文章编号:1671-4598(2024)12-0223-06 DOI:10.16526/j.cnki.11-4762/tp.2024.12.032 中图分类号:TP183 文献标识码:A

基于自适应注意力机制的轻量化语义分割网络

王艳莉，连晓峰，康毛毛
（北京工商大学 计算机与人工智能学院，北京 100048）

摘要：针对语义SLAM中语义分割速度较慢，实时性较低、占用资源过多等问题，提出一种含有自适应通道注意力机制的轻量级Mask R-CNN网络，由于原有的语义分割网络里的残差网络复杂，且应用环境在室内，环境较为简单，故该轻量级网络将原有复杂的主干网络中的ResNet-50利用深度可分离卷积与分组卷积改进为更加轻量的ResNet-DS-tiny，并加入自适应通道注意力机制提升网络精度；在自适应通道注意力模块中，利用加权方式对输入的RGB-D图像从空间和通道赋予不同的权重，增强了特征的表达能力；此外，为了轻量化特征金字塔，使用不同空洞率的空洞卷积来提取不同大小感受野的特征信息，有效地获取了多尺度的特征；相较于传统的特征金字塔，空洞卷积减少了参数量；在更充分获取RGB信息特征的同时，提升了语义分割系统的实时性并减少了资源占用。

关键词：室内语义分割；轻量化网络；注意力机制；空洞卷积

Lightweight Semantic Segmentation Network Based on Adaptive Attention Mechanism

WANG Yanli, LIAN Xiaofeng, KANG Maomao
(College of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China)

Abstract: To address the issues of slow semantic segmentation speed, low real-time performance, and high resource consumption in semantic SLAM, a lightweight Mask R-CNN network with an adaptive channel attention mechanism is proposed. Given the complexity of the residual networks in existing semantic segmentation networks and the relatively simple indoor application environments, this lightweight network replaces the original complex backbone ResNet-50 with a more lightweight ResNet-DS-tiny by incorporating depthwise separable convolutions and grouped convolutions. An adaptive channel attention mechanism is also introduced. In the adaptive channel attention module, a weighted approach is used to assign different weights to the input RGB-D images from both spatial and channel dimensions, thereby enhancing the feature representation capability. Additionally, to lighten the feature pyramid, dilated convolutions are employed to expand the receptive field, effectively aggregating multi-scale features with different dilation rates. Compared to traditional feature pyramids, the use of dilated convolutions reduces the number of parameters. This approach not only more effectively captures RGB information features but also improves the real-time performance of the semantic segmentation system while reducing resource consumption.

Keywords: indoor semantic segmentation; lightweight network; attention mechanism; dilated convolution

0 引言

为了实现室内环境下的语义建图，实时且可靠地识别周围环境中的语义信息是至关重要的。视觉感知是环境感知技术的重要组成部分，该技术不仅能够获取到周围环境的纹理、颜色、空间、形状等信息，而且安装成本低。图像语义分割是一种图像处理和视觉感知技术，旨在将图像中的每个像素分类为某一类特定的对象或区域，并对不同像素的类别进行标注，使每个像素都具有明确的类别标签，从而更好地理解环境信息。图像语义分割广泛地应用于自动驾驶、移动机器人和增强现实等领域，在无人驾驶、移动机器人和增强现实等领域有着广泛应用，该技术不仅推动了相关行业的发展，也显著改善了人们的日常生活和工作方式。

目前，深度学习的先进方法和技术已经广泛应用于语义分割领域，大幅度提高了算法的性能和精度。然而，在室内场景中，语义分割仍然面临许多挑战。室内场景下存

收稿日期:2024-06-07； **修回日期**:2024-07-19。
基金项目:重庆自然科学基金(CSTB2022NSCO-MSX1415)。
作者简介:王艳莉(2003-),女,大学本科生。
连晓峰(1977-),男,博士,副教授,硕士生导师。
引用格式:王艳莉,连晓峰,康毛毛.基于自适应注意力机制的轻量化语义分割网络[J].计算机测量与控制,2024,32(12):223-228,235.

