计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(3):19−29 [doi: 10.15888/j.cnki.csa.008376]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于稠密扩张卷积的图像语义分割模型①

张富财, 许建龙, 包晓安

(浙江理工大学 信息学院, 杭州 310018)
通信作者: 许建龙, E-mail: xujianlong126@126.com

**摘　要**: 为解决图像语义分割任务中面对的分割场景的复杂性、分割对象的多样性及分割对象空间位置的差异性问题, 提高语义分割模型的精度, 提出基于稠密扩张卷积的双分支多层级语义分割网络 (double branch and multi-stage network, DBMSNet). 首先采用主干网络提取输入图像的 4 个不同分辨率的特征图 (*De*1、*De*2、*De*3、*De*4), 其次采用特征精炼 (feature refine, FR) 模块对 *De*1 和 *De*3 这两个特征图进行特征精炼处理, 特征精炼处理之后的输出分支经过混合扩张卷积模块 (mixed dilation module, MDM) 编码空间位置特征, *De*4 分支采用金字塔池化模块 (pyramid pooling module, PPM) 编码高级语义特征, 最后将两个分支进行融合, 输出分割结果. 在数据集 CelebAMask-HQ 和 Cityscapes 中进行实验, 分别得到 *mIoU* 精度为 74.64%、78.29%. 结果表明, 本文方法的分割精度高于对比方法, 且具有更少的参数量.

**关键词**: 深度学习; 图像语义分割; 扩张卷积; 稠密连接; 多层级特征

引用格式: 张富财,许建龙,包晓安.基于稠密扩张卷积的图像语义分割模型.计算机系统应用,2022,31(3):19–29. http://www.c-s-a.org.cn/1003-3254/8376.html

## Image Semantic Segmentation Model Based on Dense Dilation Convolution

ZHANG Fu-Cai, XU Jian-Long, BAO Xiao-An

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

**Abstract**: Semantic segmentation is a very challenging task because of the complexity of parsing the scene, the diversity of segmented objects, and the differences in spatial positions of objects. To tackle this dilemma, this study proposes a novel architecture named double branch and multi-stage network (DBMSNet) based on dense dilation convolution. Firstly, four feature maps (*De*1, *De*2, *De*3, and *De*4) with different resolutions are extracted by the backbone network, and then the feature refinement maps of *De*1 and *De*3 are output through the feature refinement (FR) module. Secondly, the output branch is processed by the mixed dilation module (MDM) to extract rich spatial location features, while the *De*4 branch is processed by the pyramid pooling module (PPM) to extract multi-scale semantic information. Finally, the two branches are merged and the segmentation result is output. Comprehensive experiments are conducted on two public datasets of CelebAMask-HQ and Cityscapes, on which our model achieves mean intersection-over-union (*mIoU*) scores of 74.64% and 78.29%, respectively. The results show that the segmentation accuracy of this study is higher than that of the counterpart method, and this method has fewer parameters.

**Key words**: deep learning; image semantic segmentation; dilation convolution; dense connection; multi-stage feature

图像语义分割是为图像中的每一个像素分配一个具体的类别标签, 达到像素级别的分类, 是计算机视觉中的一项基础性工作. 图像语义分割具有广泛的实际应用场景, 如自动驾驶、城市遥感地图测绘、医学影像

① 基金项目: 浙江省重点研发计划 (2020C03094)
收稿时间: 2021-05-23; 修改时间: 2021-06-21; 采用时间: 2021-06-30; csa 在线出版时间: 2022-01-24

分析等. 在这些实际应用中, 高精度的分割结果至关重要.

随着卷积神经网络的发展, 基于深度学习的图像语义分割模型的精度得到空前的提高. Shelhamer 等提出的 FCN 模型[1], 奠定使用深度学习处理图像语义分割任务的一般性过程, 即先使用主干网络对图像做特征编码降低分辨率, 然后使用特定的解码器解码, 还原图像分辨率, 最终产生密集型的像素类别预测结果. 基于 FCN 编解码架构, 图像语义分割模型得到广泛的发展. 编码器主要使用高精度的图像分类网络, 如 VGG-Net[2]、GoogleNet[3]、ResNet[4]、DenseNet[5]、PeleeNet[6] 等, 这些主干网络具有很高的图像分类精度, 将最后的全连接分类层更换为卷积层便可以直接迁移到语义分割模型中使用. 在解码器方面, 主要关注特征图的语义信息和分割对象的空间位置信息, 出现许多处理全局特征的技术, 本质都是扩大感受野, 使模型感知全局信息, 如 Yu 等的 DilationConv[7] 提出扩张卷积的概念, 在不降低图像空间分辨率的基础上聚合图像中不同尺寸的上下文信息并且扩大感受野的范围, 精确定位分割对象; Chen 等的 DeepLab[8] 提出扩张空间卷积金字塔池化 (atrous spatial pyramid pooling, ASPP) 模块, 采用多个不同扩张率卷积的平行架构, 关注不同感受野下的对象分割; Zhao 等的 PSPNet[9] 提出金字塔池化模块 (pyramid pooling module, PPM), 使用平行的自适应池化操作获取不同感受野的分割对象.

除了通过扩大模型的感受野提高模型性能, 视觉注意力机制同样被引入图像分割任务中, 如 Fu 等提出 DANet[10], 同时使用位置注意力和通道注意力提高解码器的分割性能, 使模型有重点的关注分割对象, 但是注意力机制会耗费相当大的算力. 后来, 为了兼顾模型分割精度与推理速度, 许多模型使用轻量级的编码器和简易的解码器构建模型, 如 Paszke 等提出 ENet[11] 和 Zhao 等提出 ICNet[12], 虽然推理速度达到了实时要求, 但是精度还有待提高. 本文对前述技术进行综合考虑, 认为分割精度是语义分割模型首要考虑的因素. 本文对上述模型进行复现实验, 发现 PSPNet、DeepLab 等模型仅使用主干网络提取到的语义信息最丰富的最后一层特征图, 因为浅层的高分辨特征图依然富含大量的空间位置特征和语义特征[13], 所以通过合理的结合多级特征图依然可以提升模型性能. 因为上述模型没有充分利用浅层低级特征图的空间位置特征, 导致他们的模型虽然可以捕获丰富的高级语义特征, 但是缺乏分割对象的空间位置信息, 鉴于此, 本文提出基于编解码结构的高精度图像语义分割模型.

本文使用已有工作的主干网络 ResNet[4] 提取图像特征, 获得 4 级不同分辨率的特征图 (*De*1, *De*2, *De*3 和 *De*4), 在此基础上提出编解码架构的高精度语义分割模型: 双分支多层级语义分割网络 (double branch multi-stage network, DBMSNet), 使用双分支同时处理分割对象的空间位置特征和高级语义特征.

