基于多层级上下文信息的图像语义分割

岳师怡*
天津大学电气自动化与信息工程学院，天津 300072

摘要 随着深度学习和卷积神经网络的应用，图像语义分割的性能得到了大幅度提升。但当前图像语义分割算法在语义信息利用率及语义类别区分度方面仍有欠缺。为了进一步提升语义分割算法的性能，提出多层级的上下文信息机制，使用多层级特征对长距离的依赖关系信息和局部性较强的上下文信息进行提取，增强卷积神经网络特征的信息丰富度与类别区分度。所提多层级上下文信息机制在典型街道场景数据集 Cityscapes 验证集上的分割精度达 77.2%，实验证明了所提方法的有效性。

关键词 图像处理；语义分割；卷积神经网络；上下文信息；多层级特征
中图分类号 TP391.4 **文献标识码** A **doi**：10.3788/LOP56.241005

Image Semantic Segmentation Based on Hierarchical Context Information

Yue Shiyi*
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract The wide use of deep learning and convolutional neural networks in recent years has been one of the main reasons for performance improvement in image semantic segmentation. However, the current image semantic segmentation algorithms have certain drawbacks. For example, the semantic information is not fully used, and the discrimination between different semantic categories is not large enough. Therefore, we propose a hierarchical context information mechanism to achieve better semantic segmentation performance. The long-range dependency information and local context information (extracted from the hierarchical features) are conducive to enriching information and discriminating among different types of semantic categories. Our experiments demonstrate the effectiveness of the proposed method. The proposed method achieves a segmentation accuracy of 77.2% on Cityscapes val dataset.

Key words image processing; semantic segmentation; convolutional neural network; context information; hierarchical feature
OCIS codes 100.4996; 100.2960; 100.5010

1 引 言

图像语义分割是计算机视觉领域基础而重要的一项任务，语义分割被广泛应用于自动驾驶[1]、虚拟现实等计算机视觉任务中。图像语义分割的任务是为图像中的每个像素点分配一个类别标签，如人、车和建筑等，通过这些具有语义信息的标签完成对一幅图像的理解和解析。

卷积神经网络(CNN)[2]在计算机视觉任务中应用十分广泛，自 Krizhevsky 等[3]凭借其提出的 AlexNet 获得 ImageNet[4]图像分类竞赛冠军后，深度卷积神经网络逐渐在各类视觉任务中占据了主流地位。近年来，基于全卷积神经网络(FCN)[5]的图像语义分割模型取得了较大的进展。FCN 取消了传统卷积神经网络中的全连接层，可适应任意尺寸的输入[6-7]，并输出精细的分类结果。但由于卷积层操作的固有几何特性，基于 FCN 的语义分割模型感受野较小，只能利用局部上下文信息，类别区分性较差。

收稿日期：2019-04-25；**修回日期**：2019-05-29；**录用日期**：2019-06-13
基金项目：国家自然科学基金重点项目(61632018)
* **E-mail**：shiyiyue@tju.edu.cn

为了克服 FCN 上下文信息使用不够充分的缺点，Zhao 等[8]提出了金字塔池化模块，用以获取上下文信息；Chen 等[9-12]进一步提出了空洞空间金字塔池化（ASPP）模块，通过使用多个尺度的膨胀卷积层来提取较大范围的上下文信息，取得了较好效果。ASPP 模块使用多个具有不同膨胀率的卷积层对卷积神经网络特征进行信息提取，膨胀率越大，上下文信息范围越大，越有利于获取不同尺度范围的上下文信息，从而提高图像语义分割性能。Yang 等[13]在 ASPP 模块的基础上引入密集连接，用以获取尺度更加密集的特征。Peng 等[14]使用较大的卷积核提取上下文信息，获取较大范围的上下文信息。

近两年基于注意力机制的模型也被引入到语义分割任务中，用于提取上下文信息并提高图像语义分割性能。注意力机制模块对于长距离的依赖关系信息的提取效果较好。Wang 等[15]采用自激励注意力机制，使任意位置点的特征可接收来自其他所有位置点的特征信息，从而得到上下文信息更丰富的特征表示。Fu 等[16]使用两个注意力机制模块对卷积神经网络的特征在空间维度和通道维度上的依赖关系信息进行提取。Huang 等[17]简化了注意力模块，使用较少的计算量就可达到相近的分割性能，采用位置注意力机制模块（PAM）使各位置点的特征仅接收一定数量的其他位置点的特征信息，无需计算特征图中所有位置点的特征信息。

由于膨胀卷积操作的几何特性，膨胀卷积中卷积核的间隔较普通卷积核更大，因此膨胀卷积对特征中上下文信息的提取较普通卷积核更加稀疏，无法提取到当前特征点周围所有点的密集特征信息，并且膨胀卷积操作的感受野不能覆盖全图，无法获取远距离的上下文信息。而 PAM 模块将卷积神经网络特征图中所有点的特征信息整合到当前点中，容易引入相似类别的干扰信息。为解决上述问题，本文提出多层级上下文信息机制，将长距离的依赖关系信息提取与局部性较强的上下文信息提取相结合，增强了卷积神经网络特征的信息丰富度与类别区分度，从而达到更好的语义分割性能。

