[文章编号] 1003－4684(2026)04-0012-08

# 基于轻量化及多尺度特征提取的城市街景语义分割

丁 江[1]，崔 浩[1]，戴俊伟[1]，李利荣[1,2]，巩朋成[3]，宋海娜[1,2]

(1 湖北工业大学电气与电子工程学院，湖北 武汉 430068；
2 湖北工业大学湖北省电网智能控制与装备工程技术研究中心，湖北 武汉 430068；
3 武汉工程大学计算机科学与工程学院，湖北 武汉 430205)

[摘 要] 目前在移动端或者嵌入式设备等平台部署高性能神经网络算法是一大趋势。针对常规分割模型复杂度较高的问题，本文提出了一种基于轻量化及多尺度特征提取的城市街景语义分割算法。该算法采用轻量级网络 MobileNetV3 作为特征提取网络，引入特征金字塔网络在其横向连接处嵌入混洗注意力，加强信息融合，接着在其深层特征后堆叠三次深度不对称门控模块，进行局部信息和上下文信息的交互；然后采用多尺度膨胀金字塔模块，扩大感受野，获取多尺度特征信息，最后将不同层次的信息拼接融合；训练阶段，本文将 Ce Loss 和 Dice Loss 相结合，提高训练效率，解决了正负样本不平衡问题，避免梯度饱和。最终在 CamVid 数据集上进行了充分的验证实验，结果表明本文算法的 mIoU 达到约 70.18%，而模型参数量约为 4.44M，推理速度约为 64.1FPS，优于当前主流算法。

[关键词] 轻量化网络；城市街景；语义分割；膨胀卷积；多尺度特征
[中图分类号] TP391.4 [文献标识码] A

语义分割是视觉场景解析的一项基本任务，其目标是为给定图像中的每一个像素分配一个特定的类别标签。而随着智能化技术的不断提高，精准的语义分割已被应用到自动驾驶[1]、医学成像诊断[2]和遥感图像[3]等诸多领域。从 FCN[4] 开始，深度学习算法在语义分割领域逐渐占据主导地位，一系列基于深度学习的语义分割算法[5-8]如雨后春笋般喷涌而出。DeepLabv3[7] 通过空洞空间卷积池化金字塔(atrous spatial pyramid pooling，ASPP)来扩大感受野。PSPNet[8] 提出了金字塔池化模块，该模块通过应用平均池化得到具有不同输入分辨率的卷积层的并行分支。这些分割网络模型的发展表明，具有优秀分割性能的网络架构往往需要在不丢失重要信息的情况下大规模学习像素之间的上下文相关性和长远距离的依赖关系。尽管这些模型实现了令人满意的分割精度，但由于牺牲了太多的空间和时间复杂性，这大大限制了它们在移动设备上的应用。

近些年，针对道路场景解析，也提出了一些具有竞争力的语义分割算法。如 ERFNet[9] 和 ESP-Net[10] 提出了利用轻量级骨干的紧凑型语义分割网络。还有为了满足嵌入式设备或移动设备的需求，研究人员开发从头开始训练的复杂轻量级编码器模型。ICNet[11] 用复杂和深层路径对小尺寸输入进行编码以解析语义信息，并利用简单和浅层路径对大尺寸输入的细节进行编码。LEDNet[12]、DAB-Net[13] 引入了一个轻量级模型，以解决深度卷积神经网络使用的计算负担大的问题。此外，上下文信息也被证实对于场景解析任务非常关键。$S^2$-FPN[14] 为了对全局上下文进行编码，并对远程依赖关系进行建模，提出了注意力金字塔融合模块，并在解码阶段采用全局特征上采样来对注意力金字塔融合模块和编码器得到的特征进行融合。当前，研究人员还提出了一些基于双分支网络的出色网络架构，并在速度和准确性方面取得了先进的性能[15-17]。DDRNet23[18] 使用两个分支，一条分支保持 1/8 的高分辨率以保存更多的空间信息，另一条分支进行多次下采样获得更多的语义信息，它们之间有多个双边融合，并在整个网络架构的末端附加一个深度聚合金字塔池化模块，以扩大有效的感受野。在语义分割中为了减少参数量和计算量，得到的特征图往往是经过多次下采样的，这导致大量全局信息的损失，而膨胀卷积相比于普通卷积，在参数量一致的

[收稿日期] 2023－10－21
[基金项目] 国家自然科学基金(62071172，62202148)
[第一作者] 丁 江(1998－)，男，湖北武汉人，湖北工业大学硕士研究生，研究方向为城市街景语义分割。
[通信作者] 李利荣(1974－)，女，湖北武汉人，工学博士，湖北工业大学讲师，研究方向为计算机视觉、图像处理。

情况下具有更大的感受野，即具有更丰富的全局信息，因此本文采用了膨胀卷积来解决下采样过程中全局信息损失的问题。本文受 DeepLabV3＋启发，采用双分支结构，提出了一种基于轻量化及多尺度特征提取的城市街景语义分割算法，提高精度的同时减少整体网络的复杂度。本文算法利用轻量级网络 MobileNetV3 作为特征提取网络，结合改进的特征金字塔（improved feature pyramid network，IFPN）进行特征初步提取，同时结合提出的深度不对称门控模块（depth-wise asymmetric gating module，DAGM）和多尺度膨胀金字塔模块（multi-scale dilation pyramid module，MDPM），获得较为丰富的全局信息和上下文信息，并将浅层特征具有的丰富空间信息和得到的上下文信息进行融合，同时本文还将 Dice Loss[19] 和 Ce Loss 相结合使用提高分割效果。