本文主要工作为: (1) 提出特征精炼模块 (feature refine, FR), 通过融合高级语义特征与浅层的空间位置特征, 使模型捕获丰富的空间位置信息及全局上下文信息, 强化模型的空间位置感知能力; (2) 提出混合扩张卷积模块 (mixed dilation module, MDM), 使用已有的深度可分离卷积[14] 搭建稠密型连接模块 (轻量级模块), 充分混合不同扩张率的扩张卷积, 获取不同尺度的感受野, 强化模型对不同尺度对象的感知能力, 增强空间位置特征的解码能力, 使模型获取全局上下文信息; (3) 提出双分支的解码器, 第 1 个分支使用 FR 和 MDM 解码浅层特征的空间位置特征, 第 2 个分支使用已有的 PPM 模块解码高级语义特征; (4) 与对比方法相比, 本文提出的双分支多层级语义分割网络在公开数据集上达到更高的精度.

## 1 双分支多层级语义分割网络

### 1.1 模型总览

在这个小节, 介绍双分支多层级语义分割网络的整体结构. DBMSNet 由特征精炼模块、混合扩张卷积模块及金字塔池化模块构成. 目的是充分利用多层次特征, 通过学习丰富的空间位置特征和全局上下文特征完成高精度的图像分割任务. 如图 1 所示.

输入 3 通道图像经过 Backbone 产生 4 级特征图, 分别为不同的分辨率 (*De*1、*De*2、*De*3、*De*4 的分辨率为输入图像的 1/4、1/8、1/16、1/16). 然后使用双分支进行处理, 第 1 个分支为空间位置特征处理分支, 首先将 *De*1 和 *De*3 经过 FR 模块, 产生与 *De*1 分辨率一致的特征图, 然后经过 MDM 模块处理产生第 1 个分支的输出. 第 2 个分支为语义特征处理分支, 使用 PPM 模块处理 *De*4 特征图产生第 2 个分支的输出. 最终, 将两个分支产生的输出进行加权合并操作, 使用双线性插值算法将合并的输出上采样至输入图片的原始分辨率作为最终输出, 完成端到端的模型搭建, 既捕获抽象的高级语义特征, 又级联浅层的空间位置特征.

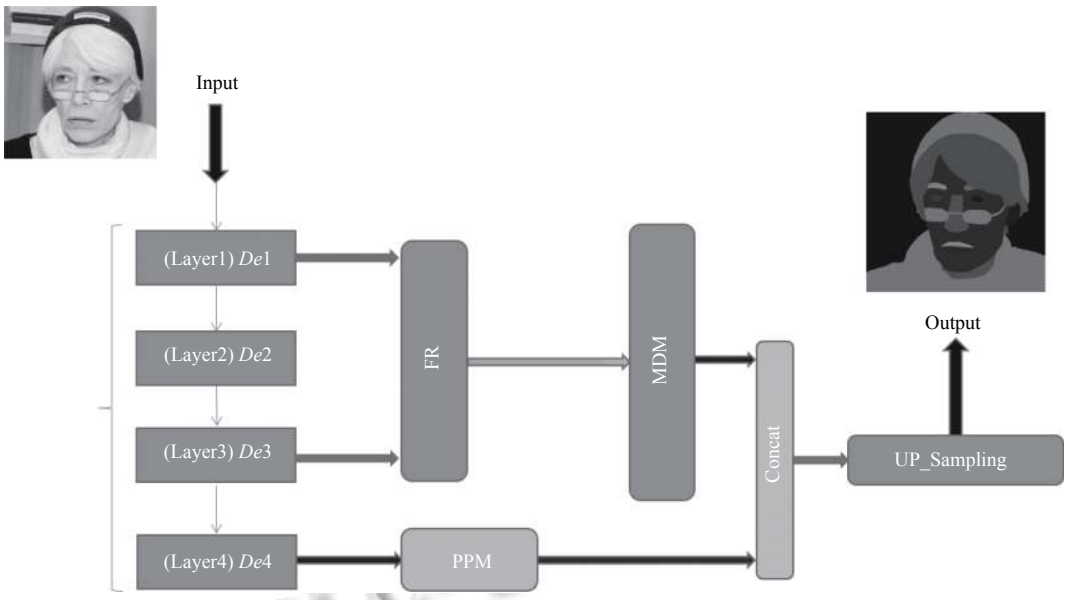



图 1 DBMSNet 网络的整体结构

### 1.2 特征精炼模块

当输入图像的分辨率为 512×512×3 时, *OS*=16 时 (*OS* 表示 output stride), 主干网络提取到的各级特征图分辨率如表 1 所示. 特征精炼模块细节如图 2.

表 1 主干网络的各级特征图分辨率

| 级别 | 分辨率 |
|---|---|
| Input | 512×512×3 |
| *De*1 | 128×128×256 |
| *De*2 | 64×64×512 |
| *De*3 | 32×32×1024 |
| *De*4 | 32×32×2048 |

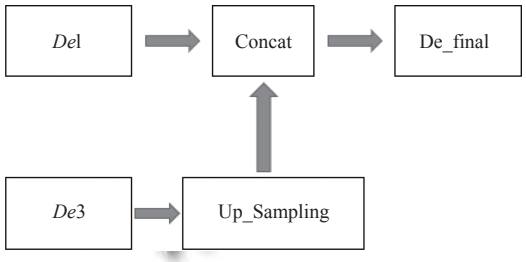



图 2 特征精炼模块细节

特征精炼模块如式 (1):

$$De_{\text{final}} = C(De1, Up(De3)) \tag{1}$$

其中, $Up(\cdot)$为双线性插值上采样函数, $C(\cdot,\cdot)$为特征图通道级联函数.

### 1.3 混合扩张卷积

#### 1.3.1 稠密型扩张卷积

稠密型扩张卷积 (dense dilation convolution, DDC) 是混合扩张卷积模块的基本组成部分, 并且深度可分离卷积[14] 和分组卷积[15] 与标准卷积的性能相似, 但是效率更高. 因此, 使用深度可分离卷积来构建轻量级 DDC 模块, 如图 3 所示的 DDC 模块细节图.