2 多层级上下文信息语义分割网络结构

2.1 多层级上下文信息网络

充分有效利用上下文信息对于图像语义分割任务十分重要。基于注意力机制的模型把卷积神经网络特征图中所有点的特征信息整合到当前点的特征之中，对于复杂场景图像，容易引入相似类别的信息造成干扰。而基于模块的 ASPP 模块的模型所提取的上下文信息不够密集，容易丢失部分上下文信息，并且由于图像场景中物体之间的尺度、光照和视角等差别较大，即使由相同语义标签像素点提取的特征也会存在差异，这些差异会影响识别精度。本文提出构建更加密集、局部性更强的膨胀卷积模块，并将长距离的依赖关系信息与局部性较强的上下文信息相结合，以增强卷积神经网络特征的信息丰富度与类别区分度，获得更好的语义分割性能。

在图像物体识别领域，使用卷积神经网络的不同层级进行目标物体的特征提取和分类有利于提升网络的性能和稳健性[18-19]，利用多层级特征进行上下文信息的提取，可以进一步丰富用于语义分割的卷积神经网络特征[20]。

He 等[21]提出的深度残差网络（ResNet）在图像分类任务中性能表现优异，被广泛用于物体检测、语义分割等任务的主干网络架构的图像特征提取，本文所提方法同样使用 ResNet 作为主干网络。本文提出的多层级上下文信息网络结构如图 1 所示。整个网络由两部分组成，使用 ResNet 作为主干网络进行图像特征提取，使用注意力机制模块和膨胀卷积模块进行上下文信息提取，并对提取的特征进行增强，将增强后的特征用于语义分割可提升语义分割的性能。图 1 中虚线表示输出预测结果。

主干网络中对 ResNet 网络的 block3 和 block4 提取到的特征使用类似特征金字塔[22]的结构进行增强，网络头部对增强后的特征进行上下文信息的提取和融合，之后输出预测结果。在卷积神经网络中，当前卷积层距离输入图像之间的卷积层个数可以表示网络的深浅。一般认为距离输入图像较近的浅层特征包含更多的细节信息（如轮廓、角点等），距离输入图像较远的深层特征经过较多的卷积层学习与抽象，包含更多的语义信息（如类别信息），即浅层特征语义信息的级别较低，深层特征语义信息的级别较高。由于语义分割任务更加依赖深层语义信息，现有的语义分割算法[10-14,16-17]大多仅使用主干网络中最后一个层级（包含的语义信息级别最高）的特征。考虑到细节信息有利于物体轮廓等细节的分割，以及语义分割任务对特征中的语义信息级别不能太低的要求，使用 Block3 和 block4 层级特征。Block3 和 block4 层级位于主干网络的最后，具备语义分割任务所需的高级别语义信息，同时在细节信息上有所差别（语义信息含量相近的情况下，block3