## 1　本文算法总体框架

本文的总体框架如图 1 所示，特征提取网络采用轻量级网络 MobileNetV3，通过 IFPN 得到特征 $P_1$ 和 $P_4$，特征 $P_1$ 作为第一分支通过 DAGM 进一步提取局部信息和上下文信息得到特征 $P_2$，再利用 MDPM 的不同膨胀率，提取特征的不同感受野信息并进行上采样，得到特征 $P_3$；将特征 $P_4$ 所处分支作为第二分支，与第一分支以拼接方式融合得到特征 $P_5$，最后加入两个连续的 3×3 卷积细化目标周围边界，并最终上采样保持与原图大小相同的输出特征图。

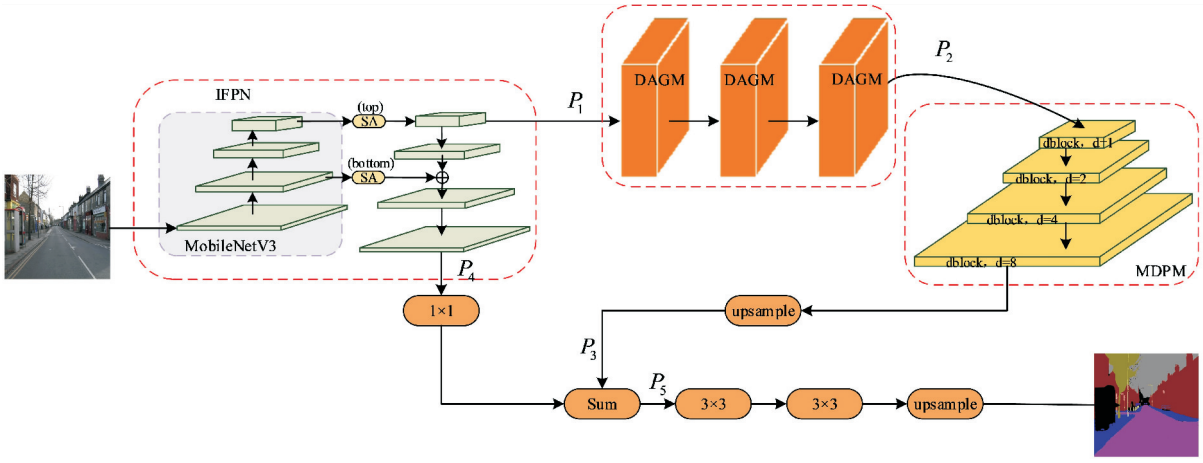


图 1　城市街景语义分割总体框架

### 1.1　改进的特征金字塔网络

本文结合 MobileNetV3 提出了改进的特征金字塔网络，如图 2 所示。特征金字塔网络主要分为自下向上的一条路径和自上到下的一条路径。自下向上就是深度卷积网络的前向提取特征的过程，自上而下则是对最后卷积层的特征图进行上采样的过程，而横向连接是融合深层特征和浅层特征的过程。

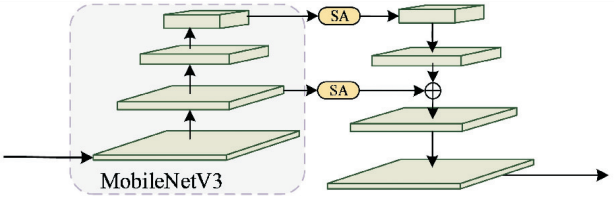


图 2　改进的特征金字塔网络

由于不同层之间存在语义鸿沟，直接融合会降低多尺度表达能力，不同层之间信息的不连续性也会造成特征提取不够充分。本文在特征金字塔网络的横向连接的两条分支上加入混洗注意力（shuffle attention，SA）[20]，以完善这一问题。SA 是在空间与通道注意力的基础上，引入特征分组与通道混洗，得到的一种超轻量型的注意力机制，具有更低的计算复杂度和参数量，因此更适用于移动端，其模型如图 3 所示。首先将初始特征 $P$ 作为 SA 的输入，并将 $P$（通道数为 $c$）沿通道维度分为 $g$ 组，每组子特征可表示为 $P_k \in \mathbb{R}^{c/g\times H\times W}$，并将 $P_k$ 分成两个子分支 $P_{k1}$、$P_{k2}$，其中分支 $P_{k1}$ 用于计算通道注意力，先采用全局平均池化（global averaging pooling，GAP）来嵌入全局信息，如：

$$m = \Gamma_{GAP}(P_{k1}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}P_{k1}(i,j) \tag{1}$$

其中，$\Gamma_{GAP}(\cdot)$ 为全局平均池化操作；$H$、$W$ 为输入特征图的高与宽。再对通过简单的缩放平移后的特征使用 Sigmoid 激活函数得到通道注意力权值向量，最后将其与 $P_{k1}$ 进行相乘运算，得到最终的通道注意力特征图 $P'_{k1}$，如：

$$P'_{k1} = \sigma(\psi(m))\cdot P_{k1} = \sigma(W_1 m + b_1)\cdot P_{k1} \tag{2}$$