在复杂的背景信息干扰、多变的光照条件导致图像不一致、物体相互遮挡使得边界模糊，以及低层次特征辨识度较弱，难以提供足够的信息进行准确分类，从而影响语义分割的整体效果。因此，在室内场景中实现高精度的语义分割仍然需要进一步的研究和优化。轻量化模型是为了解决在硬件资源有限的情况下，传统神经网络在计算资源消耗和能耗方面的问题。该方法能够在保持神经网络精度的同时，通过模型设计或模型压缩等方法减少资源占用，提高网络的运行速度。相比传统的剪枝和知识蒸馏等方法，轻量化模型设计在语义分割中使用卷积神经网络，不仅能提升分割精度，还能同时识别多个目标，是目前主流的研究方向。

文献［1］通过选择性搜索方法对图像进行分割和候选区域生成，更高效地定位感兴趣区域。在此基础上这些感兴趣区域与神经网络生成的特征图之间的对应关系。文献［2］提出了一种直接使用固定模板处理输入特征图的方法。与选择性搜索不同，固定模板通过多尺度的固定模板可以更加全面地扫描输入特征图，识别潜在的目标区域，大幅提高网络的检测精度。近年来，研究人员提出了很多基于深度学习网络的语义分割算法，文献［3］提出了空洞卷积模块，该模块通过引入空洞率使得尺寸大小相同的卷积核能获得更大的感受野，同时不会增加额外的参数。文献［4］通过结合深度卷积和空洞卷积模块来扩大模型的感受野，同时不会增加额外的参数量。这种结合保证了在使用较少参数的情况下实现高精度的分割效果。然而，作者所使用的空洞卷积的空洞率都是相同的，无法捕捉到不同大小感受野的信息。文献［5］设计了一种对称卷积结构来完成语义分割任务，该结构先使用卷积来提取特征，再使用反卷积处理特征。文献［6］提出了一种大卷积结构，该结构通过扩大感受野，更好地捕捉全局信息，同时保持定位精度，从而缓解语义分割任务中定位与分类之间的矛盾。文献［7］提出了 RefineNet 模块，采用逐层细化的方法，通过不断向网络输入更精细的像素级信息，从而得到更精准的语义分割结果。但是处理像素级的图像往往需要较大的计算资源，当前的移动设备和嵌入式系统通常无法提供足够的计算能力和存储空间来支持这种计算密集型操作。近年来，为了在移动设备上更高效地运行卷积神经网络，出现了许多兼顾运行速度和精度的网络，典型的模型包括 VGG-16[8]、GoogLeNet[9]网络等。通过这些创新方法，研究人员试图在提高分割精度的同时，优化计算效率，满足实时应用的需求。

而注意力机制最开始应用于自然语言处理领域。研究人员发现，将注意力用于计算机视觉领域也有着不错的表现。2021 年，文献［10］在语义分割任务中引入了注意力机制，取得了显著的效果。注意力机制能够有选择地对输入图像进行特征提取，使网络能够集中在更重要的信息上，从而增强特征。2018 年，文献［11］提出了 SE 注意力机制，该模块通过自适应地学习通道权重来增强特征。利用 SE 注意力机制，网络能够学习全局信息，从而选择性地提取通道里的重要特征，并抑制不重要的特征信息。2018 年，文献［12］提出了 CBAM 注意力机制，该机制能够在空间和通道上提取重要特征，来提高网络的感知能力，该模块可广泛应用于提高 CNN 网络的特征表达能力。

综上所述，对于实时语义分割任务而言，很有必要研究轻量化模型架构来满足其在现实环境中的应用需求。本文针对室内环境分割精度差、计算资源消耗等问题，在 Mask R-CNN[13]的基础上提出了一种基于自适应注意力机制的轻量化改进网络结构，本文主要工作如下：

1）提出了一种轻量化的特征提取网络 ResNet-DS-tiny。该网络在 ResNet-50[14]的基础上，结合深度可分离卷积和空洞卷积模块来替代 ResNet-50 里的传统卷积，这里使用不同空洞率的空洞卷积来提取不同大小感受野的特征信息。与传统卷积相比，改进后的卷积不仅能捕获不同区域的特征，同时还能减少计算资源消耗，提升了网络的运行速度。

