计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(3):73−84 [doi: 10.15888/j.cnki.csa.009411]

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

多尺度特征金字塔融合的街景图像语义分割①

曲海成, 王　莹, 董康龙, 刘万军

(辽宁工程技术大学 软件学院, 葫芦岛 125105)
通信作者: 王　莹, E-mail: lntuwangying@163.com

摘　要: 针对街景图像语义分割任务中的目标尺寸差异大、多尺度特征难以高效提取的问题, 本文提出了一种语义分割网络 (LDPANet). 首先, 将空洞卷积与引入残差学习单元的深度可分离卷积结合, 来优化编码器结构, 在降低了计算复杂度的同时缓解梯度消失的问题. 然后利用层传递的迭代空洞空间金字塔, 将自顶向下的特征信息依次融合, 提高了上下文信息的有效交互能力; 在多尺度特征融合之后引入属性注意力模块, 使网络抑制冗余信息, 强化重要特征. 再者, 以通道扩展上采样代替双线插值上采样作为解码器, 进一步提升了特征图的分辨率. 最后, LDPANet 方法在 Cityscapes 和 CamVid 数据集上的精度分别达到了 91.8% 和 87.52%, 与近几年网络模型相比, 本文网络模型可以精确地提取像素的位置信息以及空间维度信息, 提高了语义分割的准确率.

关键词: 语义分割; MDSDC; IDCP-LC; 属性注意力; 通道扩展上采样; 特征融合

引用格式: 曲海成,王莹,董康龙,刘万军.多尺度特征金字塔融合的街景图像语义分割.计算机系统应用,2024,33(3):73–84. http://www.c-s-a.org.cn/1003-3254/9411.html

Semantic Segmentation of Street Scenes Images Based on Multi-scale Feature Pyramid Fusion

QU Hai-Cheng, WANG Ying, DONG Kang-Long, LIU Wan-Jun

(Software College, Liaoning Technical University, Huludao 125105, China)

Abstract: This study proposes a semantic segmentation network called LDPANet to address the challenges of significant variations in target sizes and the difficulty of efficient extraction of multi-scale features in semantic segmentation tasks of street scene images. Firstly, the void convolution is combined with the deeply separable convolution introduced into the residual learning unit to optimize the encoder structure, which reduces computational complexity and alleviates the problem of gradient vanishing. Secondly, the network utilizes a layer-wise iterative void spatial pyramid to sequentially fuse top-down feature information, enhancing the effective interaction of contextual information. After multi-scale feature fusion, an attribute attention module is introduced to suppress redundant information and strengthen important features. Furthermore, channel-extended upsampling replaces two-wire interpolation upsampling as the decoder to further improve the resolution of feature maps. Finally, the accuracy of the LDPANet method on Cityscapes and CamVid datasets reaches 91.8% and 87.52%, respectively. Compared with the network model in recent years, the proposed network model can accurately extract pixel position information and spatial dimension information and improve the accuracy of semantic segmentation.

Key words: semantic segmentation; mixed depthwise separable dilated convolution (MDSDC); iterative dilated convolution pyramid with layer cascade (IDCP-LC); attribute attention; channel expansion upsampling; feature fusion

① 基金项目: 国家自然科学基金面上项目 (42271409); 辽宁省高等学校基本科研项目 (LIKMZ20220699)
收稿时间: 2023-08-31; 修改时间: 2023-09-26; 采用时间: 2023-10-09; csa 在线出版时间: 2023-12-26
CNKI 网络首发时间: 2023-12-28

语义分割[1,2]是计算机视觉场景推理任务中一个极为重要的研究方向, 其需要将图像中每一个像素分类并标签化, 以达到像素级别的解析识别. 同时也可以应用于实际视觉任务, 如在自动驾驶中以辅助系统做出障碍避让以及道路安全检测, 在医疗应用中对病理图像进行分割以辅助诊断, 以及在地理信息系统中为地信统计提供帮助.

深度卷积神经网络 (deep convolutional neural network, DCNN) 如 VGG16[3]、ResNet[4]以及 Xception[5]等为分类任务提供了一种有效的解决方案, 相比传统机器学习方法提高了分类精度. 在语义分割任务中, 基于深度卷积神经网络的 FCN 模型[6]取得开创性工作, 提出了一个完整的端到端像素分割模型. FCN 模型提出跳跃连接结构, 通过融合下采样特征图, 用浅层特征信息辅助分割结果的精细化和边界细化, 恢复同输入图像分辨率一致的分割结果. 一个缺点是 FCN 的下采样层损失了语义特征图分辨率信息, 且在获取高级特征图时感受野不够大, 使得在提取特征时会丢失空间结构信息. 为此, DeepLabv1[7]在 VGG16 模型中使用空洞卷积 (atrous convolution)[8]调整标准卷积滤波器的视野范围, 从而捕获更多范围上的上下文特征信息. 在进一步研究中, DeepLabv2[9]和 DeepLabv3[10]为了获取更加准确的预测结果, 在 ResNet 的基础上引入了空洞卷积, 作为特征提取骨干网络, 得到更高的分类精度. 在 DeepLabv3+[11]中, 以深度可分离卷积减少计算量, 训练更深层网络模型表征相应特征空间, 同时用空洞卷积来平衡精度. 一个充分的论证是空洞卷积产生网格效应 (grinding)[12], 这是由于其稀疏采样信号是从离散独立子集中得到的映射结果, 缺失像素位置相关性, 从而相邻像素丢失局部依赖关系. 近几年, 基于 DeepLabv3+模型改进的网络还有 ACFNet[13]、DECANet[14]和 DFANet[15] 在分割精度上都有所提升. 同样对于特征提取网络的研究中, 候选区域生成前景特征[16]的方法用来接收多尺度输入特征信息, 构建重叠候选集进行特征融合. DeepLabv3plus-IRCNet[17]在下采样过程引入特征图切分方法用以支撑对小尺度目标进行特征提取, 提升小目标分割任务的精度.

在编解码结构上为了获取多尺度信息, 使用一些特征融合的方法来联系语义分割中多尺度上下文的相关性. PSPNET[18]使用特征金字塔池化 (pyramid pooling module, PPM) 模块聚合全局同质上下文依赖, 不足的是忽视了类别间关联, 产生类别之间的混淆使分割效果不好. DeepLabv2、DeepLabv3 针对多尺度问题, 提出空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP) 结构捕捉上下文. Transformer[19]是借鉴语音识别中注意力机制取得成功的网络模型, 基于空间和通道轻量注意力机制 (convolutional block attention module, CBAM)[20]在经典语义分割网络中用以强化特征图的自适应性, 关注重要特征, 嵌入神经网络以增强网络模型表达能力. DANET[21]使用通道和位置双重注意力机制, 在局部特征上建模丰富的上下文依赖关系, 显著改善了分割结果. MFPNet[22]中的特征编码在多元数据中提取多层次特征, 金字塔池化在此基础上进一步提取多尺度特征, 实现了精细划分割. HANet[23]提出高度驱动注意力机制, 通过关注图像的垂直位置像素分布, 使用通道缩放提升分割性能. 此外, ENCNET[24]、DFN[25]使用全局池化, 来聚合全局上下文信息. SPAFBA[26]可以建立上下文依赖关系, 有效地增强特征表示能力, 但在改善边缘细节特征提取上还有待提高. SegFix[27]中关注分割结果的边界信息, 通过学习边界信息和内部像素的对应关系, 使用内部像素的预测去代替边缘的不可靠分割结果, 从优化边缘部分处理的角度改善了分割结果. 编解码结构中, 编码器结构基于深度神经网络 (DCNN) 提取高维特征语义信息, 解码阶段通过卷积层和双线性插值等方法, 将特征图逐步恢复为原始图像的分割结果. 插值结构对于精确表征空间结构信息而言, 其能力是有限的, 会导致数据依赖表示缺乏, 独立的像素表示没有考虑预测结果之间的相关性. DeepLabv3+融合编码阶段的冗余特征图, 用细化双线插值上采样特征, 增加了额外的模型推理时间.

