我们将整个数据集划分为训练集、验证集和测试集三个部分，分别包含 2975、500 和 1525 张图像。

为了公平比较，我们采用 ResNet-101 为主干网络。我们在测试集上比较我们的方法与现有的方法。遵循 [16]，我们只使用精细标注的数据训练网络，并将结果提交至官方服务器。结果于表6所示。很显然，我们的方法由于所有其它方法。

4.4. Pascal Context

Pascal Context dataset [40] 有 59 个类别，10103 张图片带有密集的标签标注，其中 4998 张图片用于训练，5015 张图片用于测试。定量结果于表7展示。可以看到，我们的方法优于其它的方法。

5. 总结

本文提出了一个新型空间池化操作，条带池化。它长而窄的池化窗口允许模型收集丰富的全局上下文信息，这对场景解析网络至关重要。基于条带池化和空间池化操作，我们设计了一种新的条带池化

Method	Publication	Backbone	Test mIoU
SAC [63]	ICCV'17	ResNet-101	78.1%
DUC-HDC [50]	WACV'18	ResNet-101	80.1%
DSSPN [30]	CVPR'18	ResNet-101	77.8%
DepthSeg [24]	CVPR'18	ResNet-101	78.2%
DFN [56]	CVPR'18	ResNet-101	79.3%
DenseASPP [54]	CVPR'18	DenseNet-161	80.6%
BiSeNet [55]	ECCV'18	ResNet-101	78.9%
PSANet [66]	ECCV'18	ResNet-101	80.1%
DANet [16]	CVPR'19	ResNet-101	81.5%
SPGNet [9]	ICCV'19	ResNet-101	81.1%
APNB [69]	ICCV'19	ResNet-101	81.3%
CCNet [23]	ICCV'19	ResNet-101	81.4%
SPNet (Ours)	-	ResNet-101	82.0%

表 6. Cityscapes 测试集 [11] 上与最先进方法的比较。

Method	Publication	Backbone	mIoU (%)
CRF-RNN [67]	ICCV'15	VGGNet	39.3
BoxSup [12]	ICCV'15	VGGNet	40.5
Piecewise [33]	CVPR'16	VGGNet	43.3
DeepLab-v2 [5]	PAMI'17	ResNet-101	45.7
RefineNet [32]	CVPR'17	ResNet-152	47.3
CCL [60]	CVPR'18	ResNet-101	51.6
EncNet [60]	CVPR'18	ResNet-101	52.6
DANet [16]	CVPR'19	ResNet-101	52.6
SVCNet [14]	CVPR'19	ResNet-101	53.2
EMANet [29]	ICCV'19	ResNet-101	53.1
APNB [69]	ICCV'19	ResNet-101	52.8
BFP [13]	ICCV'19	ResNet-101	53.6
SPNet (Ours)	-	ResNet-101	54.5

表 7. Pascal Context 数据集 [40] 上与最先进方法的比较。

模块来增大骨干网络的感受野，并提出了一种基于经典的带有瓶颈结构的残差块的混合池化模块。在若干个广泛使用的数据集上的大量实验证明了所提出方法的有效性。

致谢. 本研究受到 AI.SG R-263-000-D97-490，NUS ECRA R-263-000-C87-133，MOE Tier-II R-263-000-D17-112，NSFC (61922046)，国家 "万人计划" 青年拔尖人才支持计划，以及天津市自然科学基金 (17JCJQJC43700) 的支持。

参考文献

[1] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In ECCV, 2016.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI, 2017.

[3] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. Computational Visual Media, 5(2):117–150, 2019.

[4] Samuel Rota Bulo, Gerhard Neuhold, and Peter Kontschieder. Loss max-pooling for semantic image segmentation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7082–7091. IEEE, 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI, 2017.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.

[7] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In CVPR, 2016.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018.

[9] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas Huang, Wen-Mei Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In ICCV, 2019.