2）在 ResNet-50 网络的输出端，引入了自适应通道注意力模块，该模块能够有选择性地识别和提取每个通道的重要性特征，同时抑制不重要的特征，旨在提升网络对特征的表达能力。接着，将经过自适应通道注意力模块增强后的特征与原始特征图以加权的方式进行融合，从而进一步提高特征的表达能力。改进后的轻量化神经网络在语义分割任务中表现出更好的计算效率和分割精度。

1 基本原理

1.1 分组卷积

传统的卷积在所有的输入特征图上做卷积，特征图的每个通道和输入特征图的所有通道都有关，这是一种通道密集连接的方式，如图 1（a）所示。分组卷积对输入特征进行分组，然后对每个分组特征使用不同大小的卷积核分别进行卷积操作来提取特征，减少参数和计算量的同时，提高了网络的识别精度。利用分组卷积能够生成互补且特征信息不同的特征图，能更全面准确地表示输入图像特征。分组卷积的原理如图 1（b）所示，先将输入特征图进行分组，然后对每个分组使用不同的卷积核进行卷积计算。

分组卷积能够减少模型的参数量和计算量，而普通卷积的参数量如式（1）所示：

$$H \times W \times C_1 \times C_2 \tag{1}$$

其中：$H \times W$ 是输入特征图的大小，C_1 与是输入特征的通道数，C_2 是输出特征的通道数。

分组卷积的参数量如式（2）所示：

$$H \times W \times \frac{C_1}{G} \times \frac{C_2}{G} \times G = H \times W \times C_1 \times C_2 \times \frac{1}{G} \tag{2}$$

由上式可知，与普通卷积相比，分组卷积的计算量和参数量减少为 $1/G$，G 为设置的分组数。这是由于通过卷积分组减少了输入通道数，同时减少了网络计算量。

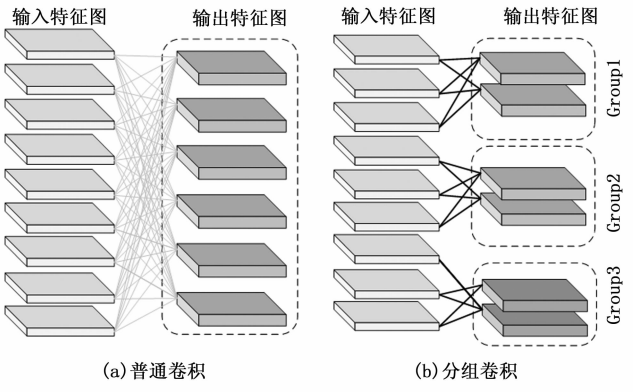

图 1　分组卷积示意图

1.2　深度可分离卷积

基于深度卷积神经网络的算法往往存在计算资源消耗，硬件要求高的问题。针对这些问题，近年来研究人员提出了一系列轻量化卷积神经网络模型，如轻量级的 MobileNet[15-17] 系列和 ShuffleNet[18-19] 系列。这两类轻量化网络架构都使用了深度可分离卷积来减少网络计算量。深度可分离卷积是一种卷积神经网络中常用的轻量化卷积操作，主要用于减少参数数量和计算量，从而加速模型训练和推理过程。主要由两部分组成：深度卷积和逐点卷积。深度卷积对每个输入通道在空间上进行了卷积操作，用来提取空间特征，卷积核数量和输入通道数相同。逐点卷积将深度卷积输出的特征图映射到新的特征空间，对深度卷积的输出使用 1×1 大小的卷积核来聚合所有通道，用来提取通道特征。普通卷积和深度可分离卷积过程对比如图 2 所示。