其中，$\psi(\cdot)$ 为缩放与平移操作；$\sigma(\cdot)$ 为 Sigmoid 激活函数；$W_1, b_1 \in \mathbb{R}^{c/2g\times 1\times 1}$ 是用于缩放和平移的可学习参数。

$P_{k2}$ 分支用于计算空间注意力。首先对 $P_{k2}$ 采

用组正则化(group normalization,GN)操作获取相关的空间信息,接着通过采用相似的缩放平移和激活函数的过程得到空间注意力权值,将权值与 $P_{k2}$ 进行相乘,得到空间注意力特征图 $P'_{k2}$ 。

$$P'_{k2}=\sigma(W_2\cdot GN(P_{k2})+b_2)\cdot P_{k2} \tag{3}$$

其中,$W_2,b_2\in\mathbb{R}^{c/2g\times1\times1}$ 是用于缩放和平移的可学习参数。

其次将得到的注意力特征图 $P'_{k1}$ 与 $P'_{k2}$ 在通道维度进行拼接,其他组的子特征均重复上述操作。最后将 $g$ 组子特征进行通道混洗,从而增强不同组特征之间的信息交互,建立远程的依赖关系,得到细化后的特征 $P'$,此时的特征 $P'$ 包含的特征信息充分,建立的远程依赖关系也能提高多尺度表达能力。

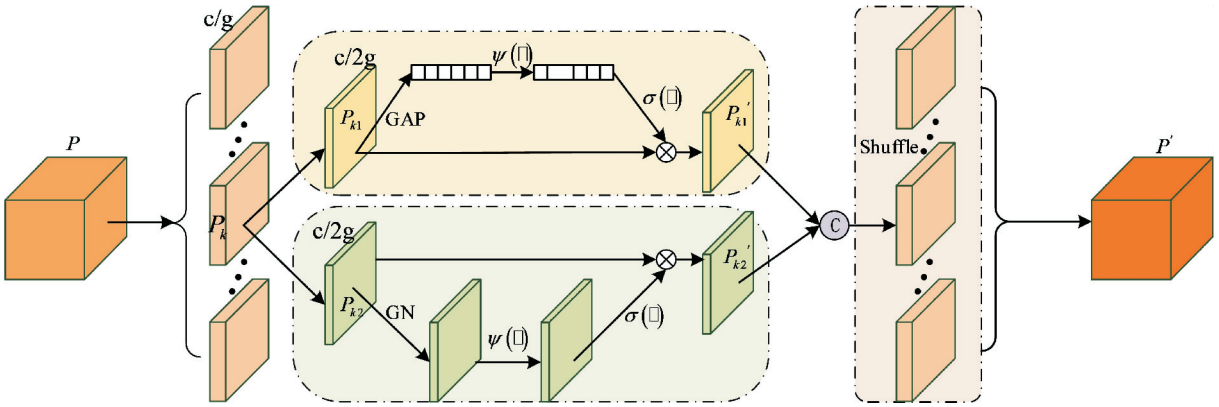


图 3 SA 结构示意图

## 1.2 深度不对称门控模块

在许多视觉任务中,感受野是一个关键影响因素,输出特征图中的每一个像素必须对应图像中足够大的区域,以捕获关于场景中的全局信息。以往常常通过卷积或者池化操作以达到增大感受野的目的。常规卷积操作可以保留特征信息,但会增加参数量,池化操作则与之相反。膨胀卷积能保持在相同参数量的情况下增大特征图的感受野,而深度可分离卷积又可以有效的减少参数量,因此设计膨胀卷积和深度可分离卷积的组合操作是可改进的工作。本文设计了一种深度不对称门控模块(DAGM),如图 4 所示,其中 DSC 表示深度可分离卷积(depthwise separable convolution, DSC),GCT[21]表示门控通道转换,$d$ 表示膨胀系数,$c$ 表示通道数。本文堆叠使用了三次深度不对称门控模块,以提取更多有用信息。

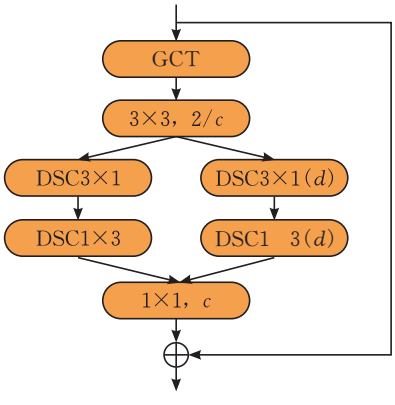


图 4 深度不对称门控模块

**1.2.1 门控通道转换** 门控通道转换(gated channel transformation,GCT)是一种简单有效的通道间关系建模体系结构,其将归一化和注意力机制结合起来,并使用轻量级的、易于分析的变量来隐式地学习网络通道间的相互关系,也可以有效和准确地进行上下文信息建模。本文算法在深度不对称模块的 3×3 卷积之前使用门控通道转换,有助于信息交互。GCT 的结构如图 5 所示。设 $x=[x_1,x_2,\cdots,x_C]$,$x_c=[x_c^{i,j}]_{H\times W}\in\mathbb{R}^{H\times W}$,$c\in\{1,2,\cdots,C\}$,其中 $x_c$ 对应 $x$ 的每个通道。大感受野有助于避免局部混淆,因此首先对各个通道的全局上下文信息进行聚合,全局上下文嵌入模块定义如下:

$$m_c=\alpha_c\parallel x_c\parallel_2=\alpha_c\left\{\sum_{i=1}^{H}\sum_{j=1}^{W}(x_c^{i,j})^2+\varepsilon\right\}^{\frac{1}{2}} \tag{4}$$

