小目标特征增强图像分割算法

任莎莎,刘 琼
(华南理工大学校软件学院,广东广州 511436)

摘 要: 在图像场景分割中存在小目标易丢失,边缘轮廓噪声大等问题.在目前的增强特征表征能力与优化空间细节的语义分割算法中,由于边缘和小目标特征的丢失,导致小目标和边缘很难被准确分割.为此,本文研究了一种小目标特征增强的图像分割算法.首先设计一种像素空间注意力模块(Pixel spatial Attention Module,PAM),来获得空间像素具有较强语义信息的特征图像.然后通过对PAM的输出进行建模提取,分别获得含有语义类别信息的边缘特征和小目标特征.最后,将特定的损失函数应用到语义分割训练中,并将多种特征进行融合,经过反复的监督学习和训练校正,可以在不影响其他类别性能的情况下提高边缘和小目标分割的性能.在Cityscapes,VOC2012,ADE20K和Camvid基线数据集上的实验表明,该算法与先进的图像分割算法相比,在小目标分割、边缘特征增强和内轮廓噪声减少等方面,其性能和效果都有明显提高,分割精度提高了2个百分点.

关键词: 场景分割;小目标特征增强;注意力模块;建模
中图分类号: TP751 **文献标识码:** A **文章编号:** 0372-2112(2022)08-1894-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211123

A Tiny Target Feature Enhancement Algorithm for Semantic Segmentation

REN Sha-sha, LIU Qiong
(*School of Software Engineering, South China University of Technology, Guangzhou, Guangdong* 511436, *China*)

Abstract: We have to face the challenge of missing small targets and severe edge noise in semantic segmentation. The existing semantic segmentation algorithms that enhance feature representation and optimize spatial details have difficulty to accurately segment the small targets and edges as the algorithms insufficiently gain detail information from tiny targets and semantic edges. This paper presents a tiny target feature enhancement algorithm for semantic segmentation. Specifically, a pixel spatial attention module(PAM) is designed to obtain strong semantic information from low-level pixel space. Semantic category information including edge features and tiny target features are obtained by modeling mask, respectively. A special loss function is designed for model training and the features gained by the model are fused with the features obtained from above way. Through edge feature enhancement, inner contour noise reduction, the segmentation performance of tiny target is improved while other segmentation categories are not degraded. Experimental results on Cityscapes, VOC2012, ADE20K and Camvid show that the proposed algorithm performance has been significantly improved by 2% in comparison with other state-of-the-art algorithms in the same scene.

Key words: scene segmentation; tiny target feature enhancement; attention module; modeling

1 引言

在场景解析中准确感知与理解图像内容,对于人工智能领域的计算机视觉至关重要[1].近年来,深度卷积神经网络,特别是VGG[2]、GoogleNet[3]、ResNet[4]等在目标识别方面取得了较大的成功,但对于神经网络的图像分割算法,其大多都由图像分类领域迁移而来,未能满足密集图像分类或分割等任务对网络特征表征能力强度的要求.对边缘、小目标等细节其语义类别关注较少.分类网络中频繁的池化操作与卷积步长的设置降低了空间分辨率[5,6],导致诸如交通信号等很多小目标被丢失.由于空间细节的丢失,又导致分割性能降低.

收稿日期:2021-08-18;修回日期:2022-06-22;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61976094);广东省自然科学基金(No.2021A1515011349)

本文在交通场景中研究类别像素数相对占比小于百分之一的小目标时发现,在复杂的现实世界中随着智能系统的应用与普及,这类小目标识别和分割需要重视. 例如,在自动驾驶的高分辨街景图像中其小目标很难被准确分割. 这严重影响了自动驾驶任务的安全行驶. 小目标分割难度在于目标小、亮度和边缘等特征浅、语义信息少、小目标和背景之间尺寸不均衡等;用较小的感受野关注其特征,很难提取全局语义信息;用较大感受野关注背景信息,小目标的特征会丢失. 在图像分割领域中人们做了大量工作. 虽取得了较好的成绩,但还不能满足对分割性能的需要.

