在本文中，我们用一种新的语境聚合方案来解决语义分割的问题，这种方案被称为对象语境，主要是为了加强对象信息的作用。由于每个像素的类别是由它所属的对象继承的，因此我们将每个像素的对象背景定义为图像中与给定像素属于同一类别的像素集合。我们用一个二元关系矩阵来表示所有像素之间的关系，其中数值1表示所选的两个像素属于同一类别，否则为0。我们建议使用密集关系矩阵作为二元关系矩阵的代用品。密集关系矩阵能够强调对象信息的贡献，因为对象像素上的关系分值往往比其他像素大。考虑到密集关系矩阵的估计需要二次计算开销和内存消耗，我们提出了一个高效的交错稀疏自我关注方案，通过两个稀疏关系矩阵的组合来模拟所有像素中任何两个像素之间的密集关系。为了捕捉更丰富的上下文信息，我们进一步将我们的交错稀疏自我关注方案与传统的多尺度上下文方案相结合，包括金字塔池（Zhao等人，2017）和无轨空间金字塔池（陈等人，2018）。我们通过在五个具有挑战性的基准上的竞争性表现实证了我们的方法的优势，包括。Cityscapes, ADE20K, LIP, PASCAL-Context和COCO-Stuff。

语义分割是计算机视觉的一个基本课题，对各种场景理解问题至关重要。它通常被表述为预测每个像素的类别的任务，即该像素所属的物体的类别。我们主要感兴趣的是通过明确地识别像素所属的物体区域来提高像素分类的准确性。自从全卷积网络（FCN）的开创性方法（Long等人，2015）以来，基于深度卷积神经网络的广泛努力已经被用来解决语义分割问题。最初的FCN方法有两个主要的缺点，包括降低了特征分辨率，失去了详细的空间信息，以及有效接收场小，不能捕捉长距离的依赖性。存在两条主要的路径来解决上述缺点：（i）提高特征图的分辨率，以提高空间精度，或在所有阶段保持高分辨率的反应图，例如，通过扩张卷积（Chen等，2018；Yu和Koltun，2016），解码器网络（Badrinarayanan等，2017；Ronneberger等，2015）或高分辨率网络（Sun等，2019a，b）。(ii) 利用全局背景来捕获长距离的依赖关系，例如ParseNet（Liu等人，2015）、DeepLabv3（Chen等人，2018）和PSPNet（Zhao等人，2017）。在这项工作中，我们专注于第二条路径，并提出一个更有效的上下文方案。我们将像素的上下文定义为一组选定的像素，如果没有指定，其上下文表示为所有选定像素表示的聚合。以前的大多数代表性研究主要是利用空间上邻近的或抽样的像素所形成的多尺度情境。例如，PSPNet（Zhao等人，2017）中的金字塔集合模块（PPM）将所有像素划分为多个区域，并选择与一个像素位于同一区域的所有像素作为其背景。DeepLabv3（Chen等人，2017年）中的无腹肌空间金字塔集合模块（ASPP）选择具有不同扩张率的像素的周围像素作为其上下文。因此，PPM上下文和ASPP（假如使用扩张卷积）上下文所选择的像素往往是物体像素、相关背景像素和不相关背景像素的混合物。由于每个像素的类别本质上是它所属的对象的类别，我们应该加强构成对象的像素。为了明确强调对象像素的贡献，我们提出了一个对象背景，旨在只收集与给定像素属于同一类别的像素作为其背景。与传统的多尺度语境方案相比，我们的对象语境更加关注必要的对象信息。尽管估计准确的对象背景并不是一件容易的事，但我们根据经验发现，对对象背景的粗略估计在各种基准上已经超过了PPM和ASPP方案。  
  
对于一个给定的像素，我们可以用一个二进制向量来记录与它属于同一类别的像素，用1，否则用0。因此，一个N×N的二元关系矩阵可以用来记录N个像素中任何两个像素之间的对等关系。由于计算二元关系矩阵是难以实现的，我们使用密集关系矩阵作为它的替代，其中每个关系值都是根据高级特征的内积相似度计算的。因此，语义上相似的像素的关系值往往更大。在我们的实现中，我们使用传统的自我注意方案（Vaswani等人，2017）来预测密集的关系矩阵，这需要O（N 2）的计算复杂性。为了解决效率问题，我们提出了一种新的交错稀疏自我关注方案，通过两个稀疏关系矩阵来近似密集关系矩阵，在保持性能的同时，显著提高了效率。为了说明我们的方法能够增强对象像素，我们在图1中展示了一些预测的密集关系矩阵的例子，其中对象像素上的关系值大于背景像素上的关系值。我们进一步说明了两个捕捉更丰富的上下文信息的扩展：（i）金字塔物体上下文，它估计了PPM之后空间金字塔分区产生的每个子区域内的物体上下文（Zhao等人，2017）。(ii) 腹式空间金字塔物体背景，它将ASPP（Chen等人，2017）与物体背景相结合。我们将我们的主要贡献总结为以下几点。- 我们提出了一个新的对象语境方案，明确地增强了对象信息。- 我们提出用高效的交错稀疏自留地来实例化对象上下文方案，与传统的自留地方案相比，它的复杂性大大降低。- 我们在三种对象上下文模块的基础上构建了OCNet，并在五个具有挑战性的语义分割基准上取得了具有竞争力的性能，包括Cityscapes, ADE20K, LIP, PASCAL-Context和COCO-Stuff。