首先, 给定一个输入图片$I^{H\times W\times C_0}$, $H$为高, $W$为宽, $C_0$为通道数. 使用通道降维 (channel reduce, CR) 层进行通道降维, 该层使用 1×1 分组卷积使通道数降低为 $C_0\times\alpha$, $\alpha$为通道降低率, 得到特征图$I'^{H\times W\times(C_0\times\alpha)}$; 其次输入到 4 个平行的深度可分离卷积层, 其中卷积操作后都进行 BatchNorm 和 ReLU 操作, 以加速模型的收敛速度、提高稳定性及解决梯度消失问题, 分别生成 4 个使用不同扩张率卷积处理的特征图, 如式 (2) 所示.

$$L = \{l_i^{H\times W\times(C_0\times\alpha),r_i}\}, 1\leqslant i\leqslant 4 \tag{2}$$

其中, $i$ 为正整数, $r_i$为不同的扩张率, $l_i$为各分支的特征图. 使用不同扩张率卷积的平行架构可以捕获不同尺度感受野对象, 在多个尺度上合并上下文信息. 虽然上述 4 个平行分支可以捕获多尺度局部语义特征, 但是缺少全局感知信息. 为了克服这个缺点, 设计自适应平均池化分支捕获全局上下文信息, 通过$I'^{H\times W\times(C_0\times\alpha)}$产生$G^{1\times1\times(C_0\times\alpha)}$ ($G$为池化分支的输出). 全局平均池化是通过计算输入的高度 $H$ 和宽度 $W$ 的平均值来进行下采样操作, 然后, 同样使用 1×1 深度可分离卷积降低通道数, 接着使用双线性插值算法恢复分辨率, 以便与 4 个平行分支产生的特征图进行通道合并. 最下面一个数据流表示残差连接.

另外, 直接将上述 5 个分支的特征图通道合并会

削弱特征表达, 所以设计通道随机打乱操作 (channel shuffle, CS), 使特征的泛化表达性更高. 整个 DDC 模块的操作如式 (3) 所示.

$$O^{H\times W\times C_0}=S(C(L_n^{H\times W\times(C_0\times\alpha)},G'^{H\times W\times(C_0\times\alpha)}))\oplus I^{H\times W\times C_0} \tag{3}$$

其中, $S(\cdot)$为通道随机混合函数, $C(\cdot,\cdot)$为通道合并函数, $L$为 4 个不同扩张率卷积分支产生的特征图, $G'$为自适应平均池化分支上采样产生的特征图, $\oplus$为元素级相加, $H$为特征图的高, $W$为特征图的宽, $C_0$为特征图通道数, $\alpha$为通道缩减率, $n$为特征图数量 1, 2, 3, 4.

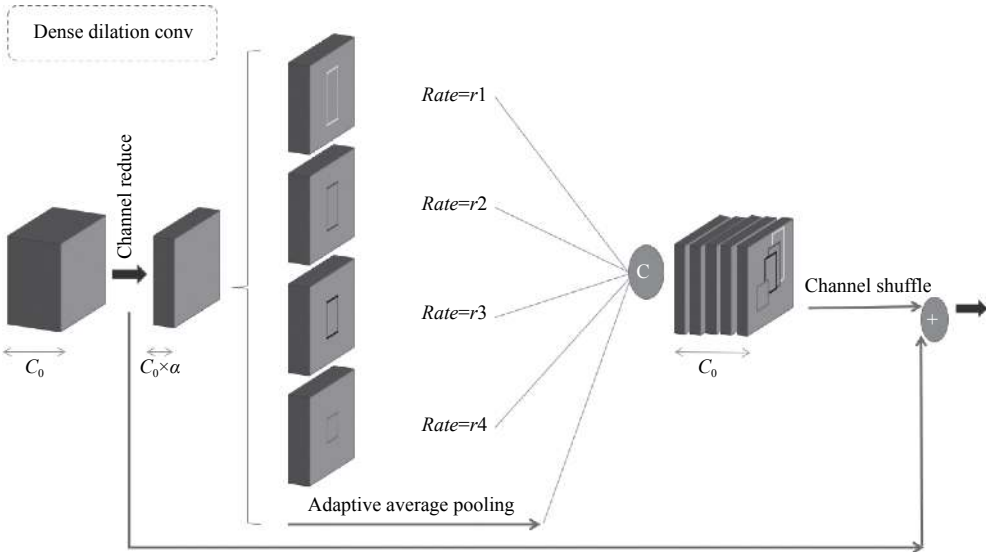



图 3 稠密扩张卷积模块的细节图

DDC 模块的详细参数如表 2 所示, 设输入为 $H\times W\times C_0$, 通道缩减率为$\alpha$. 其中, $k$为卷积核的大小, $r_i\ (i=1,2,3,4)$为不同的扩张率, $BN$ 为 BatchNorm, $H$, $W$, $C_0$为特征图的高、宽和通道数.

表 2 稠密扩张卷积模块参数

| 操作 | 参数 | 输出 |
|---|---|---|
| Channel_Reduce | $k=3, groups=4, BN$ | $(H, W, C_0\times\alpha)$ |
| Depthwise_Branch1 | $k=3, r_1, BN$, ReLU | $(H, W, C_0\times\alpha)$ |
| Depthwise_Branch2 | $k=3, r_2, BN$, ReLU | $(H, W, C_0\times\alpha)$ |
| Depthwise_Branch3 | $k=3, r_3, BN$, ReLU | $(H, W, C_0\times\alpha)$ |
| Depthwise_Branch4 | $k=3, r_4, BN$, ReLU | $(H, W, C_0\times\alpha)$ |
| Ave_Pooling | $k=1, BN$, ReLU | $(1, 1, C_0\times\alpha)$ |
| Concat | $l_{ave\_pooling}, l_i, i=1,2,3,4$ | $(H, W, C_0)$ |

### 1.3.2 混合扩张卷积模块

混合扩张卷积模块由上述 DDC 模块组成, 细节展示如图 4 所示. MDM 模块分为两部分, 分别为 DDC 组成的稠密连接分支以及全局池化分支, 用来捕获多尺度对象语义特征以及空间位置特征. 稠密连接分支使用 3 个 DDC 模块的堆叠方式而不是平行架构, 即外部为稠密连接方式, 内部为多级平行架构. 这样充分利用不同尺度的感受野, 通过对特征图的最大化利用从而达到最好的效果且有更少的参数.