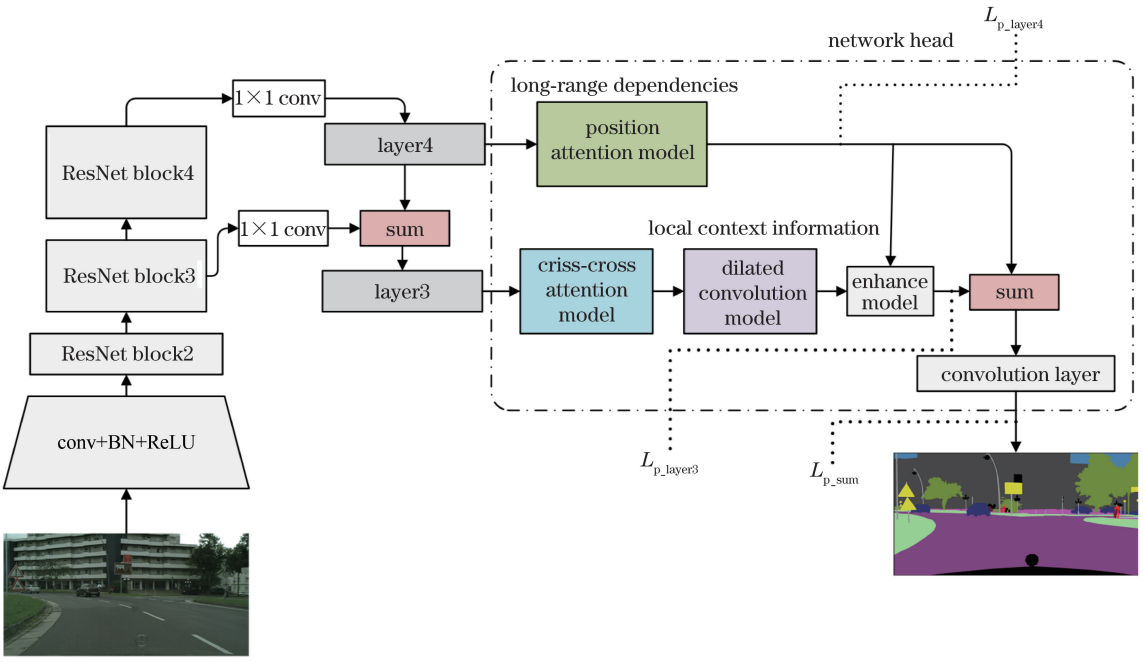

图 1 多层级上下文信息网络结构图

Fig. 1 Hierarchical context information network architecture

层级的特征包含更多的细节信息)。因此,结合 block3 和 block4 层级的特征有利于提高语义分割任务的语义分类准确度和物体轮廓分割的精确度。由于 block3 和 block4 的特征通道数量不一致,首先使用1×1 的卷积进行通道降维。Block4 的特征降维后得到 layer4,由于 block4 的特征包含的语义信息更多,之后使用来自 block4 的特征 layer4 对 block3 降维后的特征进行语义信息增强,得到 layer3。

2.2 多层级上下文信息网络头部

多层级上下文信息网络的头部分别在 layer4 和 layer3 进行上下文信息的提取融合。Layer4 使用 PAM 模块对长距离的依赖关系信息进行提取。Layer3 使用提出的更加密集、局部性更强的膨胀卷积模块对局部细节信息进行提取。

PAM 结构如图 2 所示。为了提取长距离的依赖关系信息,PAM 使用特征图中各点所含的特征信息的加权和表示当前点进行信息提取操作之后的特征(图 2 中小立方体表示当前点),权重值用当前点特征与特征图中各点特征的相似度表示。图 2 中小方块的颜色深浅表示特征图中各点对当前点的权重值。由于 PAM 整合了特征图中各点的信息,并且该信息不是通过卷积层学习得到的,而是通过计算特征相似度得到的,故 Fu 等[16]认为该模块提取的是长距离的依赖关系信息。即 PAM 模块在计算时,先对当前特征图中各点的特征向量使用向量转置相乘计算相似度,相似度较大表示参与计算的两个特征向量相似,属于同一类物体的可能性较大,反之参与计算的两个特征向量之间的差异较大。将相似度归

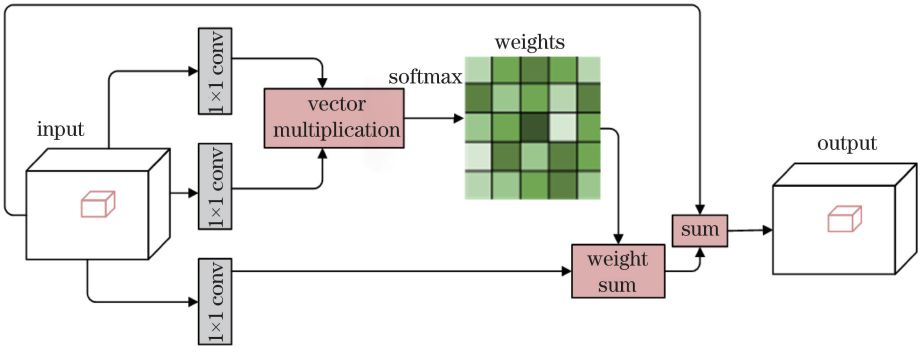

图 2 位置注意力机制模块图

Fig. 2 Position attention mechanism model

一化得到权重，使用权重计算特征图中各点特征向量的加权和，并更新各点的特征向量，则 PAM 模块将特征图中各点特征信息整合到当前点中。

膨胀卷积模块如图 3 所示。膨胀卷积模块采用多个不同膨胀率的卷积层进行卷积操作，并把卷积操作得到的特征级联到一起，之后使用 1×1 的卷积进行通道降维。膨胀卷积模块类似于 ASPP 模块，不同的是为了提取局部性较强的细节化上下文信息，膨胀卷积模块选择的膨胀率更小，不同卷积层的膨胀率选取更密集，并取消了全局池化操作。在后续实验中，膨胀卷积模块使用的膨胀率为 3、6、9，相比于 ASPP 模块中常用的膨胀率 24，膨胀卷积模块提取到的特征局部性更强，密集程度更高，有利于细节化特征的提取。

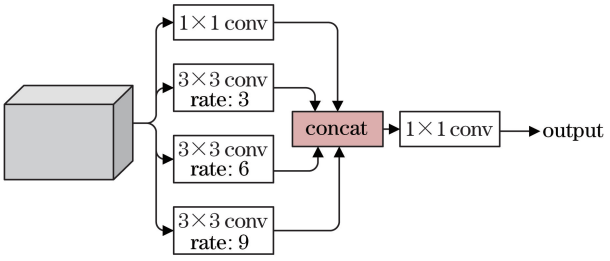

图 3 膨胀卷积模块图

Fig. 3 Dilated convolution model

PAM 模块虽然可以提取长距离的依赖关系信息，但容易造成相似类别的信息干扰。这主要是因为 PAM 模块依靠特征相似度进行上下文信息的提取，不是一个学习的计算过程，容易使得外观上相似的类别（如树冠、草地）相互干扰。此外，PAM 模块注重长距离信息的提取，对物体局部细节化的特征提取不足，对轮廓复杂的目标物体如自行车、三角形路标等的分割结果粗糙。因此，本文提出膨胀卷积模块旨在弥补 PAM 模块在信息提取上的不足。膨胀卷积模块使用较小的膨胀率提取局部性更强、密集程度更高的细节特征，可以对 PAM 模块提取的长距离依赖关系信息进行补充。