早期为了提高小目标分割精度,采用一些基于上下文的后处理矫正方法. 比如 Chen 与 Krahenbuhl 等人[7,8]在 FCN 网络之后构建基于 MRF(Markov Random Field)与 CRF(Conditional Random Field)的上下文关系来矫正分割结果,提高小目标分割精度. 然而这些后处理方法无法参与训练过程,且网络不能根据预测结果调整权重. 为了保持图像分辨率,Pohlen 等人[9]提出全分辨率残差网络,在常用的网络旁并行设计一条不带有池化和步长大于 1 的分支,两条网络在前向传播过程中交互融合,保持小目标和边缘特征分辨率的同时获取语义信息,但是高分辨特征会带来宏大的计算开销. Guo 等人[10]提出在分割网络后设计新的损失函数增大网络对小目标的关注,该损失函数通过增加一个基于类间边界共享的 ISBMetric 指标,该指标通过测量目标类别间的空间相邻性,来缓解尺度带来的损失偏差,改善小目标分割. 由于他们定义的小目标类别有限,虽然设计的损失函数能提高网络对一些小目标类别的关注度,提高整体分割性能. 但均未能解决所有小目标训练样本不均衡问题. Yang 等人提出用合成图像来实现小目标数据增强方法,提高小目标分割精度[11]. 该方法主要通过建立合成的小目标数据与分割数据集共同参与训练. 增强了模型对小目标的训练,提高了模型对小目标的表征能力. 由于合成的小目标类别有限(取决于人为定义),仍未能解决未定义小目标分割问题. 因此,我们对网络高层特征首先进行空洞卷积池化金字塔 ASPP(Atrous Spatial Pyramid Pooling)处理,用得到的全局语义信息指导浅层的高分辨图像特征进行训练. 在少量增加计算开销的情况下,保持了浅层特征的分辨率与语义信息. 再通过建模提取所有小目标特征,最后训练学习矫正小目标类别,来提高小目标分割精度并取得了更好地效果.

对边缘分割的处理是场景分割任务中的关键技术之一. 由于网络自身问题(步长与池化)导致许多信息被丢失,特别是目标轮廓存在不连续、易混淆模糊、边缘信息甚至被丢失等现象. 先前一些工作[12,13]提出用 CRF 之类的结构来改善分割性能,尤其是围绕目标边界. Zhao 等人[14~16]提出构建特征金字塔池化结构,该结构通过聚合多个尺度的特征来获得多尺度上下文,以优化目标边界细节信息. Bertasius 和 Cheng 等人[17,18]提出同时学习分割与边界特征的检测网络,恢复池化层丢失的高分辨率特征. 而在工作[19,20]中提出通过学习边界特征作为中间表征来辅助分割. Takikawa[21]在已有分割网络中通过增加一个由门控网络构成的边缘形状学习分支网络来捕获图像中的边缘特征,在网络中引进多任务的损失函数来监督网络的训练过程,同时引入多任务的正则化项来防止过拟合. 由于该网络良好的边界特征学习能力,在小目标的分割精度上有大幅度提高. 不同于在网络中通过增强边缘特征来优化目标边界的方法,Ding 等人[22]提出了一种边界感知的特征传播网络,该网络把边缘设定为一种附加类,学习图像中的边缘得分,根据其得分在边缘像素点内进行特征信息的传播等. 以上工作取得了较好的成果但存在两个不足,一是虽然增强了网络特征中已有的边缘特征,但较小的目标细节没有得到恢复. 二是未区分目标大小,对所有大小目标使用相同的边缘增强准则. 为此,我们设计了一个强化外轮廓、弱化内轮廓的带有矫正的边缘增强模块,通过建模提取所有边缘特征,最后训练学习矫正边缘类别,来获得目标边界信息. 提高边缘分割精度较明显.

在本文中,我们旨在保证其他类别分割精度的基础上,提高了小目标和边缘等目标分割精度. 本文的贡献主要包含以下几个方面. (1)设计了一种像素空间注意力模块(PAM),可以获得具有较强语义的像素空间. (2)设计了一种新的小目标特征提取方法(Tiny Target Extraction module,TTE),并且获取的小目标特征含有语义类别信息. (3)设计了一种目标边缘特征的提取方法(Edge Extraction Module,EEM),该方法获取的边缘特征含有语义类别信息. (4)设计了一种新的损失函数,在监督图像,小目标,边缘三者训练任务的同时,矫正了边缘与小目标类别,也达到了增强边缘与小目标特征的任务. 最后实验结果表明我们的方法显著提高了细小目标的分割精度,总体分割精度(mIoU)与先进算法比较,提高了 2 个百分点.