[10] Ming-Ming Cheng, Qi-Bin Hou, Song-Hai Zhang, and Paul L. Rosin. Intelligent visual media processing: When graphics meets vision. Journal of Computer Science and Technology, 32(1):110–121, 2017.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The

cityscapes dataset for semantic urban scene understanding. In CVPR, 2016.

[12] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In ICCV, 2015.

[13] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In ICCV, pages 6819–6829, 2019.

[14] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In CVPR, pages 8885–8894, 2019.

[15] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In CVPR, 2018.

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In CVPR, pages 3146–3154, 2019.

[17] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence, 2019.

[18] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In CVPR, 2015.

[19] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In CVPR, pages 7519–7528, 2019.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[21] Seunghoon Hong, Junhyuk Oh, Honglak Lee, and Bohyung Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In CVPR, 2016.

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, pages 7132–7141, 2018.

[23] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. arXiv preprint arXiv:1811.11721, 2018.

[24] Shu Kong and Charless C Fowlkes. Recurrent scene parsing with perspective understanding in the loop. In CVPR, 2018.

[25] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In NeurIPS, 2011.

[26] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In CVPR, 2006.

[27] Thuc Trinh Le, Andrés Almansa, Yann Gousseau, and Simon Masnou. Object removal from complex videos using a few annotations. Comput. Visual Media, 5(3):267–291, 2019.

[28] Xiangtai Li, Li Zhang, Ansheng You, Maoke Yang, Kuiyuan Yang, and Yunhai Tong. Global aggregation then local distribution in fully convolutional networks. In BMVC, 2019.

[29] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In ICCV, pages 9167–9176, 2019.

[30] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In CVPR, 2018.

[31] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In ECCV, 2018.

[32] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In CVPR, 2017.

[33] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In CVPR, 2016.

[34] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In CVPR, 2017.

[35] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In CVPR, 2019.

[36] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In NeurIPS, 2017.

[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[38] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. High-resolution aerial image labeling with convolutional neural networks. IEEE TGRS, 55(12):7092–7103, 2017.

[39] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In ECCV, pages 552–568, 2018.

[40] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In CVPR, 2014.

[41] Saritha Murali, VK Govindan, and Saidalavi Kalady. Single image shadow removal by optimization using non-shadow anchor values. Comput. Visual Media, 5(3):311–324, 2019.

[42] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In ICCV, 2015.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In NeurIPS, pages 8024–8035, 2019.

[44] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In CVPR, pages 4353–4361, 2017.

[45] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In CVPR, 2017.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241, 2015.

[47] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In IEEE Intelligent Vehicles Symposium, pages 1013–1020, 2018.

[48] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In CVPR, pages 3126–3135, 2019.

[49] Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, and Rama Chellapa. Gaussian conditional random field network for semantic segmentation. In CVPR, 2016.

[50] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In WACV, 2018.

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In CVPR, 2018.

[52] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In ECCV, pages 418–434, 2018.

[53] Jimei Yang, Brian Price, Scott Cohen, and Ming-Hsuan Yang. Context driven scene parsing with attention to rare classes. In CVPR, 2014.

[54] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In CVPR, 2018.

[55] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In ECCV, 2018.

[56] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In CVPR, 2018.

[57] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016.

[58] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. arXiv:1809.00916, 2018.

[59] Hang Zhang. Pytorch-encoding. https://github.com/zhanghang1989/PyTorch-Encoding, 2018.

[60] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit

Agrawal. Context encoding for semantic segmentation. In CVPR, 2018.

[61] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. In BMVC, 2019.

[62] Li Zhang, Dan Xu, Anurag Arnab, and Philip HS Torr. Dynamic graph message passing networks. In CVPR, 2020.

[63] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In ICCV, 2017.

[64] Hengshuang Zhao. semseg. https://github.com/hszhao/semseg, 2019.

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, 2017.

[66] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In ECCV, 2018.

[67] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In ICCV, 2015.

[68] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In CVPR, 2017.

[69] Zhen Zhu, Mengdu Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In ICCV, 2019.