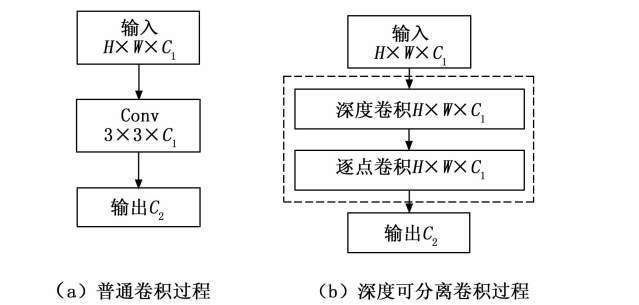

图 2　普通卷积与深度可分离卷积

普通卷积与深度可分离卷积参数量与计算量如式（3）（4）所示：

$$\frac{K_2C_1+C_1C_2}{K^2C_1C_2}=\frac{1}{C_2}+\frac{1}{K^2}\sum_{i=1}^{n}X_iY_i \tag{3}$$

$$\frac{WHK^2C_1+WHC_1C_2}{WHK^2C_1C_2}=\frac{1}{C_2}+\frac{1}{K^2} \tag{4}$$

由上式可知，深度可分离卷积的计算量与普通卷积相比，参数量仅为普通卷积的 $1/K^2$，K 为卷积核的个数。可以看出深度可分离卷积操作具有更少的计算量和参数量。另外在 MobileNet V1 中，利用深度可分离卷积代替普通卷积，特征提取精度并没有折损很多。因此，理论证明利用深度可分离卷积代替 ResNet-50 里的普通卷积不仅能减少网络的计算量和参数量，同时还能兼顾特征精度。

2　轻量级 Mask R-CNN 网络

2.1　Mask R-CNN

Mask R-CNN 是一种二阶段的实例分割算法，在第一个阶段，扫描整个图像提取特征，利用 RPN 网络生成多个候选提议框，第二阶段主要是网络对第一阶段生成的区域建议进行进一步处理分类，生成边界框和掩码。Mask R-CNN 扩展自 Faster R-CNN 框架。Faster R-CNN 是一种主流的目标检测框架，Mask R-CNN 在此基础上增加了生成分割掩码的能力，将其扩展为实例分割框架。Mask R-CNN 使用 ResNet-50 特征提取网络作为主干网进行特征提取，通过引入残差块，使得其参数量和计算量相对较少，能够在保证特征提取能力的同时，保持较高的计算效率。残差连接使得网络在深度增加时，仍能有效地进行训练。ResNet-50 由多个卷积层和残差块组成，可以提取图像的多层次特征。这些特征包含了从低层的边缘、纹理信息到高层的语义信息，为后续的目标检测和实例分割提供了丰富的信息基础。Mask R-CNN 在分割精度上都取得了不错的效果。

2.2　ResNet-DS-tiny

Mask R-CNN 网络中采用的是 ResNet-50 作为残差模块，该模块虽然改进了由于网络深度增加而引起的训练困难，其主要表现为“梯度弥散”或“梯度爆炸”。但其同时也增加了网络的复杂度，导致在分割时实时性不高。ResNet-50 网络主要是由一系列的残差模块组成，如图 3 所示。

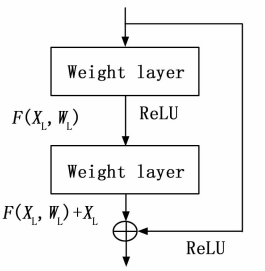

图 3　残差模块

本研究针对的是室内环境，此环境内需要被分割的对象较少，品类较为单一，在这种情况下，为保证分割系统的实时性，提出一种将 ResNet-50 采用分组卷积及深度可分离卷积的方式并引入空洞卷积以扩大感受野范围对 Mask R-CNN 网络进行轻量化改进，提出了一种轻量级的残差网络 ResNet-DS-tiny（ResNet with depthwise separable convolutions），如图 4 所示。

本文提出了一种轻量化特征提取网络 ResNet-DS-tiny，基于当前轻量化分类模型中广泛应用的深度可分离卷积模块进行模块设计，目的是提升语义分割网络的实时分割性能。该网络采用模块化设计，通过有序连接多个深度可分