为了更好地融合多层级上下文信息，本文提出了增强模块，基于增强模块对较浅层级（如 layer3）提取到的包含上下文信息的特征作进一步处理，结构如图 4 所示。首先对 layer4 的 PAM 输出特征作全局池化操作，之后对池化操作后的特征与膨胀卷积模块的输出特征作点积，点积的结果加上膨胀卷积模块的输出特征即可得增强后的特征。融合操作可以使用相加操作，也可以使用级联操作，前期的实验中，这两种融合方式的性能几乎无差异，为使模块

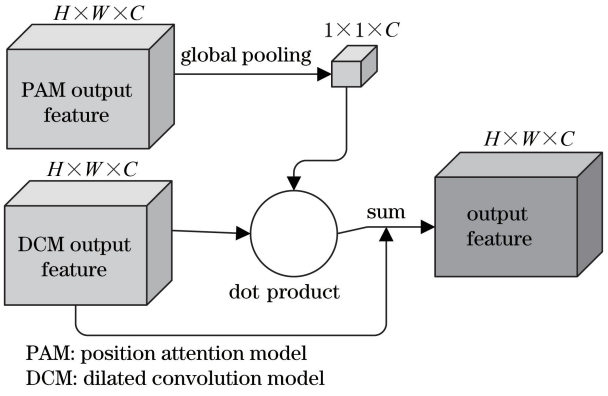

图 4 增强模块图

Fig. 4 Enhance model

简单化，使用简单相加进行特征融合。使用增强模块可充分利用较深层级（如 layer4）提取到的特征具有覆盖全局的感受野和更高级语义信息的优势，增强较浅层级（如 layer3）提取到的特征。

2.3 多层级上下文信息网络损失函数

损失函数是深度卷积神经网络训练过程中十分重要的一环，梯度下降算法通过减小损失函数逐步搜寻较优的参数，使卷积神经网络具有较好的性能。像素级的交叉熵损失函数是语义分割领域中最常用的损失函数[5,10]，损失函数惩罚各像素点的预测值与真实标签值的差。若对于一个像素点 i，网络输出的预测值为 $\hat{y}_i(l)$，表示该像素点属于第 l 类的概率，则该像素点在第 l 类上的损失函数可表示为

$$L_l^i = -\{y_l^i \log[\hat{y}_i(l)] - (1-y_l^i)\log[1-\hat{y}_i(l)]\}, \tag{1}$$

式中：y_l^i 为指示器，若像素点 i 的真实语义标签为第 l 类，y_l^i 值为 1，若像素点 i 的真实语义标签是其他类，y_l^i 值为 0。整幅图像在所有类上的损失函数为

$$L = \frac{1}{n}\sum_i \sum_l L_l^i, \tag{2}$$

式中：n 为像素总数。

多层级上下文信息网络使用了 layer3 和 layer4 的特征，为更好地监督整个网络的学习过程，采用 3 处预测结果（图 1 中虚线处表示输出预测结果）计算损失函数作为监督。整个网络的损失函数为

$$L_{\text{net}} = L_{\text{p\_sum}} + \lambda_1 L_{\text{p\_layer4}} + \lambda_2 L_{\text{p\_layer3}}, \tag{3}$$

式中：$L_{\text{p\_sum}}$ 为使用 2 个层级特征进行预测得到的损失函数；$L_{\text{p\_layer4}}$ 和 $L_{\text{p\_layer3}}$ 分别为单独使用 layer4 或 layer3 特征进行预测得到的损失函数；λ_1 和 λ_2 为不同损失函数的权重系数。$L_{\text{p\_sum}}$、$L_{\text{p\_layer4}}$ 和 $L_{\text{p\_layer3}}$ 可通过(2)式计算得到。超参数 λ_1 和 λ_2 的选择一般

可通过经验方式进行确定，人为设定几组不同的超参数值对比网络性能，从中选择一组较优的超参数。前期实验发现，不同超参数 λ_1 和 λ_2 组合对网络性能的影响较小，以下实验中，在性能差异不大的情况下参考 Huang 等[17]的超参数设计，选取 λ_1 值为 1，λ_2 值为 0.4。