其中,$\varepsilon$ 是一个小常数以避免导数为 0。使用可训练的参数 $\alpha_c$ 来控制每个通道的权值,因为不同的通道具有不同的意义。

归一化方法可以创建通道之间的竞争关系,该模块使用 $L_2$ 范数来跨通道操作,令 $m=[m_1,m_2,\cdots,m_C]$,则信道归一化公式为:

$$\hat{m}=\frac{\sqrt{C}m_c}{\parallel m\parallel_2}=\frac{\sqrt{C}m_c}{\left[\left(\sum_{c=1}^{C}m_c^2\right)+\varepsilon\right]^{1/2}} \tag{5}$$

该模块还采用一种门控机制,来适应原始特征。设门控权值 $\gamma=[\gamma_1,\gamma_2,\cdots\gamma_C]$,门控偏差 $\beta=[\beta_1,\beta_2,\cdots,\beta_C]$,则门控函数定义为:

$$\hat{x}=x_c[1+\tanh(\gamma_c\hat{m}_c+\beta_c)] \tag{6}$$

GCT 通过将归一化方法和门控机制相结合,可以对不同通道之间更多类型的关系进行建模。

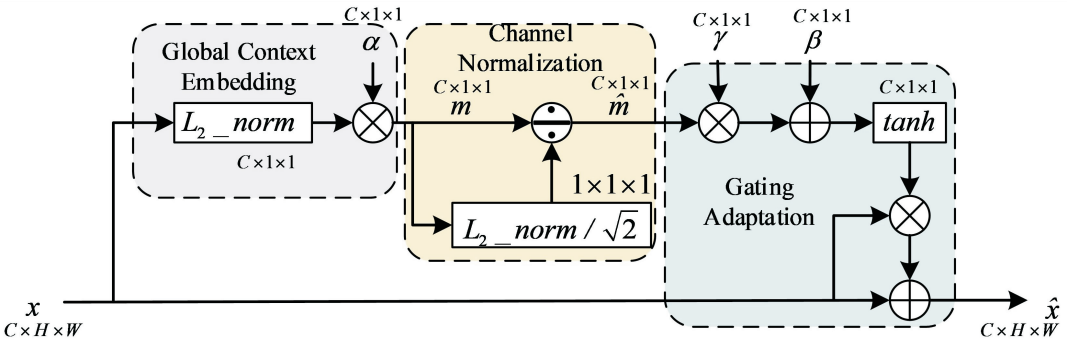


图 5　GCT 结构图

**1.2.2　深度不对称模块**　本节模块采用一种两分支结构，一个分支提取局部信息，而另一个分支提取上下文信息，然后将两个分支相加，进行信息的融合。为了提取局部信息同时进一步减少计算量，用 $n\times1$ 和 $1\times n$ 的深度卷积代替标准的 $n\times n$ 深度卷积，这种卷积方式将计算复杂度从 $O(N^2)$ 降低到 $O(N)$。为了提取更广泛的上下文信息，本模块还采用了膨胀卷积，它在不降低特征图分辨率的情况下获得更大的感受野。然而，当膨胀率变大时，也需要实现大量数据填充来保持特征图的大小，这将带来一定的计算成本。为了一定程度上避免上述矛盾，仅将膨胀卷积应用于右侧分支，因此被称为“深度不对称模块”。

### 1.3　多尺度膨胀金字塔模块

长距离的上下文信息以及多尺度信息对于分割结果十分重要，良好的预测结果通常来自不同层信息的组合，因此本节重点研究如何有效的利用多尺度信息。何凯明团队在 2020 年发表的 RegNet[22] 中提出许多网络设计架构，根据其经验将 block 设计成带有组卷积的残差结构，并分为主分支和捷径分支这两个分支。主分支按顺序是由一个 1×1 常规卷积、一个 3×3 的组卷积，再接一个 1×1 常规卷积组成，主要用于提取不同层次的特征信息；当 stride=1 时，捷径分支不做任何处理，当 stride≠1 时通过一个 1×1 卷积进行下采样。捷径分支主要用于优化梯度回传和使残差函数训练更加容易，也可以通过融合原始输入特征，使网络结合多尺度信息。本文提出在组卷积中引入膨胀卷积，减少复杂度的同时扩大特征图的感受野，并在这个带有组卷积和膨胀卷积的残差结构中加入一个 SE 注意力机制，通过一个权重矩阵，从通道域的角度赋予图像不同位置的不同权重值，从而得到更重要的特征信息，并命名为 dblock(图 6)。然后将 dblock 进行模块间级联，形成多尺度膨胀金字塔模块(图 7)。该模块采用不同的膨胀率以获得多尺度的特征信息，其中组卷积系数 $g=16$，膨胀系数 $d$ 分别为 1,2,4,8。