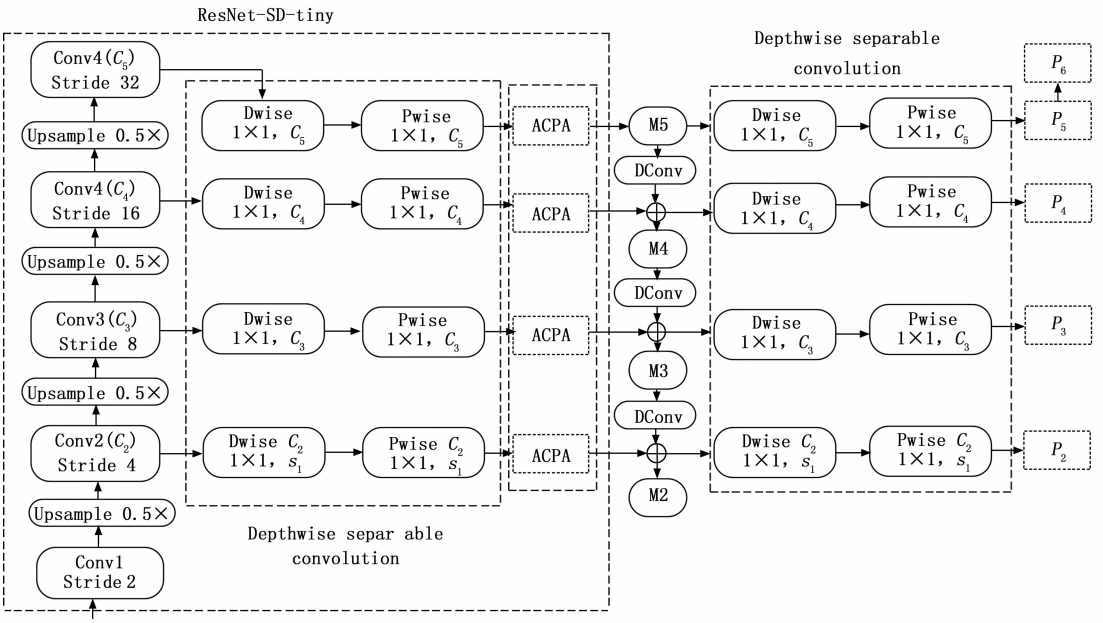

图 4 ResNet-DS-tiny 卷积结构

离卷积模块，代替原有的普通卷积构建了 ResNet-DS-tiny 整体的特征提取结构。每个深度可分离卷积模块包含两个分支，分别使用大小为 3×3 的深度卷积（卷积核数量和输入通道数相同）和卷积核大小为 1×1 的逐点卷积。利用深度卷积和逐点卷积进行卷积运算，随后在通道上对特征进行拼接。通过卷积核大小为 1×1 的卷积对获得的特征图从通道上进行融合，最终输出特征。这种设计可以从空间和通道两个特征维度提取特征，使用不同大小的卷积核可以实现多尺度特征的感受野融合，从而可以准确提取不同尺度的目标，还保证了特征提取过程中的信息保留和有效融合。总体而言，ResNet-DS-tiny 通过优化深度可分离卷积模块的设计，提升了网络在多尺度目标检测中的表现，实现了轻量化与高效性的平衡。

然而，简单地将其用来替换普通卷积可能会导致网络的特征提取性能下降。由于深度可分离卷积结构的限制，可能无法捕捉到某些复杂特征之间的高级关联性，特别是在需要长距离依赖关系的任务中。而注意力机制可以解决这个问题。因此，本文在网络输出端引入了自适应通道注意力机制 ACPA 模块，该模块能够自适应地学习通道之间的权重，有针对性地强调重要的特征信息，抑制不重要的特征信息。结合深度可分离卷积，可以更好地平衡轻量化和特征表达能力之间的关系，从而提高网络的性能。

此外，在语义分割模型中，多尺度信息的提取至关重要，因此需要在设计中充分考虑这一点。当前流行的目标检测模型通常使用深层主干网络，本文选择了 Mask R-CNN，其主干网络为 ResNet-50。深层网络的普遍问题是随着网络深度的增加，感受野随之扩大，导致高层特征图中的目标位置信息大部分丢失。原始 Mask R-CNN 通过特征金字塔在多个尺度进行采样，尽管能进行多尺度融合，但会导致信息损失。为解决这一问题，本文结合空洞卷积[20]的特点，使用不同空洞率大小的空洞卷积代替特征金字塔的池化层部分，形成空洞特征金字塔，以在增大不同感受野的同时保留更多目标位置信息。