3 实验结果与分析

3.1 数据集与实现细节

为验证所提方法的有效性，在 Cityscapes[23-24] 数据集上进行语义分割实验。Cityscapes 数据集是图像语义分割领域中使用十分广泛的数据集，包括从 50 个城市获取的 5000 张精细标注和 20000 张粗略标注的城市路面场景图像。实验仅采用了 5000 张精细标注的图像，未使用粗略标注的图像。5000 张精细标注的图像包括 2975 张训练图像数据、500 张验证图像数据和 1525 张测试图像数据。测试数据没有提供真实标签，需要提交其官方服务器进行测试。Cityscapes 数据集的图像分辨率均为 2048 pixel×1024 pixel，每个像素点对应一个类别标签，共分为 19 类（包括车、建筑和行人等）。为保证对比的公平性，仅使用训练图像进行训练，并在验证图像或测试图像上进行测试。

依照 Chen 等[10]、Yang 等[13]和 Zhang 等[25]的工作经验，本实验所使用的评价指标是语义分割领域常用的平均交并比（M_{iou}）。M_{iou} 反映预测值和真实值之间的相关度，相关度越高，M_{iou} 越大。

$$M_{\mathrm{iou}}=\frac{f_{\mathrm{TP}}}{f_{\mathrm{TP}}+f_{\mathrm{FP}}+f_{\mathrm{FN}}},\tag{4}$$

式中：f_{TP}、f_{FP} 和 f_{FN} 分别为真正率（标签为正，预测结果为正）、假正率（标签为负，预测结果为正）和假负率（标签为正，预测结果为负）。

实验采用的硬件环境为 GeForceGTX TITAN X GPU。方法的实现使用 Pytorch 框架。使用批量随机梯度下降算法[2]进行训练，依照 Chen 等[11]和 Zhang 等[25]的工作经验，且为了更加公平地进行不同方法性能的比较，采用多元学习率策略，其中初始学习率每次迭代后都乘以因子 $\left(1-\frac{N_{\mathrm{iter}}}{N_{\mathrm{total\_iter}}}\right)^{0.9}$，以此方式来减小学习率，其中 N_{iter} 为当前迭代次数，$N_{\mathrm{total\_iter}}$ 为总迭代次数。为保证实验结果对比的公平性，网络训练过程中超参数采用了 Fu 等[16]的设置。其中，初始学习率设为 0.03，使用的动量系数和权重衰减系数分别为 0.9 和 0.0001。训练过程中采用了随机水平翻转和随机缩放进行数据增广。

受 Cityscapes 图像数据过大以及计算资源的限制，在切片实验中，使用缩放后分辨率为384 pixel×384 pixel 的图像进行训练和测试。在最后的对比实验中，使用未缩放的图像进行训练和测试，并在相同参数设置下对所提方法和 Fu 等[16]提出的方法进行性能对比。

3.2 切片实验

切片实验结果如表 1 所示，其中第 1 行为 Fu 等[16]所提方法在 Cityscapes 验证集上的语义分割性能，记为 Baseline，具体结构如图 5 所示。

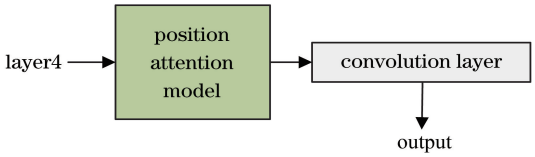

图 5 基础方法 baseline[16] 网络结构图

Fig. 5 Network structure of baseline[16]

表 1 Cityscpaes 验证集上各模块性能对比

Table 1 Performance of each model on Cityscapes validation set

Method	With/without model					M_{iou}/%
	PAM	DCM-4	DCM-3	EM	CCM	
Baseline	√					59.82
Ours-1	√	√				60.73
Ours-2	√		√			61.16
Ours-3	√		√	√		61.78
Ours-4	√		√	√	√	62.11

Note: DCM is dilated convolution model; EM is enhance model; CCM is cross-cross attention model; baseline is method in Ref. [16].

首先，探究长距离的依赖关系信息与局部性较强的上下文信息相结合的效果。表 1 中第 2 行表示在 baseline 的基础上增加本文提出的膨胀卷积模块之后的性能，记为 ours-1，具体结构如图 6 所示。对比第 1、2 两行结果可以看出，增加膨胀卷积模块之后，网络性能提高了将近 1 个百分点，证明长距离的

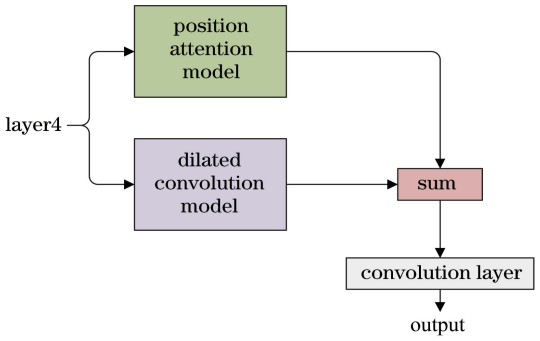

图 6 Ours-1 网络结构图

Fig. 6 Network structure of ours-1

依赖关系信息与局部性较强的上下文信息具有互补作用，有利于语义分割性能的提高。

其次，探究使用多层级特征的效果。表 1 中第 3 行表示在 baseline 的基础上，在 layer3 增加膨胀卷积模块之后的性能，记为 ours-2，具体结构如图 7 所示。对比第 2、3 两行结果可以看出，使用不同层级的特征(图 7)比使用单层级 layer4 的特征(图 6)，其网络性能提高了约 0.4 个百分点。