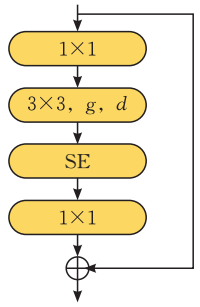


图 6　dblock 结构图

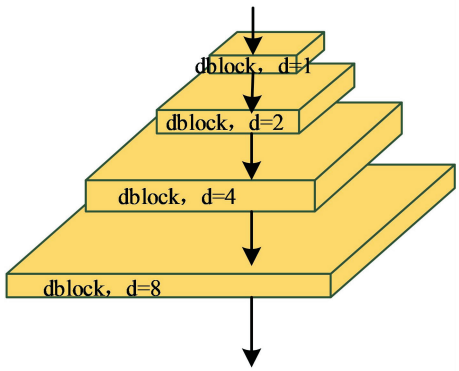


图 7　多尺度膨胀金字塔模块

### 1.4　损失函数

Dice Loss 是一个区域相关的损失值，即不管图片有多大，当正样本的区域大小一定时，其计算的损失值是不变的。该训练更倾向于挖掘前景区域，而正负样本不平衡的情况就是前景区域所占比例较小，因此采用 Dice Loss 不存在正负样本不平衡问题，但它存在梯度饱和的问题。

Ce Loss 用于分割时，当前景像素的数量远远小于背景像素的数量时，模型严重偏向背景，导致模型训练预测效果不好，因此单独使用 Ce Loss 无法处理正负样本不平衡的问题。Focal Loss 需要手动调整参数，会导致一些不稳定性；Dice Loss 对于小目标训练损失容易不稳定，另外极端情况也会导致梯度饱和现象，而 Ce Loss 中因不存在梯度饱和问题，梯度计算相对简单，这使得训练过程更加高效，减少了计算开销，因此为解决正负样本不平衡同时

避免梯度饱和问题，本文将 Ce Loss 和 Dice Loss 结合使用，命名为 C-D Loss。实验验证结合后的损失函数比单独使用其一更加有效。

## 2 实验过程与分析

### 2.1 数据集与训练策略

本文使用公开的城市街景数据集 CamVid，CamVid 是剑桥驾驶标签视频数据集，其中划分训练集、验证集和测试集分别为 367 张、101 张以及 233 张，分辨率大小为 480×360，目标分割类别为 11 类[23]，分别是道路、交通标志、汽车、天空、人行道、电线杆、围墙、行人、建筑物、自行车、树木。为了加快模型训练收敛速度，此次实验中运用了 Imagenet 图像分类数据集的预训练模型。为了合理利用显存资源和训练时间，本文采用冻结与解冻方式进行训练。一共训练 200 个周期得到检测模型，训练和测试的损失曲线如图 8 所示。

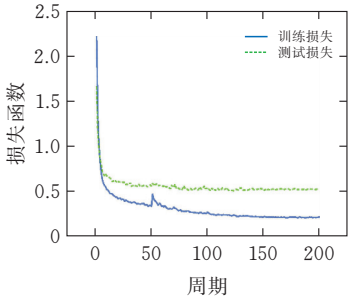


图 8 损失函数曲线

### 2.2 实验环境

本文的实验环境如下：深度学习框架为 Pytorch1.2，编程语言采用 Python3.7，实验平台采用 Ubuntu20.10 操作系统，并配置有 GeForce RTX 2080 ti 显卡(11 GB 内存)。此外网络模型训练相关参数如表 1 所示。

**表 1 训练参数设置**

| Parameters | Value / Name |
|---|---|
| Optimizer | ADAM |
| Freeze batch size | 8 |
| Unfreeze batch size | 4 |
| Epoch | 200 |
| Initial learning rate | 5e-4 |
| Momentum | 0.9 |
| Downsampling factor | 16 |

## 3 相关实验

### 3.1 消融实验

本文是以轻量级网络作为特征提取网络进行初步特征提取，因此选择多个轻量级网络作为特征提取网络进行消融实验，结果如表 2 所示。从实验结果可以发现，计算量最小的网络是 GhostNet，参数量最小的则是 MobileNetV3，而 mIoU 最高的也是 MobileNetV3，综合考虑，选取 MobileNetV3 作为本文的特征提取网络进行初步特征提取具有一定的优势。

**表 2 轻量级特征提取网络对比**

| 特征提取网络 | mIoU(%) | Params(M) | FLOPs(G) |
|---|---|---|---|
| GhostNet | 62.25 | 5.33 | 22.4 |
| GhostNetV2 | 62.73 | 6.31 | 22.5 |
| MobileNetV2 | 66.36 | 5.82 | 26.5 |
| MobileNetV3 | 66.72 | 4.84 | 31.6 |

为了验证 SA 对特征金字塔网络的影响，本文进行了相关的消融实验，实验结果如表 3 所示。从实验结果不难看出，如果只在 IFPN 的浅层特征或者深层特征处加入 SA，mIoU 都无法达到最优值，且单独加在浅层特征处相比于单独加在深层特征处可以提高 0.53%的 mIoU，说明 SA 更能充分利用浅层的空间信息，但最终两种特征层的横向连接处都应用 SA，可以使 mIoU 达到 67.28%，且 mPA 达到最优的 76.23%。