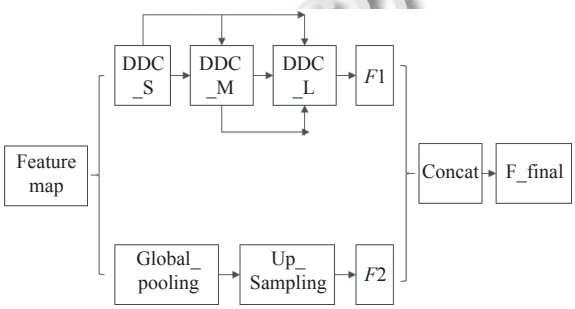



图 4 混合扩张卷积模块细节图

每个 DDC 模块拥有明确的对象捕获尺度, 第 1 个 DDC 模块使用的扩张率组合为$D_s=\{1,3,5,7\}$, 主要捕获小尺度对象; 以$D_s$的输出作为输入, 第 2 个 DDC 模块的扩张率组合为$D_m=\{5,7,11,13\}$, 主要捕获中等尺寸对象; 最后一个 DDC 模块的扩张率组合为$D_l=\{13,15,17,19\}$, 主要捕获大尺度对象. 因为在输入的时候进行通道降维操作, 并且特征图的分辨率较小, 所以设置较大的扩张率不会增加太多的计算开销. 最后, 将每一个 DDC 模块产生的输出进行元素级别的求和操

作, 共同编码多层次语义. MDM 模块的输出如式 (4) 所示.

$$O = C(F1, F2) \tag{4}$$

其中, $C(\cdot,\cdot)$为特征图通道合并操作, $F1$ 为稠密型连接分支的输出, $F2$ 为全局平均池化上采样后的输出. 感受野的计算如式 (5) 所示, 其中, $A$ 为感受野尺寸, $K$ 为卷积核尺寸, $D$ 为扩张率.

$$A = (D-1)\times(K-1)+K \tag{5}$$

最终, 堆叠三级 DDC 模块可以在理论上获得最大的感受野, 如式 (6) 所示.

$$A_{\max} = A^{s}_{\max} + A^{m}_{\max} + A^{l}_{\max} - 2 \tag{6}$$

其中, $A^{\beta}_{\max}(\beta = s,m,l)$分别为小感受野 DDC 模块, 中感受野 DDC 模块, 大感受野 DDC 模块. MDM 模块的参数配置如表 3 所示, 其中, $k$ 表示卷积核的大小, $BN$ 表示 BatchNorm, *Dilation* 表示扩张率, $H$、$W$、$C$ 分别表示特征图的高、宽和通道数.

表 3 混合扩张卷积模块参数

| 操作 | 参数 | 输出 |
|---|---|---|
| Channel_Reduce | $k$=1, $BN$, ReLU | ($H$, $W$, $C$) |
| DDC_S | *Dilation*=(1, 3, 5, 7) | ($H$, $W$, $C$) |
| DDC_M | *Dilation*=(5, 7, 11, 13) | ($H$, $W$, $C$) |
| DDC_L | *Dilation*=(13, 15, 17, 19) | ($H$, $W$, $C$) |

### 1.4 金字塔池化模块

借鉴 PSPNet 中的 PPM, 使用自适应平均池化操作, 处理第 2 个分支, 自适应即为将特征图分别池化为 (1, 1)、(2, 2)、(3, 3)、(6, 6) 的尺寸, 增强局部与全局特征的表达能力, 最后与输入图进行通道合并, 融合全局先验知识, 如图 5 所示.

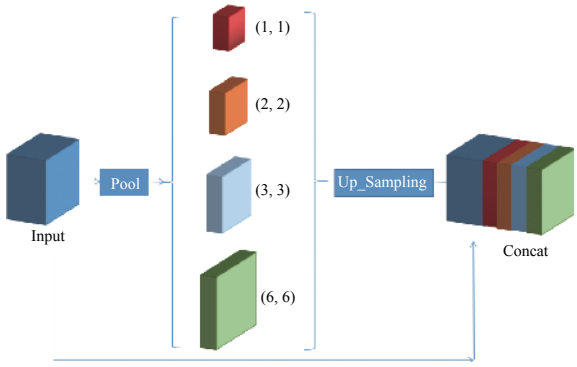



图 5 金字塔模块的细节图

模型的所有配置参数如表 4 所示, 以输入为 (512, 512, 3) 的图像为例. 表中: conv 为卷积操作, $[k\times k, conv, Number]$表示卷积核大小为 $k$, 卷积核的数量为 *Number*, $[k\times k, dwconv, r=\{1,2,3,5\}]$表示深度可分离卷积操作, 卷积核尺寸为 $k$, $r$ 为扩张率, *global_pooling* 为全局池化操作, *num_classes* 为分割的类别数.

表 4 DBMSNet 参数

| 方法 | Stages | Layer | | Output size |
|---|---|---|---|---|
| Input | — | — | | 512×512×3 |
| Backbone (ResNet50) | Stage 1 (*De*1) | BottleneckBlock × 3 | $[1\times1, conv, 256]$ | 128×128×256 |
| | Stage 2 (*De*2) | BottleneckBlock × 4 | $[1\times1, conv, 512]$ | 64×64×512 |
| | Stage 3 (*De*3) | BottleneckBlock × 6 | $[1\times1, conv, 1\,024]$ | 32×32×1024 |
| | Stage 4 (*De*4) | BottleneckBlock × 3 | $[1\times1, conv, 2\,048]$ | 32×32×2048 |
| Branch_1 | FeatureRefine | Bilinear_Interpolate | — | 32×32×1280 |
| | ChannelSpecific | SeparableConv | $[1\times1, conv, 320]$ | 32×32×320 |
| | MixedDilation | DenseDilationConv | $[3\times3, dwconv, r=\{1, 2, 3, 5\}]$ | 32×32×400 |
| | | DenseDilationConv | $[3\times3, dwconv, r=\{5, 7, 9, 11\}]$ | |
| | | DenseDilationConv | $[3\times3, dwconv, r=\{11, 13, 15, 17\}]$ | |
| | | GlobalPooling | $[global\_pooling, 1\times1, conv, 80]$ | |
| Branch_2 | PyramidPooling | AdaptivePooling | $[size=(1,2,3,6), 1\times1, conv]$ | 32×32×1024 |
| Final | Classifier | Conv | $[1\times1, conv, num\_classes]$ | 512×512×$C$ |

## 2 实验细节

DBMSNet 模型的优势在于对分割对象空间位置特征的捕获和高级语义特征的融合, 为了验证提出的 DBMSNet 的有效性, 在公开数据集中进行广泛的实验.

### 2.1 数据集

(1) Cityscapes 城市场景数据集[16]. 这是一个流行的用于城市场景对象解析的数据集, 它包含 25 000 张分辨率为 2048×1024 的标注图片. 其中精细化标注的图片数量为 5 000 张并且包含 19 个类别, 如行人、汽车、天空、建筑物等. 实验时使用 2 975 张图片作为训练集, 500 张图片作为验证集, 1 525 张图片作为测试集. 由于硬件设备的限制, 训练时将图片裁剪为 1024×512 的尺寸.