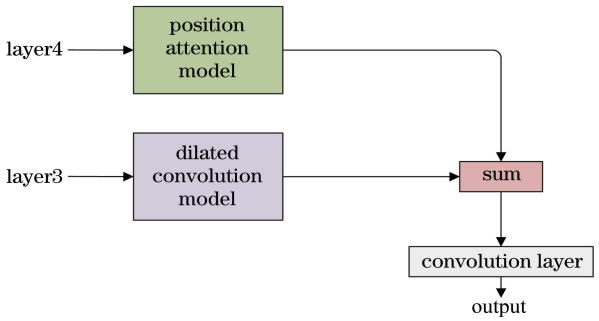

图 7 Ours-2 网络结构图

Fig. 7 Network structure of ours-2

再次，探究增强模块的效果。表 1 中第 4 行表示在 ours-2 结构的基础上，增加增强模块之后的性能，记为 ours-3，具体结构如图 8 所示。对比第 3、4 行的结果，可以看出，增强模块的加入使网络性能提高了约 0.6 个百分点。

最后，为进一步提高性能，同时尽可能地减少增加的计算量，在 ours-3 结构中插入了 Huang 等[17]提出的十字型位置注意力机制（CCM）模块，记为 ours-4，具体结构如图 1 所示。CCM 模块是 PAM 模块的简化，与 PAM 模块的计算流程相同，不同的是 PAM 使用特征图中所有的特征计算加权和，而 CCM 仅使用位于当前点同一行、同一列处的特征进行加权和的计算。表 1 中第 5 行表示 ours-4 在 Cityscapes 验证集上的语义分割的性能。可以看出，加入 CCM 模块后，网络性能提高了约 0.3 个百分点。

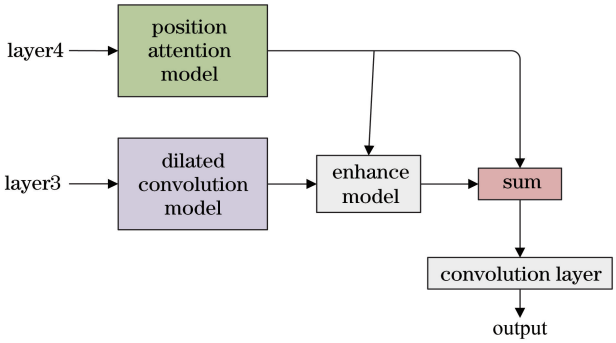

图 8 Ours-3 网络结构图

Fig. 8 Network structure of ours-3

需要指出的是，插入 CCM 模块之前，本文提出的长距离依赖关系信息与局部性较强的上下文信息相结合的方法已经使网络性能提高了近 2 个百分点，插入 CCM 模块只是为了进一步提高分割性能。

3.3 性能对比

对比所提方法与基础方法 baseline 的性能，网络中的超参数设置均与基础方法保持一致。首先采用 384 pixel×384 pixel 分辨率进行训练，使用测试集在官网进行测试，结果如表 2 所示。Cityscapes 测试集的实验结果显示，所提方法使网络的语义分割精确度从 60.17 提高到 62.19，提高了约 2 个百分点。需要说明的是，对于轮廓更加精细的物体类别，如 fence(Fen)、traffic sign(Tra)、vegetation(Veg)、person(Per)、rider(Rid)、car(Car)和 motorcycle(Mot)等，相比于 baseline 方法，所提方法的语义分割精确度显著提升，实验证明了所提方法的多层级机制以及提出的用于提取局部性更强上下文信息的 DCM 模块在语义分割任务中的有效性。

使用 768 pixel×768 pixel 分辨率进行训练和测试，受限于计算资源，选取 ResNet50 作为主干网络，实验结果如表 3 所示。其中 baseline* 表示 Fu 等[16]报道的网络性能，baseline* 使用了 4 块 GPU，采用最小批 8 张图像进行训练，最终在 Cityscapes 验证集上达到了 76.34%的精度。Fu 等[16]未报道其在测试集上的性能。受限于计算资源，baseline 和所提方法 ours-4 均使用最小批 2 张图像进行训练和测试，baseline 在验证集上的语义分割性能达 75.81%，而所提方法 ours-4 在验证集上的语义分割性能可达 77.20%，提高了约 1.4 个百分点。在测试集上，baseline 模型可获得 75.46%的精度，所提方法 ours-4 在测试集上精度可达到 76.73%，提高了约 1.2 个百分点。