以上方法在图像语义分割任务中均取得了较好的效果, 但是仍存在模型难以泛化映射多尺度信息, 使解码器在上采样过程中由于丢失特征图中的细节信息导致的预测不准确等问题, 因此本文针对以上问题基于 DeepLabv3+模型做出了以下贡献.

(1) 在提取特征时使用混合深度可分离空洞卷积 (mixed depthwise separable dilated convolution, MDSDC), 可以在保持较高的模型性能的同时, 减少模型的参数量和计算开销.

(2) 为提升多尺度目标建模能力, 本文提出了可以进行信息交互的层传递迭代空洞空间金字塔 (iterative dilated convolution pyramid with layer cascade, IDCP-

LC) 的结构来捕获语义分割类别相关性, 增强上下文信息相互关联程度, 融合不同尺度特征信息.

(3) 在多尺度特征融合之后, 加入了一种属性注意力 (class activation mapping attention, CAM attention)[20] 使网络更好地关注模型的重要信息.

(4) 最后, 利用通道扩展上采样解码器代替双线插值上采样, 对融合后的特征结果图进行通道压缩转置, 补全空间结构信息, 增强解码器的数据相关性依赖, 恢复预测图分辨率, 得到预测结果.

1 本文方法

1.1 LDPANet 网络结构

针对复杂场景中语义分割难以提取多尺度信息以及存在目标比例多样性变化, 本文提出了一种基于多尺度特征金字塔融合的语义分割网络 (layered dilated pyramid attention net, LDPANet), 如图 1 所示. 在编码器阶段, 构建基于混合深度可分离空洞卷积的深度神经网络提取特征, 以 1/8 倍下采样提取高级抽象语义特征信息, 获得准确的像素预测分类. 模型通过结合上下文信息, 对高级特征结果提出重采样的迭代空洞空间金字塔 (IDCP-LC), 实现了每一层次特征之间的有效交互, 更好地融合多尺度信息, 在跨尺度特征整合之后, 模型引入了属性注意力模块 (CAM attention), 以进一步提高模型对关键信息的感知能力. 在编码器末端, 使用通道压缩操作来减少模型的参数数量. 最后, 为了提高分割结果的分辨率, 解码器部分采用了通道扩展上采样的方式, 将特征图 $F^{h\times w\times c}$ 进行通道转置扩展, 利用通道维度信息, 代替双线插值方法, 恢复 s 倍上采样预测结果图 $Y^{H\times W\times N}$, 其中 N 为 1×1 卷积输出通道数. 最后使用非线性激活单元输出预测结果, 使整体网络模型泛化能力得到提升, 提高了分割结果的精度.

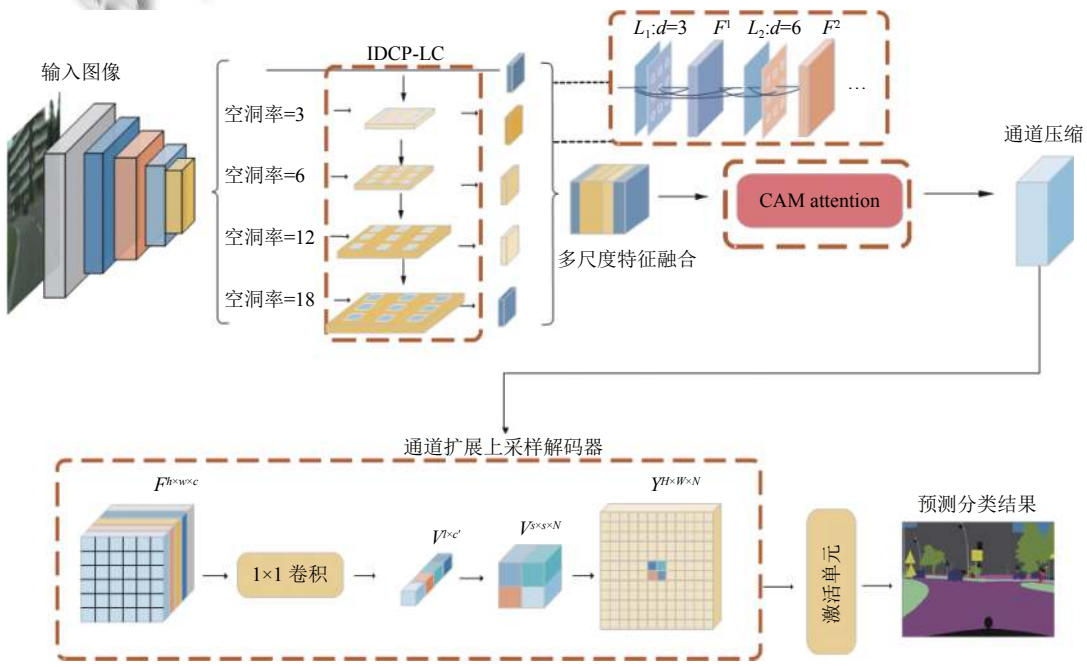

图 1 总体网络结构模型

1.2 混合深度可分离空洞卷积

深度可分离卷积解决了标准卷积中的计算效率低和参数数量膨胀问题. 然而深度可分离卷积在提取特征图的特征信息过程中, 会丢失部分空间结构信息, 因此在逐深度卷积过程中引入空洞卷积, 称为深度可分离空洞卷积, 该卷积在扩大感受野的同时还不丢失特征图的分辨率, 但是达到一定深度后必然引起网络退化, 出现梯度消失. 残差学习单元可以解决深度神经网络中的梯度消失和模型退化问题, 其结构中引入了跳跃连接, 允许网络直接将输入传递到输出, 以便更好地传播梯度和保留重要的特征信息. 因此本文将深度可分离空洞卷积和残差学习单元有效结合, 提出了混合深度可分离空洞卷积 (mixed depthwise separable dilated convolution, MDSDC). 可以在降低计算复杂度的同时

提高模型的表达能力和学习能力.