(2) CelebAMask-HQ 人脸分割数据集[17]. 该数据集包含 30 000 张高分辨率的人脸图片, 每一张图片拥有详细的标注信息.该数据集的标注图的尺寸为 512×512 并且分为 19 个类别, 分别为面部皮肤、鼻子、眼镜、左右眼、左右眉毛、左右耳朵、牙齿、上下嘴唇、头发、帽子、耳环、项链、脖子和衣服. 提取了一部分数据进行 DBMSNet 的实验, 其中训练集 24 184 张, 验证集 300 张, 测试集 2 824 张. 为了验证模型的泛化能力, 继续在城市场景解析数据集中进行了广泛的实验.

### 2.2 评价指标

(1) 语义精度. *mIoU* 平均交并比是一个广泛用于图像分割领域的评价分割精度的指标. 假设 $b$ 代表语义分割的类别数, 则 *mIoU* 如式 (7) 所示.

$$mIoU = \frac{1}{b+1}\sum_{z=0}^{b}\frac{p_{zz}}{\sum_{v=0}^{k}p_{zv}+\sum_{v=0}^{b}p_{vz}-p_{zz}} \tag{7}$$

其中, $p_{vz}$为真实值为$z$, 被预测为$v$的数量, $b+1$为类别个数 (包含背景忽略类). $p_{zz}$为正确的数量. $p_{zv}$、$p_{vz}$分别为假正和假负.

(2) 像素精度 (pixel accuracy, *PA*). 预测正确的像素数占总像素数的比例, 如式 (8) 所示.

$$PA = \frac{\sum_{z=0}^{b}p_{zz}}{\sum_{z=0}^{b}\sum_{v=0}^{b}p_{zv}} \tag{8}$$

其中, $p_{zv}$为像素总数, $p_{zz}$为预测正确的像素数, $b$为类别数.

### 2.3 消融实验

实验中使用开源深度学习框架 PaddlePaddle[18] 搭建 DBMSNet 模型, 实验设备为 Tesla V100 单卡 32 GB 显存, 操作系统为 Ubuntu 16.04. 在训练之前, 进行数据预处理操作: ① 使用步长为 0.25, 范围为 0.75 到 1.5 的随机尺寸缩放; ② 设置随机水平翻转与随机垂直翻转概率为 0.5; ③ 设置随机旋转角度为 (−10, 10); ④ 设置随机对比度变化范围 0.4, 随机亮度变化范围 0.4, 随机色彩饱和度变化范围 0.4; ⑤ 像素归一化处理.

在训练时使用“Poly”学习率衰减策略, *power* 为 0.9, 终止学习率为 0, 如式 (9) 所示.

$$lr = 1-\left(\frac{iter}{max\_iter}\right)^{power} \tag{9}$$

其中, *lr* 为学习率, *iter* 为迭代次数, *max_iter* 为最大迭代次数, 使用随机梯度下降 (stochastic gradient descent, SGD) 优化算法, 动量为 0.9, 权重衰减为 $4e^{-5}$. 对于两个分支的输出, 第 1 个分支的权重为 0.4, 第 2 个分支的权重为 1. 最后使用像素级的交叉熵损失为损失函数. 本文在 Cityscapes 数据集中进行广泛的消融实验, 设置 *Batch_Size*=6, *iters*=40000.

#### 2.3.1 双分支的有效性消融实验

为验证本文提出的双分支模型的有效性, 首先构造只使用 *De*4 特征图, 并使用 PPM 解码的单分支模型作为 Baseline. 实验结果如表 5 所示.

表 5 双分支有效性实验

| Model | *mIoU* (%) | Acc (%) | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| *De*4(PPM) | 75.27 | 94.90 | 65.6 | 65.0 |
| *De*4(PPM)+*De*3(Two) | 75.05 | 95.26 | 66.5 | 65.8 |
| *De*4(PPM)+*De*3\2(Two) | 75.04 | 95.01 | 66.7 | 68.7 |
| ***De*4(PPM)+*De*3\1(Two)** | **75.78** | **95.15** | **66.6** | **78.8** |
| *De*4(PPM)+*De*3\2\1(Two) | 75.56 | 95.07 | 66.8 | 81.2 |

表 5 中, *De*4(PPM) 表示只使用 De4 特征图, 用 PPM 处理; *De*3(Two) 表示 *De*3 使用 TwoDecoder 处理, TwoDecoder 表示本文提出的 FR 和 MDM 组合的解码器; *De*3\2 表示 FR 模块的输入为 *De*3 和 *De*2; Params 表示模型的参数量; FLOPs 表示模型的浮点运算总量 (输入为 360×640 估算). 例如 *De*4(PPM)+*De*3\2\1(Two) 表示: *De*4 分支使用 PPM 处理, *De*3\2\1 为第 1 个分支的输入, 使用 TwoDecoder 处理.

实验结果表明, 在单独使用 *De*4(PPM) 分支的 Baseline 的基础上, 添加第 1 个分支会对模型性能产生影响. 其中, 添加 *De*3 或 *De*3\2 为第 1 个分支的输入时会对 Baseline 产生负面影响, 精度分别降低 0.22 和 0.23; 添加 *De*3\1 或 *De*3\2\1 为第 1 个分支的输入时会对 Baseline 产生正面影响, 精度分别提高 0.51 和 0.29; 并且, 添加 *De*3\1 作为第 1 个分支的输入取得最好性

能 75.78, 相比 Baseline 精度提升 0.51, 证明本文提出的双分支解码器是有效的.

2.3.2 PP_Out 参数的消融实验

为相对减少模型参数, 分别设置 PP_Out=1024、512、256 和 128 进行对比实验, PP_Out 表示 PPM 模块的输出通道数, 选择的基准模型为实验 1 中的最佳配置: *De*4(PPM)+*De*3\1(Two), 结果如表 6 所示.

表 6 PP_Out 参数影响

| PP_Out | *mIoU* (%) | Acc (%) | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| 1 024 | 75.05 | 95.26 | 66.5 | 65.8 |
| 512 | 75.42 | 95.34 | 47.6 | 59.7 |
| 256 | 75.24 | 95.11 | 38.1 | 51.0 |
| **128** | **75.76** | **95.39** | **33.0** | **35.0** |

实验结果表明, 在基准模型一致的前提下, 设置不同的 PP_Out 会对模型产生不同的影响. 设置 PP_Out=128 的参数量为 PP_Out=1024 的 50%, FLOPs 为 PP_Out=1024 的 61%, 证明 PP_Out 设置为 128 可以取得更好的结果, 且拥有更少的 Params 和 FLOPs.