我们用四个小节来介绍我们的方法。首先，我们介绍了情境表示的一般数学公式和物体情境的定义（第3.1节）。第二，我们用传统的自我注意（SA）（Vaswani等人，2017）和我们的交错稀疏自我注意（ISA）来实例化对象语境（第3.2节）。第三，我们介绍了对象背景的金字塔扩展（第3.3节）。最后，我们说明了OCNet的整体管道和实施细节（第3.4节）。

公式1：用一组相关像素的加权平均表示来描述一个像素

公式2：ASPP中Li的选定（和本文内容无关，展示用）

公式3：OC中Li的选定（只要同一类的都选）

其中li和lj分别是像素i和j的标签。我们可以（怎么判断是同一类）用一个N×N的二元关系矩阵来表示N个像素中任何两个像素之间的成对关系（在地面真相对象背景中编码），其中第i行用1记录所有与像素i属于同一类别的像素，否则为0。特别是，二元关系矩阵只编码了地面真实物体背景的部分信息，即（同一类别）标签共现关系。换句话说，所有的类别都可以被替换，而二元关系矩阵将是不变的。考虑到估计二元关系矩阵是难以实现的，我们建议1.使用密集关系矩阵来作为二元关系矩阵的替代。我们期望属于同一物体类别的像素之间的关系值比属于不同类别的像素大，因此，物体像素的贡献得到加强。在下面的讨论中，我们首先说明了密集关系方案的表述，即直接估计大小为N×N的密集关系矩阵W。其次，2.为了提高效率，我们提出了一个稀疏关系方案，将密集关系矩阵分解为两个稀疏关系矩阵的组合，包括Wl和Wg，其中两个稀疏关系矩阵的大小为N×N。更多细节说明如下。

密集关系。密集关系方案估计了每个像素i和图中所有像素之间的关系。我们说明了基于密集关系的上下文表述。

其中wij是像素i和j之间的关系值，即坐标（i，j）处的W元素。由于我们需要直接估计i和I中所有像素之间的关系，估计W的计算复杂度与输入大小成二次方。O(N 2)。

公式4——和公式1相比，权重关系更明确

稀疏关系。稀疏关系方案只按照 "交错（a.k.a. interleaving）方法"（Greenspun，1999；Roelofs and Koman，1999）估计像素i和两个选定像素子集之间的关系。我们举例说明基于稀疏关系的上下文表示。

我们使用上标g/l来分别标记与全局/局部关系阶段相关的运算符和操作。例如，zg i / zl i表示第i个像素的全局/局部关系阶段后的上下文表示。wg ij / wl ij是像素i和像素j之间的关系，分别属于Ig i / Il i。Ig i和Il i是被选中的背景像素。

公式5和6就是公式4的变体，把单一的W变为Wg和Wl（全局部分与局部部分）

接下来又回归到了“选哪些像素”的问题

其中Ig i / Il i是一个像素的子集，其余数/商数分别与像素i除以P时相同。

其中我们用P表示一个大小为N×N的置换矩阵，确保两个稀疏关系矩阵的像素排序相匹配，P>是P的转置。我们在附录B中说明了P中每个值pi,j的定义，以及为什么稀疏关系方案比密集关系方案在附录C中更有效。

上述插图中的i和j都代表一维情况下像素i和j的空间位置。P代表全局关系阶段的组号（第3.2节），它决定了上下文像素的选择。稀疏关系方案的主要优点在于我们只需要估计像素i和Ig i∪Il i之间的关系值，而不是I，因此，节省了大量的计算成本。考虑到具有相同余数/商数的像素共享相同的Ig ∗ / Il ∗，因此，我们忽略了下标，用Ig / Il分别代表所有共享相同余数/商数的像素组。我们分别在Ig / Il中的每组像素中计算（部分）稀疏关系矩阵Wg / Wl。特别是，Wg和Wl都是稀疏块矩阵，我们可以将原始的密集关系矩阵近似为这两个稀疏关系矩阵的乘积。

3.2 实例 我们分别基于自我注意和交错稀疏自我注意来解释密集关系和稀疏关系的具体实例。

自我注意。基于自我注意的密集关系方案的实现说明如下。

解释一下公式10和公式11（自注意力+密集关系矩阵）

X∈RN×Cin是输入表示，W∈RN×N是密集关系矩阵，Z∈RN×Cout是输出表示。为方便起见，我们在下面的讨论中假定Cin=Cout=C。θ和φ是两个不同的函数，将输入转化为低维空间，θ（X），φ（X）∈RN×C/2（通道数变为原来的一半）。根据Vaswani等人（2017）的研究，低维空间的内积用于计算密集关系矩阵W，缩放系数d用于解决softmax函数的小梯度问题，我们设定d=C 2。自我关注使用函数ρ和δ来学习更好的嵌入，我们有ρ（-）∈RN×C和δ（-）∈RN×C 2。根据Vaswani等人（2017）对self-attention的原始描述，我们也可以把θ、φ和δ分别称为query-、key-和valuetransform函数。我们用两组连续的1×1 conv→BN→ReLU来实现θ（-）和φ（-）。BN是批量归一化的缩写（Ioffe和Szegedy，2015），可以同步统计。具体来说，θ(-)、φ(-)和δ(-)将输入通道减半，而ρ(-)将输入通道加倍。