对于二维卷积输入特征向量$X^{H,W}$由式 (1) 计算可得输出结果$y_{[i,j]}$来表示映射学习:

$$y_{[i,j]}=\sum_{i=1}^{H}\sum_{j=1}^{W}x[i+d\cdot L,j+d\cdot L]k[L,L] \tag{1}$$

其中, $k[L,L]$为卷积核权重参数. 相比普通二维卷积计算并不复杂, 在计算过程中以无效 0 填充卷积核, 形成空洞率d, 以调整适应感受野范围, L为滤波器尺寸. 在二维的特征图映射中, 通过滤波器间填充孔洞形成二维空洞卷积, 用以保留下采样特征图的空间分辨率. 其中滤波器算子感受野分辨率为:

$$f=L+(L-1)\times(d-1) \tag{2}$$

当空洞率d为 1 时为普通卷积. 单个卷积核$k^{d\times d}$同时将输入特征图$F^{h\times w\times c}$的通道相关性和空间相关性映射到输出特征, 需要滤波器同时具有通道维度和空间维度, 用C_{in}和C_{out}表示输入特征图通道和卷积输出通道变化, 单个卷积核输出特征图的面积为M^2, 每一层的卷积计算时间复杂度为:

$$T\sim \text{O}(M^2\cdot L^2\cdot C_{\text{in}}\cdot C_{\text{out}}) \tag{3}$$

将空间相关性和通道相关性从普通卷积中分离, 分别独立映射表示, 即先将特征图的每个通道映射表示, 然后以1×1卷积逐点覆盖特征图, 利用特征图的不同通道位置的空间信息, 从而将单个普通卷积进行优化处理, 过程如图 2 所示.

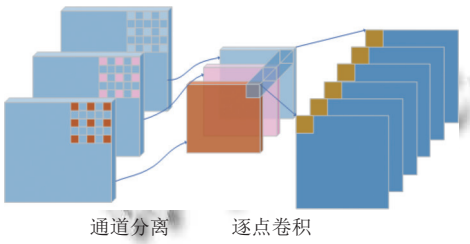

图 2 空洞深度可分离卷积

图 2 中通道映射以空洞卷积控制整个过程的感受野, 形成空洞深度可分离卷积, 以解耦方式优化时间复杂度:

$$T\sim \text{O}(M^2\cdot L^2\cdot C_{\text{in}}+M^2\cdot C_{\text{in}}\cdot C_{\text{out}}) \tag{4}$$

在深度可分离卷积中将通道维度单独映射后, 整个卷积过程中空间复杂度为:

$$S\sim \text{O}(L^2\cdot C_{\text{in}}+C_{\text{in}}\cdot C_{\text{out}}) \tag{5}$$

为了进一步加深网络, 在特征提取网络结构中通过残差结构构成 3 层残差学习单元. 通常在一个卷积模块中连接映射特征空间可表示为$H(x)$, 则一个残差层i的学习空间可表示为:

$$y_i=H(x_i)+F(x_c,w_i) \tag{6}$$

在直接映射结构中连接映射$H(x)\to x$, F是构建的残差部分, 映射部分通过1×1卷积降维, 其中连接映射表示为$H(x_i)=\omega_i x_{i-1}$, 因此对于残差映射单元的优化目标不是 H, 转而学习残差部分$H(x)-x$, 对于构建的多层$L=\{1,2,\cdots,l\}$残差映射单元则表示为:

$$X_L=H(x_i)+\sum_{i=1}^{L}F(x_c,w_i) \tag{7}$$

本文提出的结合残差学习单元的混合深度可分离空洞卷积, 其特征提取结构如图 3 所示. 其中S为卷积步长, C为卷积层输出通道数. 在加深特征提取网络, 同时中间层嵌套串联 8 个残差模块来提取多尺度信息, 使网络在浅层和深层调整特征图分辨率, 有效地提升对分割物体的预测精度.

1.3 层传递的迭代空洞卷积金字塔

由于同一分类目标在近景远景中会出现不同尺度问题, 针对这一问题一般会引入金字塔结构, 金字塔池化结构通过全局平均池化来获取高级特征图理解上下文信息, 但因其参数量过大而无法拓展到卷积层来增强上下文理解. 空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP) 模块, 即使以空洞卷积提升感受野范围, 但多尺度目标还是存在空间分辨率和几何边界的矛盾, 浅层网络特征图的分类精度低, 影响预测分类的准确性, 深层网络虽然准确分类却丢失目标边界信息. 所以提出了层传递的迭代空洞空间金字塔 (iterative dilated convolution pyramid with layer cascade, IDCP-LC), 其结构如图 4 所示. 结合了依赖上下文信息的特点, 通过在空洞空间金字塔中加入可以互相反馈信息的层传递结构对特征图重采样, 进而解决原有模型难以映射多尺度信息的问题. 具体地, 将上一层的输出与当前层的输入相加, 使每一层的信息都可以依次下传, 实现了多尺度特征的有效交互, 从而保留更多的上下文信息, 并减轻梯度消失的问题. 此外, 通过在模型的后处理阶段中使用多个尺度的预测结果, 进一步提高了模型的准确性.