2.3.3 TwoDecoder 的消融实验

TwoDecoder 表示第 1 个分支的解码器 (由 FR 和 MDM 组成). 在实验 1 证明 *De*4(PPM) 基础上添加 *De*3\1(Two) 取得最佳性能后, 充分调试 TwoDecoder 的最佳配置, 实验中设置 PP_Out=128, 结果如表 7 所示.

表 7 中, 设置第 1 个分支为 *De*3\1, 依次验证 TwoDecoder 的配置, Small 表示 MDM 模块中 DDC 小扩张率 *r*=[1, 3, 5, 7], Middle 表示 MDM 模块 DDC 中扩张率 *r*=[5, 7, 11, 13], Large 表示 MDM 模块中 DDC 大扩张率 *r*=[13, 16, 18, 20], Global_Pooling 表示 MDM 中的全局池化层, Attention 表示 MDM 中的注意力层, 在实验中增加注意力机制进行实验效果的探索; 第 2 个分支 *De*4(PPM) 保持不变. 如 *De*3\1+Small+Middle 表示: 第 2 个分支保持 *De*4(PPM) 不变, 第 1 个分支使用 *De*3\1 为输入, MDM 模块包含 Small 和 Middle 两个扩张率组合.

表 7 TwoDecoder 最佳配置

| 模型 | *mIoU* (%) | Acc (%) | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| *De*3\1+Small | 75.58 | 95.23 | 33.2 | 44.3 |
| *De*3\1+Small+Middle | 76.10 | 95.39 | 33.2 | 44.7 |
| *De*3\1+Small+Middle+Large | 76.10 | 95.33 | 33.3 | 45.0 |
| ***De*3\1+ Small+Middle+Large+Global_Pooling** | **76.37** | **95.42** | **33.4** | **46.9** |
| *De*3\1+ Small+Middle+Large+Global_ Pooling+Attention | 75.98 | 95.40 | 33.5 | 48.4 |

实验结果表明, 通过改变不同 TwoDecoder 配置, 在 *De*4(PPM) 和 *De*3\1 两个分支保持不变的前提下, 第 1 个分支 TwoDecoder 配置为 Small+Middle+Large+Global_Pooling 时精度最高为 76.37, 相反, 添加 Attention 层后的精度下降 0.39.

2.3.4 单分支的消融实验

在实验 1、2 得出 PP_Out=128, 第 1 个分支输入为 *De*3\1 时, 模型得到最高精度后, 继续验证单分支的性能, Two 的配置为实验 3 中的最佳配置, 实验结果如表 8 所示.

表 8 单分支消融性实验

| Branch | *mIoU* (%) | Acc (%) | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| *De*4(PPM) | 75.34 | 95.16 | 32.5 | 34.6 |
| *De*3\1(Two) | 74.69 | 95.63 | 24.5 | 42.5 |
| ***De*4(PPM)+*De*3\1(Two)** | **76.37** | **95.42** | **33.4** | **46.9** |

表 8 中, *De*4(PPM) 表示只使用 *De*4 分支, PPM 解码; *De*3\1(Two) 表示只使用第 1 个分支, *De*3\1 作为输入, 使用 TwoDecoder 解码; 设置 PP_Out=128. 实验结果表明, 单独使用 *De*4 分支比单独使用 *De*3\1 分支的精度高 0.65, 结合两个分支取得最佳精度 76.37, 均高于单独使用一个分支的精度, 证明两个分支结合的有效性.

2.3.5 *OS* 的消融实验

*OS* 为主干网络的输出步长, 表示提取特征图的缩放比例. 在基准模型为 *De*4(PPM)+*De*3\1(Two) 的前提下, 测试不同 *OS* 对模型精度的影响. PP_Out 设置为 128, 实验结果如表 9 所示.

实验结果表明, 在基准模型相同的情况下, *OS*=16 取得最佳精度 76.37%, 比 *OS*=8 高 0.35, 实验数据表明 *OS* 的改变不会影响模型的参数量, 但是会影响浮点运算总量, *OS*=16 的浮点运算总量仅为 *OS*=8 的 40%, 且精度更高, 所以设置模型 *OS*=16.

表 9 *OS* 消融性实验

| *OS* | *mIoU* (%) | Acc (%) | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| 8 | 76.02 | 94.36 | 33.4 | 118.4 |
| **16** | **76.37** | **95.42** | **33.4** | **46.9** |

2.3.6 主干网络消融实验

通过改变不同的 Backbone 网络, 验证本文模型可以得到的最佳精度. 结果如表 10 所示.

该实验使用上述实验得出的最佳模型配置: *De*4(PPM)+*De*3\1(Two), Two 为 Small+Middle+Large+Glo-bal_Pooling. 实验结果表明, 使用主干网络 ResNet101 取得最佳精度 78.29%, 比 ResNet50 高 1.32, 比 ResNet152 高 0.14.

表 10 主干网络的影响

| Backbone | *mIoU* (%) | Acc (%) | Params (M) | FLOPs (G) |
|---|---|---|---|---|
| ResNet50_vd | 76.97 | 95.61 | 33.4 | 46.9 |
| **ResNet101_vd** | **78.29** | **95.98** | **52.5** | **64.4** |
| ResNet152_vd | 78.15 | 95.91 | 68.1 | 81.8 |

## 2.4 对比实验结果

(1) 与当前主流的高精度图像语义分割模型对比性能, Cityscapes 实验结果如表 11 所示, 其他模型的数据均来自公开论文中的数据. 实验结果表明, 本文提出的模型精度均高于对比模型. Cityscapes 数据集可视化结果如图 6 所示.

表 11 其他模型性能对比数据

| 方法 | Backbone | *mIoU* (%) |
|---|---|---|
| SegNet[19] | — | 56.10 |
| GUNet[20] | — | 70.40 |
| FCN8s[1] | VGGNet16 | 65.30 |
| DeepLab[8] | ResNet101 | 70.40 |
| WASPNet[21] | ResNet101 | 70.50 |
| RefineNet[22] | ResNet101 | 73.60 |
| PGCNet[23] | ResNet18 | 75.78 |
| DenseASPP[24] | ResNet101 | 76.20 |
| MPDNet[25] | ResNet50 | 76.80 |
| DBMSNet (Ours) | ResNet50 | 76.97 |
| **DBMSNet (Ours)** | **ResNet101** | **78.29** |
| DBMSNet (Ours) | ResNet152 | 78.15 |

(a) 原图 (b) 真值图 (c) 预测图

图 6 Cityscapes 可视化结果

(2) 模型参数量对比. 为体现本文提出的模型的先进性, 继续对比模型的参数量与浮点运算总量, 对比结果如表 12 所示.

实验结果表明, 本文提出的模型取得最高的精度 78.29, 拥有最少的参数量 33.4 M 和浮点运算数 421.8 G, 充分证明本文提出模型的先进性.