（交错稀疏自注意力+密集关系矩阵）

稀疏关系方案的实现，即交错稀疏自我关注，首先将所有像素分为多个子组，然后在每个子组上应用自我关注，计算稀疏关系矩阵，即Wg和Wl，以及上下文表示。我们用图2中的一个二维例子来说明交错稀疏自我关注方案的整体管道，我们用全局关系模块估计Wg，用局部关系模块估计Wl。通过这两个稀疏关系矩阵的组合，我们可以近似估计所有像素中任何两个像素之间的密集关系，附录D中用一个例子解释了这一点。全局关系。根据Ig的定义，我们把所有的位置分成多个组，每个组由一个采样位置的子集组成。考虑到每组内的像素是根据余数除以组数P来采样的，并且它们分布在整个图像范围内，因此，我们称之为全局关系。

关于图2

说明交错式稀疏自我关注的情况。我们的方法是由一个全局关系模块和一个局部关系模块组成。在最左边/最右边的特征图是输入/输出。首先，我们用四种不同的颜色给输入特征图X着色。我们可以看到，有4个局部组，每个组由4种不同的颜色组成。对于全局关系模块，我们将X中具有长空间间隔距离的所有相同颜色的位置进行互换和划分，从而输出Xg。然后，我们将Xg分成4组，并对每组独立应用自注意力。我们将每组的所有更新的特征图合并在一起作为输出的Zg。对于局部关系模块，我们对Zg进行处理，将原来附近的位置分组，得到Xl。然后，我们按照与全局关系相同的方式进行划分和应用自我关注，得到最终的特征图Zl。通过全局关系模块和局部关系模块的结合，我们可以将所有输入位置的信息传播到每个输出位置。具有相同颜色饱和度的位置标记的特征图的值保持不变。我们只在用自我关注操作更新特征图时增加颜色的饱和度。

图5展示与讲解

我们在图5（a）中说明了我们的OCNet的整体管道。更多细节说明如下。骨干网。我们使用ResNet-101（He等人，2016）或HRNetV2-W48（Sun等人，2019b）在ImageNet数据集上进行预训练作为骨干。对于ResNet-101，我们按照PSPNet（Zhao等人，2017）做了一些修改：用扩张率为2和4的扩张卷积取代最后两个块内的卷积，这样输出的步长就变成了8。对于HRNetV2-W48，我们直接将我们的方法应用于输出跨度为4的最终串联特征图。  
基础OC。在将特征图送入OCP之前，我们应用降维模块（3×3卷积），将ResNet-101和HRNetV2-W48从骨干网输出的特征图的通道减少到512。然后，我们将更新的特征图送入OCP，并将OCP的输出特征图与OCP的输入特征图连接起来。我们进一步进行1×1卷积，将串联后的特征图的通道数从1024个减少到512个，这不包括在图5（b）中。  
金字塔-OC。我们首先应用3×3卷积将通道减少到512，然后将降维后的特征图送入Pyramid-OC，对输入的特征图进行四种不同的金字塔分割（1×1区域、2×2区域、3×3区域和6×6区域），并将四个平行OCP输出的四种不同的输出物体背景特征图串联起来。四个对象背景特征图中的每一个都有512个通道。我们应用1×1卷积，将输入特征图的通道从512增加到2048，并将其与所有四个物体背景特征图串联。最后，我们用1×1的卷积法对4096个通道的串联特征图进行处理，并产生最终的512个通道的特征图，这不包括在图5（c）中。  
ASP-OC。我们只在对象上下文池分支内进行降维，在这里我们使用3×3卷积，将通道减少到256。来自对象上下文池模块的输出特征图有256个通道。对于其他四个分支，我们完全遵循原始的ASPP模块，在上述第二个分支内应用1×1卷积，在其余三个平行分支内应用不同扩张率（12、24、36）的3×3扩张卷积。我们在所有这四个分支中按照原来的设置将输出通道设置为256（Chen等人，2017）。最后，我们将这五个平行输出的特征图串联起来，并使用1×1卷积，将串联后的特征图的通道从1280减少到256，这不包括在图5（d）中。  
讨论。在我们的另一项工作中也讨论了物体上下文的概念：物体上下文表征（OCR）（Yuan等人，2019）。主要的区别在于，这项工作主要是对像素和像素之间的密集关系进行有效建模，而OCR主要是利用粗略的分割图来构建一套物体区域表征，并对像素和物体区域之间的密集关系进行建模。