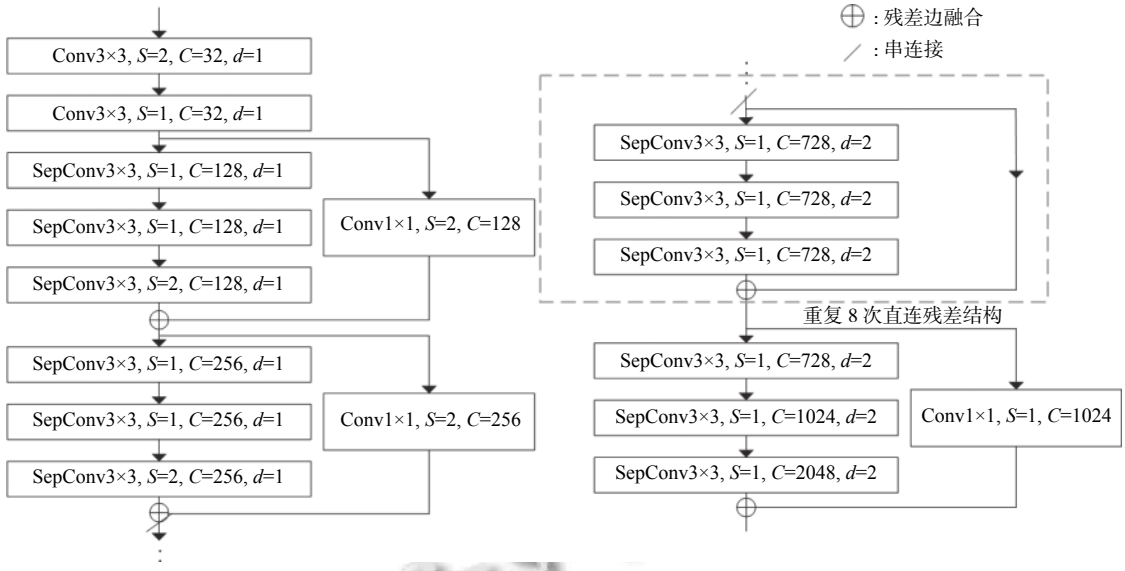

图 3 特征提取结构

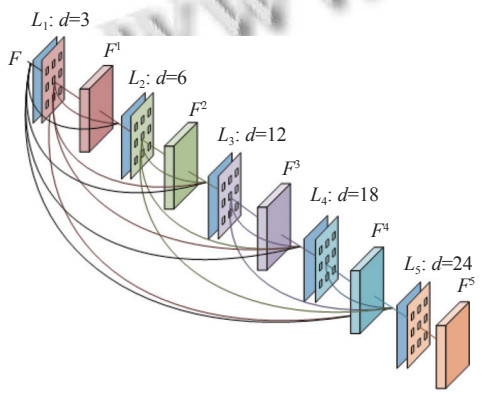

图 4 迭代空洞卷积金字塔

在下采样得到特征图映射F后, 通过不同膨胀率d的空洞卷积构成金字塔层结构P_i, 将高层金字塔输出特征图F^i迭代到其他一些感受野较大的层中, 共享采样信号, 在每层结构P_i中, 首先用1×1的卷积层映射融合特征图, 减少输入特征图的通道, 降低计算复杂度, 再用3×3空洞卷积层捕获特征图的多尺度上下文相关性, 其中金字塔层的特征图映射关系为:

$$F^i = P_i(F^1, F^2, \cdots, F^{i-1}) \tag{8}$$

以级联方式共享每层不同感受野特征, 整个金字塔结构融合的多尺度特征图感受野分辨率Δ_f为:

$$\Delta_f = [f(P_1), f(P_1)+f(P_2)-1, \cdots, \sum_1^n f(P_i)-i+1] \tag{9}$$

其中, f由式 (2) 可得. 每层的卷积核尺寸为 3, 空洞卷积膨胀率为$\Phi = \{3, 6, \cdots, 6(i-1), \cdots\}$ $(i \neq 1)$, 相对于普通并行结构空洞金字塔池化 (ASPP) 方法:

$$\Delta_{\text{ASPP}} = [f(P_3), f(P_6), f(P_{12})] \tag{10}$$

层传递的迭代空洞金字塔能以更多的感受野密集覆盖特征图, 加权映射多尺度信息, 提升网络分割效果.

1.4 属性注意力模块

注意力机制在计算机视觉和自然语言处理等领域中被广泛应用, 它本质上是通过矩阵乘法的方式来对输入进行加权处理, 从而使得模型可以更加关注重要的信息. 注意力机制可以用于确定输入中各个元素 (例如图像中的像素或文本中的词) 之间的依赖关系, 并根据这些依赖关系调整它们的权重, 从而实现对不同元素的加权聚焦, 本文提出了属性注意力模块, 如图 5 所示.

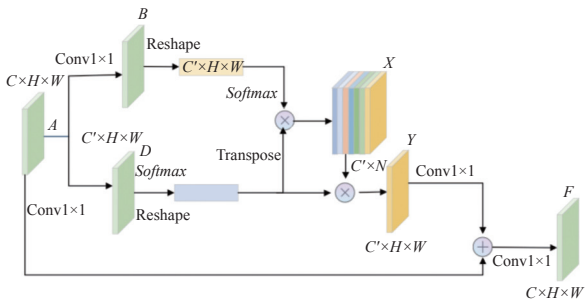

图 5 属性注意力结构

属性注意力模块可以从输入特征中选择性地聚焦于与目标任务相关的属性或类别信息. 这种模块可以通过对输入特征的通道维度进行建模, 来捕捉远程上

下文信息, 即图像中每个类别与输入特征每个通道之间的关系, 并根据这些关系选择一个最相关的类别或标签. 这种机制能够提高模型的分类准确率, 并提供一种直观的方式来理解模型如何进行分类进而增强每个类别之间的上下文信息依赖性. 首先将主干网络提取的特征经过融合之后的输出特征定义$A \in R^{C\times H\times W}$, 其分别通过两个$1\times1$卷积生成特征图$B \in R^{C'\times H\times W}$和属性关注度特征图$D \in R^{N\times H\times W}$, 其中 C'是B降维后的通道数, N代表图像分类中的类别数. 接下来将B转换变为 $B \in R^{C'\times HW}$, 同时D经过激活函数 *Softmax* 后变换得到$D \in R^{N\times HW}$, 将$B \in R^{C'\times HW}$和$D \in R^{N\times HW}$的转置相乘通过 *Softmax* 函数生成聚集所有属性的相似性映射图$X \in R^{C'\times N}$, 具体运算如式 (11) 所示:

$$a_{u,k}=\sum_{i=1}^{HW}B_{u,i}\frac{e^{D_{k,i}}}{\sum_{j=1}^{N}e^{D_{j,i}}},\ x_{u,k}=\frac{e^{a_{u,k}}}{\sum_{j=1}^{N}e^{a_{u,j}}} \tag{11}$$

其中, $B_{u,i}$表示特征图B第u个通道的第i个像素值, $D_{k,i}$表示特征图D第k个通道的第i个像素值, $a_{u,k}$表示B_u和D_k之间的属性特征关联矩阵$x_{u,k}\in X$表示属性之间的影响因子, $u\in[1,2,\cdots,C']$, $k\in[1,2,\cdots,N]$. 将$X\in R^{C'\times N}$和$D\in R^{N\times H\times W}$相乘得到$Y\in R^{C'\times H\times W}$, 将其通过$1\times1$卷积和$A$相加, 最终输出的属性增强特征图如式 (12) 所示:

$$F_u=f\left(f\left(\sum_{k=1}^{N}\left(x_{u,k}\cdot\frac{e^{D_k}}{\sum_{j=1}^{N}e^{D^J}}\right)+A_u\right)\right) \tag{12}$$

其中, F_u 表示表示输出特征 $F\in R^{C\times H\times W}$第$u$个通道, $f(\cdot)$表示1×1卷积-BN-ReLU 系列运算. 如式 (11) 显示每个通道的最终输出是属性特征注意图中所有通道基于类别的加权和, 表示特征图之间基于类别的语义依赖, 也就是提出的 CAM attention 直接提高了类别级信息的感知和辨别能力.