**表 3 IFPN 中 SA 的不同位置的效果对比**

| SA(top) | SA(bottom) | mIoU(%) | mPA(%) |
|---|---|---|---|
| | | 65.52 | 74.76 |
| √ | | 65.59 | 74.23 |
| | √ | 66.12 | 74.63 |
| √ | √ | 67.28 | 76.23 |

本文对损失函数进行相关改进，因此也探讨了不同损失函数对本文网络分割性能的影响，结果如表 4 所示。

**表 4 不同损失函数消融实验**

| Ce Loss | Focal Loss | Dice Loss | mIoU(%) | mPA(%) |
|---|---|---|---|---|
| √ | | | 67.28 | 76.23 |
| | √ | | 66.67 | 76.28 |
| | | √ | 66.41 | 74.19 |
| √ | √ | | 65.90 | 75.14 |
| √ | | √ | 67.85 | 76.95 |
| | √ | √ | 66.98 | 74.41 |

单独使用 Ce Loss 效果优于单独使用 Focal Loss 及单独使用 Dice Loss，但由于 Ce Loss 无法解决正负样本不平衡问题，Focal Loss 存在不稳定性，Dice Loss 能解决样本不平衡问题，但也存在梯度饱和问题，因此单独使用某个损失函数总会存在一些状况，无法达到最优值。综合考虑结合其中两种损失函数做了对比实验，实验表明 Ce Loss＋Dice Loss 的组合能达到最佳性能，能有效解决样本不平衡问题，也能减少训练时间，并取得了 67.85%的 mIoU。

为了验证 DAGM 和 MDPM 对网络性能的影响,本文单独对这两个模块进行消融实验。实验数据如表 5 所示。仅仅加入 DAGM 关注局部信息和上下文信息时,mIoU 可以达到 68.67%,而仅仅使用 MDPM,利用不同的膨胀率获得不同层次,不同大小的感受野时,mIoU 最优可以达到 69.27%,可以发现,MDPM 对网络的助力更大,而由于 MDPM 也是一个提取多尺度特征的模块,因此也可以从侧面说明有效提取特征图的多尺度信息对分割结果十分重要。最后同时采用这两个模块可以达到 70.18%的 mIoU,比单独使用一个模块时效果更突出。

**表 5　DAGM 和 MDPM 对比实验**

| DAGM | MDPM | mIoU(%) | mPA(%) |
|---|---|---|---|
| √ | | 68.67 | 76.50 |
| | √ | 69.27 | 76.77 |
| √ | √ | 70.18 | 78.60 |

有效利用多尺度特征对分割性能的提升尤为关键,不同的感受野代表不同的尺度信息,因此本文也探讨了 MDPM 的膨胀率对模块性能的影响,结果如表 6 所示。本文一共做了 5 组膨胀率的相关实验,结果表明膨胀卷积系数选择 1,2,4,8 时效果最好,更加适应本文的分割网络。

**表 6　不同膨胀率的对比试**

| dilation rate | mIoU(%) | mPA(%) |
|---|---|---|
| 1,3,5,7 | 70.02 | 78.41 |
| 1,2,4,8 | 70.18 | 78.60 |
| 2,4,6,8 | 70.10 | 78.29 |
| 3,5,7,9 | 69.78 | 77.54 |
| 1,4,8,12 | 69.99 | 77.42 |

为了检验本文提出的模块对整体网络性能的影响,进一步对本文所提出的模块进行消融实验,不同模块的帧率与推理时间对比如表 7 所示。

**表 7　不同模块的 FPS 与推理时间对比**

| Method | FPS | 时间/ms |
|---|---|---|
| IFPN | 82.7 | 12.1 |
| IFPN+C-D Loss | 80.5 | 12.4 |
| IFPN+C-D Loss+DAGM | 79.2 | 12.6 |
| IFPN+C-D Loss+DAGM+MDPM | 64.1 | 15.6 |

由表 7 可以看出最终本文算法的推理速度可以达到 64.1FPS。mIoU 与参数量的实验结果如图 9 所示(其中 $L_{C-D}$ 为 C-D Loss),加入 IFPN,整个网络相比于 DeepLabV3+参数量下降了约 19.8%,mIoU 增长了0.56%;加入 C-D Loss 之后 mIoU 又上升了0.57%,加入 DAGM 之后,mIoU 再次提高了 0.82%;再加入 MDPM 之后,mIoU 达到了 70.18%,相比 DeepLabV3+的 mIoU 上升了 3.46%,同时参数量下降了 0.4 M。