表 2 Cityscapes 测试集上各模块性能对比

Table 2 Performance of each model on Cityscapes test set %

Method	Accuracy																		M_{iou}
	Roa	Sid	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter	Sky	Per	Rid	Car	Tru	Bus	Tra	Mot	Bic	
Base-line	96.5	72.4	85.8	38.1	35.7	30.6	41.7	50.6	86.3	58.9	90.4	65.2	45.8	88.0	55.1	64.5	49.8	36.3	51.9	60.17
Ours-2	96.7	73.5	86.1	38.2	36.8	33.3	42.7	51.3	87.1	61.1	91.6	66.1	47.8	90.3	54.8	67.2	50.3	36.7	53.0	61.30
Ours-3	96.7	73.0	86.4	44.2	37.5	30.7	41.5	51.5	87.4	58.4	91.7	66.9	47.3	90.5	56.6	68.0	56.9	40.3	52.4	61.99
Ours-4	96.9	74.5	86.7	44.2	38.2	31.2	42.8	52.1	87.8	59.7	91.7	66.7	48.3	90.4	55.1	65.4	55.8	40.5	53.5	62.19

表 3 Cityscapes 数据集上网络性能对比

Table 3 Network performance on Cityscapes dataset

Model	Backbone	M_{iou} / %	
		Val set	Test set
Baseline*	ResNet-50	76.34	—
Baseline	ResNet-50	75.81	—
Ours-4	ResNet-50	77.20	—
Deeplab-V2[10]	ResNet-101	—	70.40
Refinenet[20]	ResNet-101	—	73.60
Gcn[14]	ResNet-101	—	76.90
Baseline	ResNet-50	—	75.46
Ours-4	ResNet-50	—	76.73
Baseline	ResNet-101	—	76.67
Ours-4	ResNet-101	—	78.01

在 Cityscapes 测试集上对比 3 个在语义分割任务中使用广泛的经典方法，结果表明，所提方法在使用性能较弱的主干网络(ResNet50)的情况下，其性能仍能够超过或接近多个经典方法使用性能较强主干网络(ResNet101)的性能。换用 ResNet101 作为主干网络后，baseline 模型可获得 76.67%的精度，所提方法 ours-4 在测试集上精度可达 78.01%，提高了约 1.3 个百分点。实验证明，使用多层级的特征和长距离的依赖关系信息与局部性较强的上下文信息相结合的方法，对提高语义分割任务的精度是有效的。

3.4 定性结果分析

使用定性方式对比所提方法与基础方法 baseline 的性能，结果如图 9 所示。其中第 1 行为 Cityscapes 验证集的图像，第 2 行为 Cityscapes 验证集图像的语义标签，第 3 行为使用基础方法 baseline 对 Cityscapes 验证集的图像进行语义分割的结果，第 4 行是使用所提方法 ours-4 对 Cityscapes 验证集的图像进行语义分割的结果。

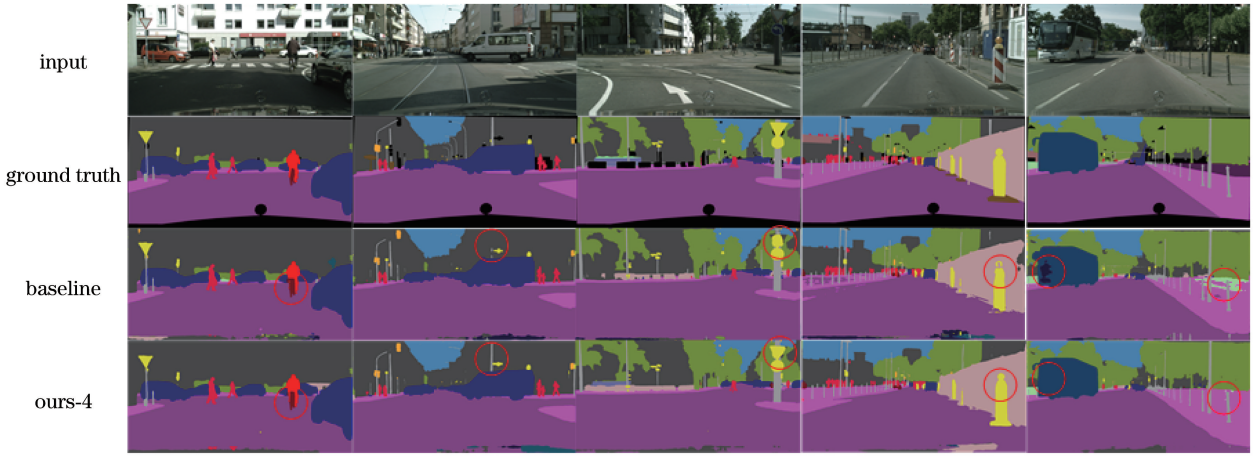

图 9 Baseline 和 ours-4 的定性对比

Fig. 9 Qualitative comparison between baseline and ours-4

对比图 9 中第 1 列的分割结果，在图像的自行车部分(红色圆圈处)，所提方法对自行车的轮廓分割相比于基础方法更加精细。同样地，第 2、3、4 列图像中，在电线杆、三角形路标和柱形路障处，所提方法的语义分割结果对于物体轮廓的处理更加贴合。这是因为基础方法仅使用长距离的依赖关系信息，缺少局部的上下文语义信息，加入局部上下文信息可以增强网络对于物体轮廓细节的分割性能。在图 9 的最后一列图像中，基础方法将大型客车的一部分(左侧圆圈处)错误地分类为小汽车，而所提方法没有出现对客车的误分类情况。这是因为客车的外观特征与小汽车类似，基础方法使用长距离的依赖关系信息，相似类别存在信息干扰，所提方法通过增加局部信息可以改善这一问题。在图 9 的最后一