1.5 通道扩展上采样

解码器结构中, 通常输入特征图尺寸为网络模型输入图像的1/8或者是1/16, 然后解码器去恢复预测图像空间分辨率. 最为简单有效的方式是通过插值方法上采样, 但存在的问题是没有考虑预测分类像素间的相关性, 其重建恢复能力有限同样跳跃连接低级特征图的方式约束限制特征空间, 二次特征聚合导致解码器难以优化. 由此提出解耦空间分辨率和通道冗余特征的通道转置上采样结构, 具体来说, 使得到的融合特征图$F^{h\times w\times c}$, 利用通道c的大量标签特征信息, 重构至输出标签特征和空间分辨率, 得到恢复预测特征图$Y^{H\times W\times N}$如图 6 所示. 同样将真值标签通过独热编码 (one-hot) 为$G^{H\times W\times N}$, 和预测图结果图$Y^{H\times W\times N}$作多类别交叉熵损失, 其中N为语义分割类别数, 上采样倍数为$s=H/h$, 为了避免复杂的计算, 使用1×1逐点卷积, 输出通道为$c'=s\times s\times N$, 映射多尺度特征图$F'\in\mathbb{R}^{h\times w\times c'}$, 这样将特征图$F'$的每个通道维度作为向量$V_i^{1\times c'}$, 遍历特征图$F'$的每个通道, 将向量$V_i^{1\times c'}$转置并重排结构为三维张量$\mathbb{V}_i^{s\times s\times N}$, 组成整个预测结果特征图$Y^{H\times W\times N}$. 相比于双线插值的方法, 通道压缩的方式利用数据相关性, 增强数据依赖, 重组结构以缩放尺度s进行上采样, 并且额外只需少量的逐点卷积运算学习映射.

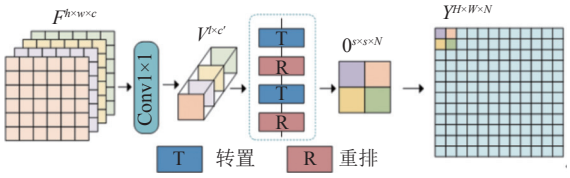

图 6 通道扩展解码器

通过扩展通道数和上采样操作, 通道扩展上采样可以帮助网络从低分辨率输入图像或特征图中恢复高分辨率的细节信息, 生成更准确的预测结果.

2 实验结果与分析

2.1 实验环境配置与数据集

本文实验基于 TensorFlow 2.5.0 深度学习框架搭建整体网络模型, 操作系统为 Ubuntu 18.04, CPU 为 Intel Core i9 10900K, 利用 NVIDIA RTX TITAN 24 GB GPU 训练模型, 内存为 32 GB. 在训练网络模型时选取实验参数为 epochs=30、Steps_per_epoch=400、validation_steps =10.

本文网络的有效性和泛化性用 Cityscape 和 CamVid 数据集进行验证.

自动驾驶城市街景数据集 Cityscapes, 包含 5 000 张分辨率为1024×2048的街景图像, 其中 2 975 张训练集, 500 张验证集和 1 525 张测试集, 划分 34 类语义分割目标, 提供了像素标注. 经过设置背景类, 实验在 19

个精细注释类别中训练网络模型. 为了扩充数据多样性, 在模型训练过程中使用对输入图像减均值的方法, 同时对图像和标签数据在水平和垂直方向随机翻转, 随机裁剪图像或填充, 避免训练过拟合, 使得模型输入图像分辨率为512×512.

CamVid 数据集是从视频序列中提取著名街景数据集, 从驾驶汽车的角度拍摄, 增加了观察对象类的数量和异质性. 它包含 701 张带注释像素级别的语义分割和图像分割, 32 个 ground truth 语义标签, 数据集中图片和标签的分辨率为 480×360. 数据集在训练或测试之前, 先对图片进行预处理, 将分辨率为 (480, 360) 图片归一化处理为 (512, 512). 当进行模型训练时, 将原始数据集分割成为训练集, 验证集和测试集 3 个类别: 367 张图片用于训练、101 张图片用于验证、233 张图片用于测试. 实验选取图像中的 11 个类别进行语义分割, 包含天空、建筑、柱杆、道路、树、人行道、围栏、车辆、行人、骑车的人、其他.

对于网络模型训练, 使用 Adam 优化器, 初始化学习率为 $init\_lr=3\mathrm{E}-4$, 通过“poly”学习策略调整学习率:

$$lr=init\_lr\times\left(1-\frac{epoch}{max\_epoch}\right)^{decay} \tag{13}$$

其中, $decay$为衰减指数, 调整学习率下降速度, 实验过程中$decay$设置为 0.9 学习更多最优值.

2.2 评价指标

采用语义分割标准度量方法类别平均交并比 (mean intersection over union, $MIoU$) 和平均像素精度 (mean pixel accuracy, MPA) 定量分析实验结果, MPA 计算每个类的被正确预测的比例, 体现为像素分类准确率, 而 $MIoU$ 严格地计算预测类别中真实值的交集和并集的比值平均值, 更好表征语义分割效果, 两者分别定义为:

$$MPA=\frac{1}{K+1}\sum_{i=0}^{K}\frac{p_{ii}}{\sum_{j=0}^{K}p_{ij}} \tag{14}$$

$$MIoU=\frac{1}{K+1}\sum_{i=0}^{K}\frac{p_{ii}}{\sum_{j=0}^{K}p_{ij}+\sum_{j=0}^{K}p_{ji}-p_{ii}} \tag{15}$$

其中, K为分类标签数, $K+1$为加上背景以后总标签类别, 其余忽略标签为背景, p_{ii}表示为预测分类正确的像素数, p_{ij}表示为错误分类像素数.

2.3 消融实验分析

设计消融实验系统地验证提出模块对图像分割任务的有效性: 1) 金字塔层数选择; 2) 不同编码结构选择; 3) 不同解码器上采样选择; 4) 多尺度上下文依赖模块性能; 5) 加入不同模块对性能的影响.

2.3.1 金字塔层数选择

在迭代空洞卷积金字塔模块中, 由式 (8) 看出增加的金字塔层提升重采样感受野, 以额外的计算代价得到更多的融合特征. 在训练输入图像分辨率512×512的限制下, 编码器1/8倍下采样后输入金字塔特征图分辨率为64×64, 当金字塔层结构超过P_6时, 式 (2) 计算得到P_7层 (空洞率=36) 的卷积分辨率为73×73, 超出特征图范围, 空洞卷积退化为中心点的1×1逐点计算形式, 重采样无效且带来额外计算, 以此对$P=\{P_3,P_4,P_5,P_6\}$结构分析重采样感受野对实验结果影响, 分析得到层级结构的相应实验结果见表 1. 增设的级联层结构得到更大感受野, 平均交并比和平均像素精度相应得到提高, 分别在P_5和P_6得到最好效果. P_6结构中, 平均像素精度仅提升 1.3%, 但同时平均交并比略降 0.4%, 此时第 6 层金字塔卷积分辨率为61×61, 重采样得到额外特征信息有限, 综合平衡计算复杂度和模型效果, 以P_5结构作为实验迭代金字塔模块. 对于级联到更多感受野的特征图, 此时包含大量的通道数, 相应需要大量额外计算, 为此, 级联到新的层级, 首先通过1×1卷积对特征图通道降维, 以此降低时间复杂度和访存占用.

表 1 多尺度金字塔感受野实验结果

金字塔层	最大感受野	$MIoU$ (%)	MPA (%)
P_3	43	78.9	85.7
P_4	79	79.4	88.2
P_5	129	81.3	90.5
P_6	189	80.9	91.8

2.3.2 不同编码结构选择

编码器结构实验中, 通过空洞深度可分离卷积优化模型参数, 两类残差结构进行特征编码, 学习残差映射加深网络模型, 避免模型退化, 从而下采样提取高级语义特征, 对比残差网络 ResNet 骨架网络以及参数量变化实验结果见表 2.