(3) CelebAMask-HQ 实验结果.

CelebAMask-HQ 数据集中的实验结果如表 13 所示. 首先在自己的实验环境中复现了表中所有的对比模型, 并且使用相同的训练准则进行模型训练. 通过实

验数据可以清晰地观察到, DBMSNet 模型在捕获左眼和右眼、左眉毛和右眉毛、左耳和右耳的类别 IoU 方面取得了绝对的领先, 这归功于提出的特征精炼 (FR) 模块以及混合空洞卷积 (MDM) 模块, 不但可以提取多尺度的高级抽象语义信息, 而且可以完美地获取解析对象的空间位置信息, 在其他的类别预测精度相差无几的情况下, 精确的空间位置信息可以完美地反应人脸中对称的对象, 如左右眉毛、左右眼睛等, 这些对象具有绝对相似的外观, 但是空间位置不同, DBMSNet 模型完美地解决了这个难题. 与 PSPNet 相比精度提升了 0.76%, 像素精度提升了 0.02%, 达到了最高的分数.

可视化结果比较如图 7 所示. 通过可视化的比较结果, 可以清晰地看到, UNet、DeepLab v3+等模型由于特征表达能力不够, 会出现类别混淆、对象空间位置混淆、对象类别错分的现象, 并且对于大多数对象的空间位置预测出错. 而 PSPNet 虽然也达到了很高的分割精度, 但是在空间位置准确度预测方面依然不如本文的模型.

表 12 模型参数对比

| 方法 | Backbone | *mIoU* (%) | Params (M) | GFLOPs (1024×2048) |
|---|---|---|---|---|
| FCN-8s[1] | VGG-16 | 65.30 | 134.5 | 1 319.1 |
| Dilation10[7] | VGG-16 | 67.10 | 134.4 | 997.2 |
| DeepLab[8] | ResNet101 | 70.40 | 43.9 | 1 481.4 |
| DUC[26] | ResNet101 | 77.60 | 163.7 | 1 135.7 |
| RefineNet[22] | ResNet101 | 73.60 | 118.0 | 6 748.2 |
| DBMSNet (Ours) | ResNet50 | 76.97 | 33.4 | 421.8 |
| **DBMSNet (Ours)** | **ResNet101** | **78.29** | **52.5** | **577.4** |
| DBMSNet (Ours) | ResNet152 | 78.15 | 68.1 | 733.1 |

表 13 CelebAMask-HQ 实验结果 (%)

| 方法 | Backbone | 脸 | 鼻子 | 眼镜 | 右眼 | 左眼 | 右眉毛 | 左眉毛 | 右耳 | 左耳 | 牙齿 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PSPNet[9] | ResNet50 | **93.61** | 88.56 | **87.01** | 59.78 | 59.30 | 55.88 | 54.61 | 57.43 | 53.82 | **87.62** |
| DANet[10] | ResNet50 | 93.10 | 88.88 | 82.94 | 32.46 | 27.41 | 29.27 | 28.82 | 41.87 | 9.77 | 85.02 |
| FCN[1] | HRNet | 93.60 | **89.21** | 85.96 | 39.84 | 41.12 | 37.49 | 36.85 | 41.96 | 12.63 | 86.93 |
| BiseNet[27] | — | 92.49 | 88.40 | 80.08 | 34.77 | 25.94 | 33.68 | 24.64 | 42.55 | 4.66 | 84.18 |
| UNet[28] | — | 91.38 | 87.25 | 75.65 | 31.28 | 25.76 | 29.04 | 24.48 | 40.97 | 5.86 | 80.66 |
| AttentionUnet | — | 91.70 | 87.55 | 77.25 | 31.18 | 27.45 | 27.26 | 24.33 | 37.50 | 16.57 | 81.68 |
| EHANet[29] | ResNet18 | 91.22 | 87.35 | 75.08 | 31.69 | 35.36 | 36.87 | 21.86 | 37.24 | 27.06 | 79.29 |
| DeepLab v3+[8] | MobileNetV2 | 90.99 | 85.59 | 72.85 | 17.89 | 33.47 | 11.06 | 34.50 | 31.64 | 21.78 | 74.51 |
| Ours | ResNet50 | 93.59 | 88.6 | 82.61 | **67.54** | **67.91** | **61.46** | **60.30** | **65.61** | **65.52** | 86.63 |
| 方法 | Backbone | 上唇 | 下唇 | 头发 | 帽子 | 耳环 | 项链 | 脖子 | 衣服 | *mIoU* | PAcc |
| PSPNet | ResNet50 | 80.75 | 83.69 | 93.34 | 78.00 | 54.94 | 49.04 | 86.10 | 86.38 | 73.88 | 95.60 |
| DANet | ResNet50 | 80.05 | 83.18 | 91.50 | 75.65 | 45.82 | 68.30 | 83.35 | 76.92 | 60.82 | 94.23 |
| FCN | HRNet | 82.43 | 84.89 | 92.34 | 81.11 | 53.65 | 15.81 | 85.29 | 81.84 | 65.09 | 94.92 |
| BiseNet | — | 79.67 | 82.79 | 90.13 | 68.43 | 40.62 | 0.21 | 81.77 | 68.53 | 58.66 | 93.37 |
| UNet | — | 74.68 | 79.80 | 89.07 | 49.86 | 36.75 | 0.00 | 79.25 | 59.18 | 55.27 | 92.25 |
| AttentionUnet | — | 75.94 | 80.58 | 89.43 | 52.78 | 38.97 | 0.00 | 80.25 | 61.86 | 56.42 | 92.57 |
| EHANet | ResNet18 | 75.80 | 77.77 | 89.22 | 65.74 | 37.56 | 0.00 | 79.25 | 63.39 | 57.99 | 92.67 |
| DeepLab v3+ | MobileNetV2 | 70.35 | 75.07 | 89.19 | 65.90 | 36.57 | 0.00 | 78.69 | 68.37 | 55.21 | 92.55 |
| Ours | ResNet50 | 80.30 | 83.5 | 93.15 | **83.88** | 51.37 | 24.17 | 84.97 | 83.41 | **74.64** | **95.62** |