列图像中右侧圆圈处，基础方法将距离树冠的较近处误分类成草地，同样存在相似类别信息干扰，加入局部上下文信息可以避免相似类别之间信息干扰的情况发生。

4 结 论

在多层级上下文信息语义分割网络中，使用多层级的特征、长距离的依赖关系信息与局部性较强的上下文信息相结合的方式，增强了卷积神经网络特征的信息丰富度与类别区分度，获得了更好的语义分割性能。设计了简单有效的膨胀卷积模块，以提取更加密集的局部性较强的上下文细节信息。为更充分地利用不同层级特征所含的信息，达到信息的互补，引入了增强模块，通过进一步丰富特征的语义信息提高类别区分度。在 Cityscapes 数据集上的实验结果表明，多层级上下文信息语义分割网络的分割精度达到 77.20%，比基础方法提高了约 1.4 个百分点。此外，所提方法模块可以插入到不同结构的主干网络，以实现语义分割性能的提升。未来将采用深度可分离卷积等方法降低模型计算量、提高模型速度，以促进语义分割模型在智能驾驶等领域的实际应用。

参 考 文 献

[1] Guo C C, Yu F Q, Chen Y. Image semantic segmentation based on convolutional neural network feature and improved superpixel matching [J]. Laser & Optoelectronics Progress, 2018, 55(8): 081005.
郭呈呈, 于凤芹, 陈莹. 基于卷积神经网络特征和改进超像素匹配的图像语义分割[J]. 激光与光电子学进展, 2018, 55(8): 081005.

[2] LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[3] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[4] Deng J, Dong W, Socher R, *et al*. ImageNet: a large-scale hierarchical image database [C] // 2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009: 248-255.

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 3431-3440.

[6] Fang X, Wang G H, Yang H C, *et al*. High resolution remote sensing image classification combining with mean-shift segmentation and fully convolution neural network [J]. Laser & Optoelectronics Progress, 2018, 55(2): 022802.
方旭, 王光辉, 杨化超, 等. 结合均值漂移分割与全卷积神经网络的高分辨遥感影像分类[J]. 激光与光电子学进展, 2018, 55(2): 022802.

[7] Wang L, Liu Q. A multi-object image segmentation algorithm based on local feature [J]. Laser & Optoelectronics Progress, 2018, 55(6): 061002.
王琳, 刘强. 基于局部特征的多目标图像分割算法[J]. 激光与光电子学进展, 2018, 55(6): 061002.

[8] Zhao H S, Shi J P, Qi X J, *et al*. Pyramid scene parsing network [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6230-6239.

[9] Chen L C, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs[J/OL]. (2016-06-07) [2019-03-30]. https://arxiv.org/abs/1412.7062.

[10] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[11] Chen L C, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation[J/OL]. (2017-12-05) [2019-03-30]. https//arxiv.org/abs/1706.05587.

[12] Chen L C, Zhu Y K, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation [M] // Ferrari V, Hebert M, Sminchisescu C, *et al*. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 833-851.

[13] Yang M K, Yu K, Zhang C, *et al*. DenseASPP for semantic segmentation in street scenes [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 3684-3692.

[14] Peng C, Zhang X Y, Yu G, *et al*. Large kernel

matters: improve semantic segmentation by global convolutional network[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1743-1751.

[15] Wang X L, Girshick R, Gupta A, *et al*. Non-local neural networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7794-7803.

[16] Fu J, Liu J, Tian H J, *et al*. Dual attention network for scene segmentation[C]// The IEEE Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 3146-3154.

[17] Huang Z L, Wang X G, Huang L C, *et al*. CCNet: criss-cross attention for semantic segmentation [J/OL]. (2018-11-28) [2019-03-30]. https// arxiv.org/abs/1811.11721.

[18] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot MultiBox detector[M]// Leibe B, Matas J, Sebe N, *et al*. Computer vision - ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[19] Cai Z W, Fan Q F, Feris R S, *et al*. A unified multi-scale deep convolutional neural network for fast object detection[M]// Leibe B, Matas J, Sebe N, *et al*. Computer vision - ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 354-370.

[20] Lin G S, Milan A, Shen C H, *et al*. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5168-5177.

[21] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[22] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 936-944.

[23] Cordts M, Omran M, Ramos S, *et al*. The cityscapes dataset [C]// CVPR Workshop on the Future of Datasets in Vision, June 7-12, 2015, Boston, Massachusetts. New York: IEEE, 2015.

[24] Cordts M, Omran M, Ramos S, *et al*. The cityscapes dataset for semantic urban scene understanding [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 3213-3223.

[25] Zhang H, Dana K, Shi J P, *et al*. Context encoding for semantic segmentation [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7151-7160.