卷积计算的结构映射大幅度减少模型参数, 可加深网络模型来提高模型精度, 同时空洞填充得到的感受野捕获更多边界信息.

表 2 不同的编码结构的分割性能

方法	*MIoU* (%)	*MPA* (%)	参数量 (M)
ResNet50+P_6	74.7	82.2	2.92
ResNet101+P_6	78.9	87.5	4.68
ResNet152+P_6	80.9	88.4	6.25
本文	**81.3**	**90.5**	**2.87**

注: 加粗字体为最优对比结果

2.3.3 解码器结构选择

为得到进一步的实验提升, 验证通道扩展方法上采样的可集成和有效性, 对解码器设置了 3 组消融实验, 对比分析插值方法和本文方法实验结果见表 3, 相比 1/16 特征图, 在 1/8 倍下采样过程中, 由于保留特征图更大分辨率需要额外的内存消耗, 同时通道扩展方法比插值方法提升平均交并比 1.6%, 分割效果可见于分割模型的对比实验. 通过特征图通道重构, 恢复预测图分辨率, 同时可附加于其他分割模型上采样模块, 获得更好的数据依赖.

表 3 不同解码器上采样结果 (%)

方法	*MIoU*	*MPA*
1/8特征图+双线插值上采样	81.3	90.5
1/8特征图+本文方法	**82.9**	**91.2**
1/16特征图+本文方法	81.1	86.6

注: 加粗字体为最优对比结果

2.3.4 多尺度上下文依赖模块性能

基于以上两组消融对比实验, 验证了编码器结构和迭代金字塔的可行性方案, 设置了一组多尺度模块对比实验, 以补充说明迭代金字塔结构的优越性. 具体实验结果和相应方法见表 4. 同基准方法空洞空间金字塔 (ASPP) 模块相比, IDCP-LC 结构的分割结果 *MIoU* 同比提高 3.7%, *MPA* 提高 2.6%. 而池化金字塔模块在分割表现上欠佳, 一个重要原因是空洞卷积相比池化能更好地扩展感受野, 但比直接上采样具有较大分割效果提升.

表 4 多尺度上下文依赖模块性能 (%)

方法	*MIoU*	*MPA*
直接上采样	73.5	83.4
池化金字塔 (PPM)	77.1	84.1
空洞空间金字塔 (ASPP)	79.2	87.9
迭代空洞卷积金字 (IDCP-LC)	**82.9**	**90.5**

注: 加粗字体为最优对比结果

2.3.5 不同模块性能

为了验证各模块对 LDPANet 在整个网络中所起的作用, 将 MDSDC 模块、IDCP-LC 模块和 CAM attention 模块融合在模型上. 实验测试在 5 种情况下对同一组图像进行语义分割, 不同模块性能比较见表 5 分析.

表 5 LDPANet 模型不同模块性能比较 (%)

New modules			*MIoU*	*MPA*
MDSDC	IDCP-LC	CAM attention		
—	—	—	80.5	89.02
√	—	—	80.9	89.98
√	√	—	81.2	90.69
√	—	√	80.9	90.72
√	√	√	**82.9**	**91.80**

注: 加粗字体为最优对比结果, “√”表示选择的模块

从表 5 中数据可以发现, 加入 MDSDC 模块后 *MIoU* 和 *MPA* 分别提升了 0.4% 和 0.96%, 再将 MDSDC 模块和 IDCP-LC 模块结合后 *MIoU* 值高达 81.2%, *MPA* 为 90.69%, 比不加该模块分别增加了 0.7% 和 1.67%. 再与 CAM attention 模块结合 *MIoU* 和 *MPA* 分别提升了 0.4% 和 1.7%. 从整体定量结果中得到, 3 个模块共同加入后, 本文实验平均交并比 *MIoU* 结果为 82.9%, 平均像素精度 *MPA* 结果为 91.80%, 分别提升了 2.4% 和 2.78%. 通过以上实验结果, 可以得出结论, 本文在改进特征提取编码器的基础上再引入属性注意力机制和融合多尺度特征的模块对于网络的表征能力和特征提取融合具有明显的改善效果.

2.4 分割模型对比实验

为了验证本文算法的有效性设计了对比实验, 将本文算法与近年相关研究对比, 定量分析不同模型在相同条件下对于图像语义分割多分类精度和平均交并比的影响, 相同条件下不同语义分割算法的分类精度、平均交并比、预测时间等对比结果见表 6. 分析表 6 中数据可以得出, 本文提出的算法对比其他模型在提高准确性的同时还兼顾了对模型效率的提升, 可以使理论更好地应用在实际中.

表 6 不同语义分割算法性能比较

Module	*MIoU* (%)	*MPA* (%)	Param (M)	Time (s)
DeepLabv3+	74.28	85.32	156.6	285.3
ACFNet	78.98	84.81	124.9	293.5
DECANet	76.31	87.52	122.6	267.3
DFANet	76.0	89.34	125.3	273.6
LDPANet	**81.3**	**91.8**	**109.8**	**251.8**

注: 加粗字体为最优对比结果

在同样的实验环境下, 比较本文网络模型和其他分割网络方法分割效果. 图 7 所示为各种网络模型算

法在 Cityscapes 数据集中的部分可视化结果, 其中注释为数据集 19 类标签和训练被忽略的无效标签, 具体分割类别交并比和像素精度见表 7 和表 8. FCN-8S 和 SegNet 这两种模型忽视多尺度上下文信息, 对交通灯和杆状物体这类小尺度目标预测恢复能力有限. PSPNET 和 DeepLabv3+分别通过 PPM 和 ASPP 处理多尺度特征对多尺度信息建模, 显著提升分割效果. 对比分割可视化效果, 本文模型分割效果明显, 在栅栏和杆状物等小尺寸目标中恢复了可信的结果. 如在第 7 行第 2 幅图中的杆状物, 实验中恢复了横跨细节部分, 相对于 DeepLabv3+在分割标准类别交并比上提高 4.0% *MIoU*, 上采样结果更加接近真值标签. 同样类似的分割表现有交通标志, 地形、人行道和行人等, 比基准网络 DeepLabv3+交并比 *MIoU* 分别提升 11.9%、6.4%、1.9%、4.7%, 类别像素精度 *MPA* 分别提高 10.5%、4.0%、1.0% 和 2.0%, 在第 7 行第 4、第 5 幅图可看出准确预测人行道及地形区域, 而从第 3 幅图中可看出, 只有本文实验结果预测出树木背后交通标志部分, 表明融合重采样多尺度特征图, 能恢复更多上采样需要的分辨率信息. 同样从定量分析中可看出, 本文迭代空洞卷积金字塔对于这种比例变化的有效性, 从自行车和摩托车的分割评测中可看出, 在类别交并比中分别提高 4.1% 和 1.2%, 即使在摩托车的像素精度上几近一致, 依然在同比例目标自行车中提高 3.1% *MPA*. 在消极表现上, 本文方法在墙体的预测上 *MIoU* 有所降低, 这对于本文来说, 对重采样特征图扩充更多的感受野, 而迭代空洞金字塔较大将会带来额外计算和过度采样问题, 因为大尺度目标更需要宽阔感受野, 由于本文迭代反馈的特殊性, 迭代过多更会造成采样重复及冗余. 在具体分类目标中, 本文方法在交通标志、杆状物、地形以及栅栏等有更加明显的分割效果.