通过结合深度可分离卷积和空洞卷积的方式，在两个分支上的卷积层中使用不同空洞率的空洞卷积代替池化层来扩大模型感受野，用于增强模型在多尺度信息获取方面的能力。本文分别设置特征金字塔中的空洞卷积的空洞率大小依次为 2、4、8、16，通过逐步扩大特征金字塔感受野来获取多尺度特征信息。然后，合并两个分支，恢复原始的通道数。普通卷积如图 5（a）所示，空洞卷积如图 5（b）所示。

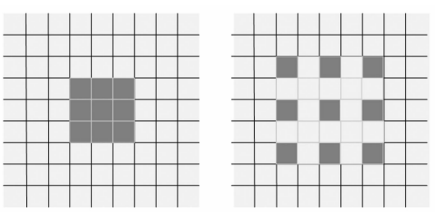

图 5 卷积率＝2 空洞卷积示意图

由于两个分支分别使用不同空洞率大小的空洞卷积来扩大感受野，以获取多尺度的信息。因此使用 1×1 的卷积对可以融合两个分支的特征信息，有助于促进信息交流，增强特征之间的表达能力。本文结合深度可分离卷积和空洞卷积，既扩大了模型的感受野，使其能够有效地获取多尺度的信息，又减少了网络的计算量和参数量，从而提高了运行效率。在输出端引入了自适应注意力机制，可以根

据特征的重要性加权特征图，增强特征的表达能力，进一步提升了网络的精度和性能。

3　实验结果与分析

实验平台硬件配置为：Inter（R）corei5 3.50 GHz，NVIDIA GeForce RTX 3050 Laptop 显卡，软件系统为 Ubuntu 20.04，Tensorflow 深度学习框架。

实验数据集为语义分割常用的 COCO 数据集。本实验使用 COCO 数据集中图像分割部分作为测试集，该部分包含 91 种类别（banner，blanket，branch，bridge，building-other，bush，cabinet 等）。

3.1　评价指标

实验采用了图像语义分割常用的 5 个性能评价指标：交并比（IoU，intersection ratio）、平均交并比（mIoU，average intersection ratio）、像素精度（PA，pixel accuracy）、平均像素精度（MPA，average pixel accuracy）、模型运算帧率（FPS，frames processed per second）。

在图像语义分割中，交并比（IoU）是衡量预测分割区域与真实分割区域重叠程度的指标，则平均交并比（mIoU）对每个类别的交并比取平均值，以衡量模型整体分割性能。如式（5）所示：

$$mIoU\% = \frac{1}{K+1}\sum_{i=0}^{K}\frac{P_{ii}}{\sum_{j=0}^{K}P_{ij}+\sum_{j=0}^{K}P_{ji}-P_{ii}}\times 100\% \quad (5)$$

式中，P_{ii} 表示属于类 i 且被正确分类为分类 i 的像素数量，称为真正（True Positives）；P_{ij} 表示属于类 j 却被错误分类为类 i 的像素数量，称为假负（False Negatives）；P_{ji} 表示属于类 i 却被错误分类为类 j 的像素数量，称为假正（False Positives）；P_{jj} 表示属于类 j 且被正确分类为类 j 的像素数量，称为真负（True Negatives）。其中，类 i 为正类，类 j 为负类，$K+1$ 代表分割类别的总数量，其中 1 表示一个背景类。

像素精度（PA）和平均像素精度（MPA）用于评估图像分割算法的性能。像素精度表示所有像素中正确分类的百分比，而平均像素精度则是各类物体像素精度的平均值。像素精度和平均像素精度计算式如式（6）（7）：

$$PA\% = \frac{\sum_{i=0}^{K}P_{ii}}{\sum_{i=0}^{K}\sum_{j=0}^{K}P_{ij}}\times 100\% \quad (6)$$

$$MPA\% = \frac{1}{K+1}\sum_{i=0}^{K}\frac{P_{ii}}{\sum_{j=0}^{K}P_{ij}}\times 100\% \quad (7)$$