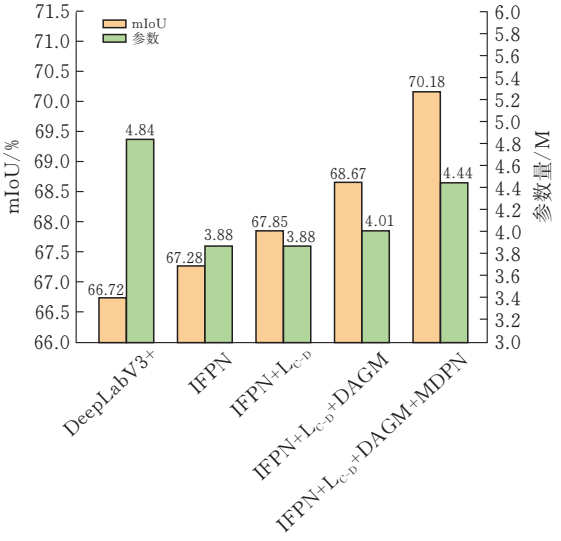


图 9　mIoU 与参数量对比

## 3.2　对比实验

为了验证本文算法的优势,将本文算法同当前主流语义分割算法进行对比实验,不同算法的分割效果对比如表 8 所示,本文主要以 MobileNetV3 作为特征提取网络进行一系列实验,本文算法的参数量只有 4.44M,比 $S^2$-FPN18 的参数量低了近 75%,而 mIoU 也高出 0.68%,因此本文的算法在分割精度提高的情况下,参数量大幅度下降,是其优势所在。

**表 8　不同算法在测试集上的对比**

| Methods | mIoU(%) | Params(M) | Flops(G) |
|---|---|---|---|
| DeepLab | 61.60 | 262.10 | 457.8 |
| PSPNet | 69.10 | 250.80 | 412.2 |
| U-Net | 69.29 | 43.90 | 92.2 |
| DeepLabV3+ | 66.72 | 4.84 | 31.6 |
| DFANet-A[24] | 64.70 | 7.80 | 3.4 |
| BiSeNet1 | 65.70 | 5.80 | 14.8 |
| BiSeNet2 | 68.70 | 49.00 | 54.0 |
| ICNet | 67.10 | 26.50 | 28.3 |
| DABNet | 66.40 | 0.76 | 10.4 |
| AGLNet[25] | 69.40 | 1.12 | 13.9 |
| LBN-AA[26] | 68.00 | 6.20 | 49.5 |
| $S^2$-FPN18 | 69.50 | 17.80 | 29.1 |
| Ours | 70.18 | 4.44 | 39.3 |

为了使本文算法的效果更加直观,将本文算法在 CamVid 数据集上的可视化效果展示如图 10 所示。可以发现,本文算法基本能将所有目标分割出来,对较小目标,如路灯和红绿灯也有良好的分割效果。同时本文算法的推理速度可以达到 64.1FPS,参数量为 4.44M,可以满足移动端或嵌入式设备的要求。

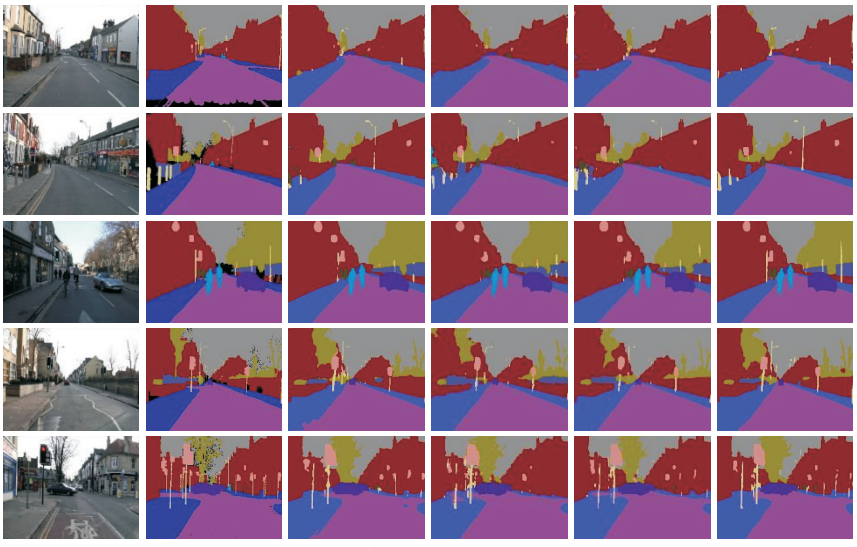


图 10 本文算法可视化

## 4 结论

本文提出基于轻量化及多尺度特征提取的城市街景语义分割算法，采用改进的特征金字塔网络，加强不同层之间的信息融合，深度不对称门控模块在减少计算量的同时融合局部和上下文信息，提出的 Ce Loss 和 Dice Loss 的结合损失函数也减少了样本不平衡问题带来的影响，而多尺度膨胀卷积模块则利用不同膨胀率的卷积组合，获得不同的感受野大小和不同尺度的信息，从一系列实验可以清楚发现，本文提出的这种算法可以实现精度、参数量和速度的平衡。

在未来的工作中可以考虑对数据集中小目标的有效特征提取进行研究，提高网络的特征提取能力。

[ 参 考 文 献 ]

[1] Feng D, Haase-Schütz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges [J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(03): 1341-1360.

[2] Asgari Taghanaki S, Abhishek K, Cohen J P, et al. Deep semantic segmentation of natural and medical images: a review [J]. Artificial Intelligence Review, 2021, 54: 137-178.

[3] Yuan X, Shi J, Gu L. A review of deep learning methods for semantic segmentation of remote sensing imagery[J]. Expert Systems with Applications, 2021, 169: 114417.