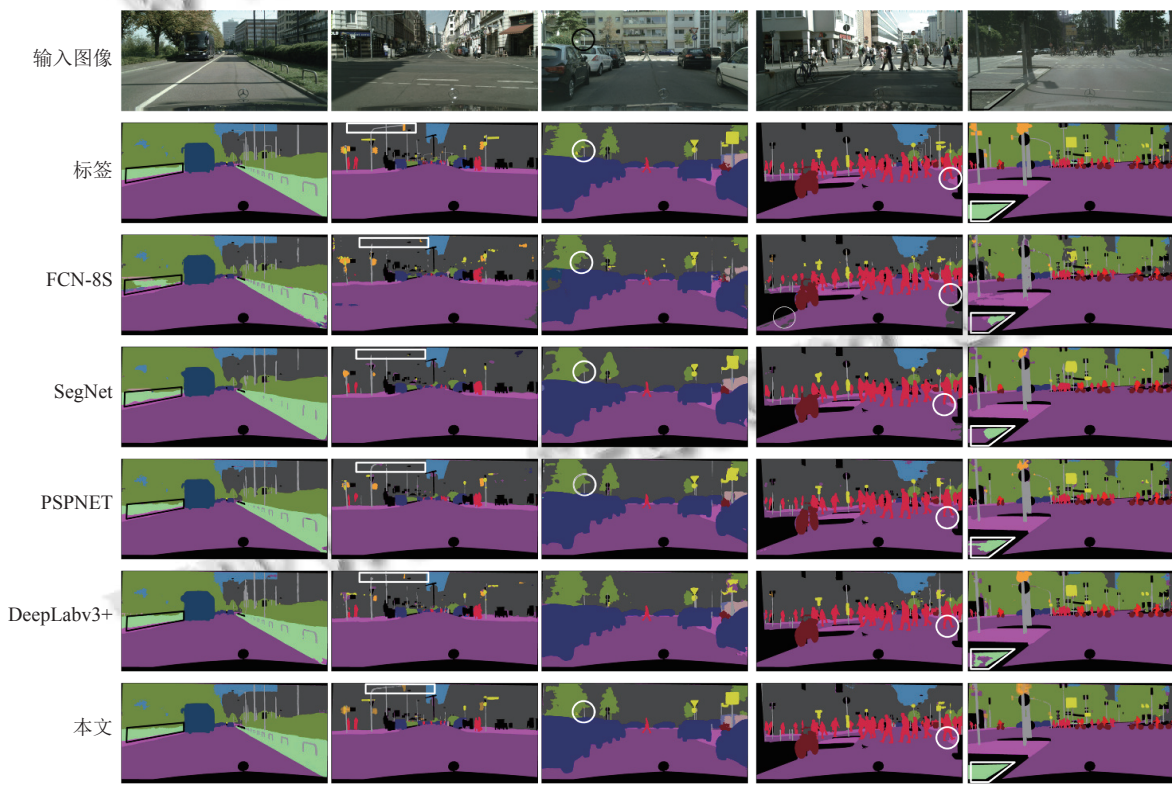

图 7 不同分割模型在 Cityscapes 数据集上的可视化结果

2.5 在 CamVid 上的实验分析结果

为了进一步验证该算法的泛化性, 实验在 CamVid 数据集上对改进后网络图像分割的效果进行测验, 记录的 *MPA* 值见表 9.

从表 9 可以看出, 绝大部分物体都可以得到更好的分割, 本文网络的 *MPA* 相比于 DeepLabv3+提高了 5.20%, 并且相较于其他先进模型都有提升, 进一步表明了该算法的普适和有用.

表 7 不同分割模型在 Cityscapes 数据集中类别平均交并比 (*MIoU*) 结果 (%)

方法	道路	人行道	建筑物	墙壁	栅栏	杆状物	交通灯	交通标志	植被	地形	天空	行人	骑手	汽车	卡车	公共汽车	火车	摩托车	自行车	平均*MIoU*
FCN-8S	96.1	78.6	84.7	47.9	51.1	44.0	33.7	54.4	88.9	61.7	91.3	67.8	40.7	88.2	59.8	76.7	52.1	31.4	64.4	63.9
SegNet	96.8	86.6	91.0	73.4	69.9	50.7	35.7	62.2	91.6	77.5	92.7	74.7	50.9	93.4	80.2	81.6	57.9	33.2	69.5	72.1
PSPNET	96.2	87.0	92.3	76.9	74.0	61.5	52.9	71.9	93.2	79.1	90.3	80.6	66.5	94.5	88.5	92.2	82.7	61.8	75.4	79.8
DeepLabv3+	97.4	90.5	92.8	81.9	80.1	63.6	51.5	68.7	93.3	79.8	93.7	81.7	64.1	95.3	89.9	90.9	79.7	57.4	77.7	80.5
本文	**97.8**	**92.4**	**94.5**	77.1	**84.2**	**67.6**	51.7	**80.6**	**94.0**	**86.2**	**93.2**	**86.4**	**68.4**	**96.2**	86.7	**94.4**	**84.7**	58.6	**81.8**	**82.9**

注: 加粗字体为最优对比结果

表 8 不同分割模型在 Cityscapes 数据集中类别像素精度 (*MPA*) 结果 (%)

方法	道路	人行道	建筑物	墙壁	栅栏	杆状物	交通灯	交通标志	植被	地形	天空	行人	骑手	汽车	卡车	公共汽车	火车	摩托车	自行车	平均*MPA*
FCN-8S	98.6	84.1	97.3	60.6	64.5	53.0	43.4	62.2	91.8	66.2	94.4	78.2	45.3	91.4	63.3	84.9	59.1	35.1	76.2	71.1
SegNet	99.4	91.2	96.7	81.9	75.9	58.3	39.4	68.8	95.0	83.4	96.4	85.6	62.5	96.3	89.0	88.6	68.4	37.2	80.9	78.7
PSPNET	99.6	90.8	96.6	81.2	77.6	69.7	57.1	77.1	96.6	82.7	92.0	88.7	75.0	97.1	93.7	95.2	86.3	65.9	79.8	84.4
DeepLabv3+	99.4	94.1	95.5	87.1	87.3	74.9	57.5	75.3	96.9	87.9	97.1	90.6	83.2	97.8	97.0	96.9	86.1	65.9	86.3	87.2
本文	**99.6**	**95.1**	96.8	**95.8**	**90.2**	74.2	**57.7**	**85.8**	96.5	**91.9**	**97.6**	**92.6**	76.0	**98.4**	94.2	**97.8**	**96.4**	65.2	**89.4**	**89.0**