模型运算帧率（FPS）是反映网络运行速度的重要指标，运算帧率越高模型的性能越好。计算公式如式（8）：

$$FPS = \frac{N}{\sum_{i=1}^{N}T_i} \quad (8)$$

其中：N 为图像数量，T_i 为网络处理第 i 帧图像所用的时间。

3.2　实验结果与分析

参数量与计算量是评价网络是否轻量的重要指标，首先测算 ResNet-DS-tiny 与其他常见卷积网络的参数量及浮点数进行对比（浮点数反应网络的计算量）。假设输入特征图大小为 $H_i\times W_i\times C$，卷积核的尺寸大小为 $k\times k$，输出特征的通道数为 C_2，输出特征图的大小为 $H_o\times W_o\times C_2$，深度可分离卷参数量 P 和每秒浮点运算次数 F 如式（9）（10）：

$$P = P_{DW}+P_{PW} = k_H\times k_W\times C_1 + C_1\times C_2 \quad (9)$$

$$F = P\times H_o\times W_o = (k_H\times k_W\times C_1 + C_1\times C_2)\times H_o\times W_o \quad (10)$$

式中，P 为总参数量，P_{DW} 为深度卷积的参数量，P_{PW} 为逐点卷积的参数量，F 为浮点数。

表 1　网络参数量、浮点数对比

	参数量（$\times 10^6$）	浮点数（$\times 10^8$）
ResNet-50	42.5	73.4
Mobile Net-v2	3.34	3.19
ResNet-DS-tiny	2.28	2.11
ResNet-DS-tiny＋ACPA	2.3	2.13

本文将 ResNet-DS-tiny、ResNet-50、Mobile Net-v2 以及加入 ACPA 模块的 ResNet-DS-tiny 参数量、浮点数在 COCO 数据集上进行对比实验，实验结果具体见表 2。

由表 2 可知，Mask R-CNN 中所使用的特征提取网络 ResNet-50 地参数量与浮点数相较于其他轻量级的语义分割网络所使用的主干网络高出一个数量级。本文所提的 ResNet-DS-tiny（无 ACPA 模块）相较于经典轻量化网络 Mobile Net-v2 参数量减少了 31.7%，浮点数减少了 33.9%，可见 ResNet-DS-tiny（无 ACPA 模块）满足轻量级语义分割网络标准。

对于语义分割网络，分割精度是十分重要的评价网络性能指标。本文对 Mask R-CNN 与 Mobile Net-v2 以及轻量级 Mask R-CNN（含 ACPA 模块）、轻量级 MaskR-CNN（不含 ACPA 模块）的 IoU、$mIoU$、PA、MPA 进行对比，实验结果见表 3。

表 3　网络分割精度对比

模型	编码网络	自适应通道注意力机制	IoU/%	$mIoU$/%	PA/%	MPA/%
Mask R-CNN	ResNet-50	×	76.3	44.8	73.1	46.2
Mobile Net-v2	Mobile Net	×	68.5	40.6	69.4	38.7
轻量级 Mask R-CNN	ResNet-DS-tiny	×	69.8	41.8	69.5	39.5
轻量级 Mask R-CNN	ResNet-DS-tiny＋ACPA	√	76.2	43.9	73.1	45.9

由表 3 可知，非轻量化的 Mask R-CNN 的 *mIoU* 与 *MPA* 比众多轻量级网络精度略高，本文所提轻量级 Mask R-CNN 网络相较于 Mobile Net-v2 *mIoU* 提升了 1.2%，*MPA* 提升了 1.2%。含有 ACPA 模块的轻量级 Mask R-CNN 网络精度与原始 Mask R-CNN 网络基本持平，相较于不含 ACPA 模块的网络指标也均有提升。综合表 2，含有 ACPA 模块的 ResNet-DS-tiny 网络相较于无此模块的网络参数量与浮点数略有提升，在网络复杂度提升不大的情况下，获取更好的网络性能是值得的。