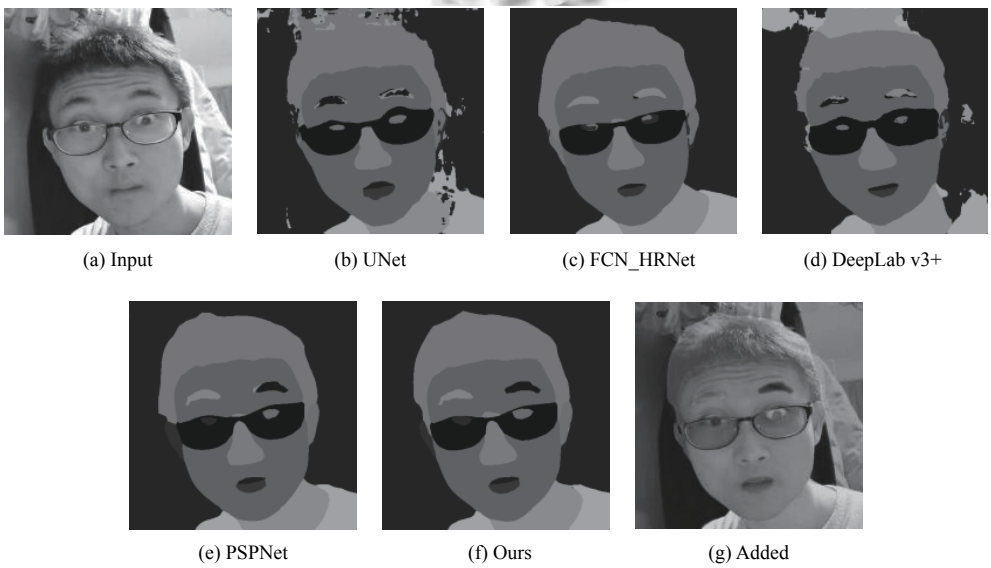
(a) 输入图 (b) 真值图 (c) UNet (d) PSPNet (e) DeepLab v3+ (f) FCN_HRNet (g) Ours

图 7 CelebAMask-HQ 可视化结果

使用作者本人的图片进行模型测试, 结果如图 8 所示.可以看到, UNet、DeepLab v3+等模型无法做到语义类别的准确预测, 无法清晰的辨别对象空间位置信息; FCN、PSPNet 由于未充分使用主干网络的特征图, 对于空间位置的解析不够精确; DBMSNet 达到了最佳的预测结果, 兼具语义类别的准确性与空间位置的精确性.

## 3 结论

本文提出一种高精度语义分割网络称为双分支多层级语义分割网络 (DBMSNet). 首先使用残差网络提取到 4 级分辨率由大到小的特征图 (*De*1, *De*2, *De*3 和 *De*4); 其次将 *De*1 和 *De*3 通过本文提出的 FR 模块与 MDM 模块, 充分混合空间位置特征的同时编码上下文信息及多尺度感受野, 此输出为第一分支; 然后将 *De*4 通过 PPM 模块, 目的是提取高级语义信息, 此输出为第二分支; 最后将两个分支进行融合输出, 达到空间位置特征与高级语义信息融合的目的, 完成高精度的图像分割任务. 文中多组消融实验充分表明了本文所提模块的有效性. 最终实验结果表明, 本文所提模型在相同数据集的精度明显优于文中列出的现有模型, 在人脸解析数据集 CelebAMask-HQ 取得最高精度 74.64%, 在 Cityscapes 数据集取得 78.29% 的精度. 所提模型兼顾分割对象的空间位置特征与高级语义特征, 具有较好的性能.



(a) Input (b) UNet (c) FCN_HRNet (d) DeepLab v3+

(e) PSPNet (f) Ours (g) Added

图 8 真人输入可视化结果

## 4 结语

在后续的研究中, 继续两方面的内容: (1) 加快模型的推理速度, 同时需保持高的解析精度, 本文认为分割精度是语义分割工作的核心要义. (2) 提升分割对象的边界准确率, 边界问题仍然存在于本文提出的模型中, 需要进一步改进.

## 参考文献


1 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

2 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015. 1–14.

3 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1–9.

4 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

5 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269.

6 Wang RJ, Li X, Ling CX. Pelee: A real-time object detection system on mobile devices. Proceedings of the 32nd Conference on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 1967–1976.

7 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. Proceedings of the 4th International Conference on Learning Representations. San Juan, 2016. 1–13.

8 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

9 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6230–6239.

10 Fu J, Liu J, Tian HJ, *et al*. Dual attention network for scene segmentation. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3141–3149.

11 Paszke A, Chaurasia A, Kim S, *et al*. ENet: A deep neural network architecture for real-time semantic segmentation. arXiv: 1606.02147, 2016.

12 Zhao HS, Qi XJ, Shen XY, *et al*. ICNet for real-time semantic segmentation on high-resolution images. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 418–434.

13 Xiong ZQ, Wang ZC, Li J, *et al*. Using features specifically: An efficient network for scene segmentation based on dedicated attention mechanisms. IEEE Access, 2020, 8: 217947–217956. [doi: 10.1109/ACCESS.2020.3041748]

14 Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1800–1807.

15 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856.

16 Cordts M, Omran M, Ramos S, *et al*. The cityscapes dataset for semantic urban scene understanding. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3213–3223.

17 Lee CH, Liu ZW, Wu LY, *et al*. MaskGAN: Towards diverse and interactive facial image manipulation. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5548–5557.

18 Liu Y, Chu LT, Chen GW, *et al*. Paddleseg: A high-efficient development toolkit for image segmentation. arXiv: 2101.06175, 2021.

19 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495. [doi: 10.1109/TPAMI.2016.2644615]

20 Mazzini D. Guided upsampling network for real-time semantic segmentation. Proceedings of the British Machine Vision Conference. Newcastle: BMVA Press, 2018. 117.

21 Artacho B, Savakis A. Waterfall atrous spatial pooling architecture for efficient semantic segmentation. Sensors, 2019, 19(24): 5361. [doi: 10.3390/s19245361]

22 Lin GS, Milan A, Shen CH, *et al*. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5168–5177.

23 Bai X, Zhou J. Parallel global convolutional network for semantic image segmentation. IET Image Processing, 2021, 15(1): 252–259. [doi: 10.1049/ipr2.12025]

24 Yang MK, Yu K, Zhang C, *et al*. DenseASPP for semantic segmentation in street scenes. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3684–3692.

25 Bai X, Zhou J. Efficient semantic segmentation using multi-path decoder. Applied Sciences, 2020, 10(18): 6386. [doi: 10.3390/app10186386]

26 Wang PQ, Chen PF, Yuan Y, *et al*. Understanding convolution for semantic segmentation. Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE, 2018. 1451–1460.

27 Yu CQ, Wang JB, Peng C, *et al*. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 334–349.

28 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich: Springer, 2015. 234–241.

29 Luo L, Xue DY, Feng XL. EHANet: An effective hierarchical aggregation network for face parsing. Applied Sciences, 2020, 10(9): 3135. [doi: 10.3390/app10093135]