[4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[5] 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, 2017, 40(04): 834-848.

[6] Poudel R P, Liwicki S, Cipolla R. Fast-scnn: Fast semantic segmentation network[J]. arXiv preprint arXiv:1902.04502, 2019.

[7] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation [J]. arXiv preprint arXiv:1706.05587, 2017.

[8] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 6230-6239.

[9] Romera E, ÁLvarez J M, Bergasa L M, et al. ERFNet: Efficient residual factorized convNet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(01): 263-272.

[10] Mehta S, Rastegari M, Caspi A, et al. ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation [M] // Ferrari V, Hebert M, Sminchisescu C, et al. Computer Vision-ECCV 2018: Vol. 11214. Cham: Springer International Publishing, 2018:561-580.

[11] Zhao H, Qi X, Shen X, et al. Icnet for real-time semantic segmentation on high-resolution images[C]// Proceedings of the European conference on computer vision (ECCV). 2018: 405-420.

[12] Wang Y, Zhou Q, Liu J, et al. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation[C]// 2019 IEEE international conference on image processing (ICIP). IEEE, 2019: 1860-1864.

[13] Li G, Yun I, Kim J, et al. DABNet: Depth-wise

asymmetric bottleneck for real-time semantic segmentation[J]. Computer Science, 2019. DOI: 10.48550/arXiv.1907.11357.

[14] Elhassan M A, Yang C, Huang C, et al. S^2-FPN: Scale-ware strip attention guided feature pyramid network for real-time semantic segmentation[J]. arXiv e-prints, 2022: arXiv-2206.

[15] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation [C]// Proceedings of the European conference on computer vision (ECCV). 2018: 325-341.

[16] Yu C, Gao C, Wang J, et.al. BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129(11): 3051-3068.

[17] Fan M, Lai S, Huang J, et al. Rethinking bisenet for real-time semantic segmentation[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9716-9725.

[18] Hong Y, Pan H, Sun W, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes[J]. IEEE Transactions on Intelligent Transportation Systems,2023(24): 3448-3460.

[19] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]// 2016 fourth international conference on 3D vision (3DV). Ieee, 2016: 565-571.

[20] Zhang, Q L, Yang Y B. SA-Net: Shuffle attention for deep convolutional neural networks[Z]. 2021.

[21] Yang Z, Zhu L, Wu Y, et.al. Gated channel transformation for visual recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 11791-11800.

[22] Radosavovic I, Kosaraju R P, Girshick R, et al. Designing network design spaces[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10428-10436.

[23] 赵迪，孙鹏，陈奕博，等. 基于高度有效驱动注意力与多层级特征融合的城市街景语义分割[J]. 光电子·激光，2022，33(10)：1038-1046.

[24] Li H, Xiong P, Fan H, et al. Dfanet: Deep feature aggregation for real-time semantic segmentation[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9522-9531.

[25] Zhou Q, Wang Y, Fan Y, et al. AGLNet: Towards real-time semantic segmentation of self-driving images via attention-guided lightweight network[J]. applied soft computing, 2020, 96: 106682.

[26] Dong G, Yan Y, Shen C, et al. Real-time high-performance semantic image segmentation of urban street scenes[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(06): 3258-3274.

# A Semantic Segmentation Algorithm for Urban Streetscape Based on Lightweight and Multi-scale Feature Extraction

DING Jiang[1], CUI Hao[1], DAI Junwei[1], LI Lirong[1,2], GONG Pengcheng[3], SONG Haina[1,2]

(1 *School of Electrical and Electronic Engineering, Hubei Univ. of Tech., Wuhan* 430068, *China*;
2 *Hubei Power Grid Intelligent Control and Equipment Engineering Tech. Research Center, Hubei Univ. of Tech., Wuhan* 430068, *China*;
3 *School of Computer Science and Engineering, Wuhan Engineering Univ., Wuhan* 430205, *China*)

**Abstract**: It is currently a major trend to deploy high-performance neural network algorithms on platforms such as mobile or embedded devices. Aiming at the problem of high complexity of conventional segmentation models, this paper proposes a semantic segmentation algorithm for urban streetscape based on lightweight and multi-scale feature extraction. The algorithm adopts the lightweight network MobileNetV3 as the feature extraction network, introduces the feature pyramid network to embed the shuffle attention in its lateral connection to strengthen the information fusion, then stacks three times the depth-wise asymmetric gating module after its deep features to interact with the local and contextual information; and then employs the multi-scale dilation pyramid module to expand the receptive field to obtain the multi-scale feature information, and finally splices the different levels of information. Finally, the information at different levels is spliced and fused; in the training stage, this paper combines Ce Loss and Dice Loss to improve the training efficiency, solve the positive and negative sample imbalance problem, and avoid gradient saturation. Finally, full validation experiments are carried out on CamVid dataset, and the results show that the mIoU of this paper's algorithm reaches about 70.18%, while the number of model parameters is about 4.44M, and the inference speed is about 64.1FPS, which is better than the current mainstream algorithms.

**Keywords**: lightweight networks; urban streetscape; semantic segmentation; dilated convolution; multi-scale features

[责任编校：赵　瞳]