语义分割网络的运行时间、内存占用及每秒系统可处理的图像帧数也是评价网络是否轻量的重要指标。本文实验验证了 Mask R-CNN 与 Mobile Net-v2 以及轻量级 Mask R-CNN（含/不含 ACPA 模块）的上述指标，实验结果见表 4。由表 4 可知，在分割时间方面 Mask R-CNN 比轻量级的网络高出一倍，内存占用比轻量级网络高出一个数量级，FPS 是其他网络的一半。本文所提轻量化 Mask R-CNN 网络相较于 Mobile Net-v2 在分割时间、内存占用、帧数这 3 个指标分别提升了 13.3、27.2、4.3%。综上，可见本文所提轻量化 Mask R-CNN 网络在网络精度与性能提高的前提下，满足轻量化网络的标准。

表 4 网络运行时间、内存占用、帧数对比

模型	分割时间（单帧）/ms	内存占用 /MB	帧数 /s
Mask R-CNN	63	32.85	30
Mobile Net-v2	34	5.84	69
轻量级 Mask R-CNN(无 ACPA)	29	4.13	73
轻量级 Mask R-CNN	30	4.25	72

4 结束语

本文提出一种含有自适应通道注意力机制的轻量级 Mask R-CNN 网络，通过引入深度可分离卷积、分组卷积以及不同空洞率的空洞卷积对 Mask R-CNN 进行了轻量化改进。特别设计了轻量化特征提取网络 ResNet-DS-tiny，有效减少了参数量和计算量。并加入自适应通道注意力机制。在自适应通道注意力模块中，利用加权方式对输入的 RGB-D 图像从空间和通道赋予不同的权重，增强了特征的表达能力。此外，本文还提出了空洞特征金字塔网络，有效聚集了多尺度特征，并相较于传统特征金字塔池化模块减少了参数量，提高了 RGB 信息特征的获取充分性，增强了分割系统的实时性并减少了资源占用。相比于 Mask R-CNN，本文提出的语义分割模型在分割时间和精度上均有显著提升。

参考文献：

[1] GIRSHICK R. Fast R-CNN [C] //Proceedings of the IEEE international conference on computer vision (ICCV). Santiago, Chile: IEEE, 2015: 1440 - 1448.

[2] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (6): 1137 - 1149.

[3] CHEN L, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets and fully connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40 (4): 834 - 848.

[4] WANG Y, ZHOU Q, LIU J, et al. Lednet: a lightweight encoder-decoder network for real-time semantic segmentation [C] //Proceedings of the IEEE International Conference on Image Processing. Taibei, Taiwan: IEEE, 2019: 1860 - 1864.

[5] NOH H, HONG SEUNGHOON, HAN BOHYUNG. Learning deconvolution network for semantic segmentation [C] //Proceedings of the International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 1520 - 1528.

[6] PENG C, ZHANG X Y, YU G, et al. Large kernel matters: Improve semantic segmentation by global convolutional network [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 1743 - 1751.

[7] LIN G S, MILAN A, SHEN C, et al. RefineNet: Multi-path refinement networks for high resolution semantic segmentation [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 5168 - 5177.

[8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C] //Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Kuala Lumpur, Malaysia: IEEE, 2015: 730 - 734.

[9] SZEGEDY C, LIU W, JIA Y G, et al. Going deeper with convolutions [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, 2015: 1 - 9.

[10] YUAN Y, HUANG L, GUO J, et al. OCNet: object context network for scene parsing [EB/OL]. (2021 - 03 - 15) [2024 - 7 - 19]. https://arxiv.org/ abs/1809.00916.

[11] HU J, SHEN L, SUN G. Squeeze and excitation networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 7132 - 7141.

[12] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C] //Proceedings of the European Conference on Computer (ECCV). Munich, Germany: Springer, 2018: 3 - 19.

[13] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [C] //Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2961 - 2969.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770 - 778.

（下转第 235 页）