注: 加粗字体为最优对比结果

表 9 不同语义分割算法 *MPA* 比较 (%)

Module	*MPA*
DeepLabv3+	82.32
ACFNet	82.71
DECANet	83.45
DFANet	84.56
LDPANet	**87.52**

注: 加粗字体为最优对比结果

为了更好地验证提出方法模型的优越性, 选取了 4 种不同的场景模型的效果图与 DeepLabv3+模型进行了比较, 语义分割效果对比如图 8 所示. 从选取的第 2 组场景图可以发现, 通过人类视觉很难辨别出远处红框内的道路旁边的行人, DeepLabv3+未分割出行人, 本文提出的算法不仅分割到了远处行人, 而且将行人的轮廓清晰分割出来. 图 8 中选取的第 4 组场景图是在弱光下采集的图像, 从效果图可以发现 DeepLabv3+模型未能分割出正确的类别, 算法实现了类别精确分割. 分析实验效果对比图可以发现, 在不同场景模式下本文提出算法分割效果比 DeepLabv3+模型分割物体的类别更准确, 边界更清晰, 分类效果更好. 实验将测试集语义分割的图像进行可视化与原图进行比较, 模型可视化效果图如图 9 所示. 从图 9 效果图可以发现, 网络模型将每个类别清晰的分割出来, 进一步验证了算法的优越性.

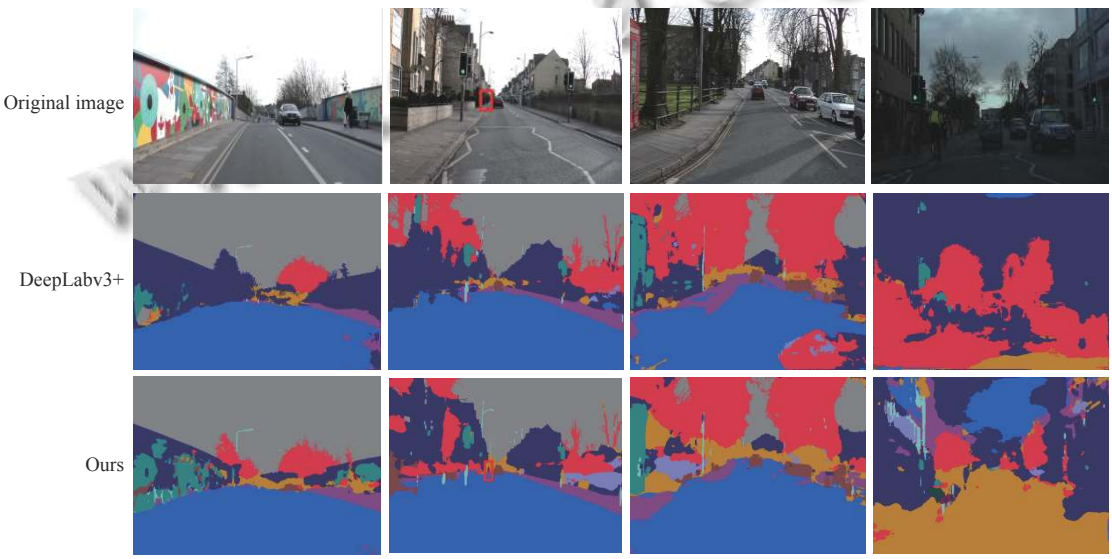

图 8 模型效果对比图

图 9 模型可视化效果对比图

3 结论

本文针对街景图像语义分割中的多尺度目标比例变化问题, 提出了一个端到端的网络模型. 通过引入混合深度可分离空洞卷积得到高级语义特征图, 再利用迭代级联空洞空间金字塔模块对高级特征图密集重采样, 获得多尺度特征融合的结果, 加入属性注意力机制捕捉关键信息来细化目标边缘像素和实现类别的高精度分割, 最后将得到扩充特征图通道信息, 使用通道扩展上采样解码器恢复特征图分辨率, 得到预测分类结果. 为评价本文网络模型的性能, 在城市街景 Cityscapes 和 CamVid 数据集上设置了多组定量分析实验, 证明了提出模型的有效性. 相对于其他分割方法, 本文算法可以预测到更多像素信息, 使街景图像实现更高精度的分割. 下一步研究将专注实现语义分割模型的轻量化, 使理论更好地部署在应用中.

参考文献

1 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.

2 Asgari Taghanaki S, Abhishek K, Cohen JP, *et al*. Deep semantic segmentation of natural and medical images: A review. Artificial Intelligence Review, 2021, 54(1): 137–178. [doi: 10.1007/s10462-020-09854-1]

3 陈铭, 梅雪, 朱文俊, 等. 一种新型 Mobile-Unet 网络的肺结节图像分割方法. 南京工业大学学报 (自然科学版), 2022, 44(1): 76–81, 91.

4 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440.

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

6 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015.

7 张焯林, 赵建伟, 曹飞龙. 构建带空洞卷积的深度神经网络重建高分辨率图像. 模式识别与人工智能, 2019, 32(3): 259–267. [doi: 10.16451/j.cnki.issn1003-6059.201903007]

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 刘漳辉, 占小路, 陈羽中. 基于语义传播与前/背景感知的图像语义分割网络. 模式识别与人工智能, 2022, 35(1): 71–81.

10 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 833–851.

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

12 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2881–2890.

13 Zhang F, Chen YQ, Li ZH, *et al*. ACFNet: Attentional class feature network for semantic segmentation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6797–6806.

14 唐璐, 万良, 王婷婷, 等. DECANet: 基于改进 DeepLabv3+ 的图像语义分割方法. 激光与光电子学进展, 2023, 60(4): 92–100.
15 Li HC, Xiong PF, Fan HQ, *et al*. DFANet: Deep feature aggregation for real-time semantic segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9514–9523.
16 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
17 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.
18 Fu J, Liu J, Tian HJ, *et al*. Dual attention network for scene segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3146–3154.
19 Wang ZM, Wang JS, Yang K, *et al*. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with DeepLabv3+. Computers & Geosciences, 2022, 158: 104969.
20 Choi S, Kim JT, Choo J. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9373–9383.
21 Zhang H, Dana K, Shi JP, *et al*. Context encoding for semantic segmentation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7151–7160.
22 Yu CQ, Wang JB, Peng C, *et al*. Learning a discriminative feature network for semantic segmentation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1857–1866.
23 Wu P, He XT, Tang MQ, *et al*. HANet: Hierarchical alignment networks for video-text retrieval. Proceedings of the 29th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2021. 3518–3527. [doi: 10.1145/3474085.3475515]
24 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
25 赵斐, 张文凯, 闫志远, 等. 基于多特征图金字塔融合深度网络的遥感图像语义分割. 电子与信息学报, 2019, 41(10): 2525–2531.
26 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]
27 Yuan YH, Xie JY, Chen XL, *et al*. SegFix: Model-agnostic boundary refinement for segmentation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 489–506.

(校对责编: 孙君艳)