2　相关工作

2.1　ASPP

将多尺度特征纳入深度卷积神经网络 DCNNs(Deep Convolutional Neural Networks)是使语义分割达到最优性能的关键因素之一. Chen 等人[7,16]提出了一种多尺度特征提取方案,通过 ASPP 来扩展空间接收场. ASPP 结构一般由不同膨胀率的空洞卷积并行组

成. 空洞卷积是一种常见的信号优化算法,由 Holschneider 等人在文献[23,24]中为实现非抽样小波变换的高性能和高效计算而提出. 设二维图像信号经过主干网络后,每一个位置 i 上的输入特征 x,经过卷积滤波器 w 得到对应的输出特征 y,对特征图 x 上进行空洞卷积的具体过程为:

$$y[i]=\sum_k x[i+rk]w[k] \tag{1}$$

其中 r 为空洞卷积的膨胀率,它表示对输入信号采样的步幅大小. 当 r=1 时为标准常规卷积. 通过修改 r 的值来获得适合不同尺度的目标感受野. 主干网络输出的特征,经过带不同膨胀率卷积的 ASPP 模块处理,增强了网络的感知能力,输出具有较高的上下文语义信息. 再与浅层特征融合,不仅能增加部分细节信息,获得满足不同尺度目标的语义信息,还在一定程度上缓解了膨胀卷积带来的栅格效应.

2.2 空间注意力机制

在目前的图像语义分割模型中,由 DCNNs 输出高层图像特征具有较高的语义信息,但缺少细节信息,而浅层的图像特征细节信息丰富但缺少语义信息,高层特征与低层特征简单融合很难使分割精度提高. 为此 Deeplabv3[16]和 PSPNet[14]使用多尺度特征提取方案来扩展空间接受场. 这些方案只关注局部特征关系,产生的上下文语义信息有限. 近期,CCNet[25]和 EMANet[26]采用空间稀疏注意力机制得到上下文信息,在不降低网络性能的前提下,降低了模型的计算复杂度,也提高了空间上下文语义信息. Zhong 等人[27]提出一个高效的压缩注意力网络结构(SANet),通过增强网络表征能力,使网络关注更多的细节. 然而,他们也没有考虑到像素和类别之间的关系来直接构建空间上下文信息. 而这些关系不仅有助于降低上下文中的噪声信息,还能使空间上下文更具解释性. 因此,这些基于空间上下文的方法在表征学习中如果未考虑有效通道信息,就不能获得较好语义信息. 为了得到像素空间具有较强的语义信息,我们设计了一种把空间注意力和通道注意力有机结合的像素空间注意力模块.

3 本文方法

3.1 图像的多重特征提取分割算法

本文算法结构如图 1 所示. 在主干网络 ResNet101 输出到 ASPP 模块,采取不同膨胀率的卷积来获得不同尺度的特征图. 很明显 ASPP 输出虽然可以得到较好的语义信息,但是最后一层网络特征图无法恢复丢失的所有信息. 所以我们把它输入到我们设计的一种像素空间注意力模块 PAM,可以得到适应不同尺度下的特征映射和具有较强的语义关系的像素空间特征. 使其输出到边缘特征提取模块 EEM 得到具有类别信息的边缘特征 $\boldsymbol{y}_{\text{edge}}$,并用边缘实况图对此特征进行监督学习. 同时输出另一路到小目标提取模块 TTE 得到具有类别信息的小目标特征 $\boldsymbol{y}_{\text{tiny}}$,并用小目标实况图对此特征进行监督学习. 并设计专门的损失函数. 最后,将得到地小目标特征 $\boldsymbol{y}_{\text{tiny}}$、边缘特征 $\boldsymbol{y}_{\text{edge}}$、ASPP 特征输出 $\boldsymbol{y}_{\text{ASPP}}$ 与主干网络浅层特征 $\boldsymbol{y}_{\text{res1}}$ 融合输出,经过反复的监督学习和训练修正,我们可以在不降低其他类别性能的前提下,提高边缘和小目标的分割性能. 具体公式如下:

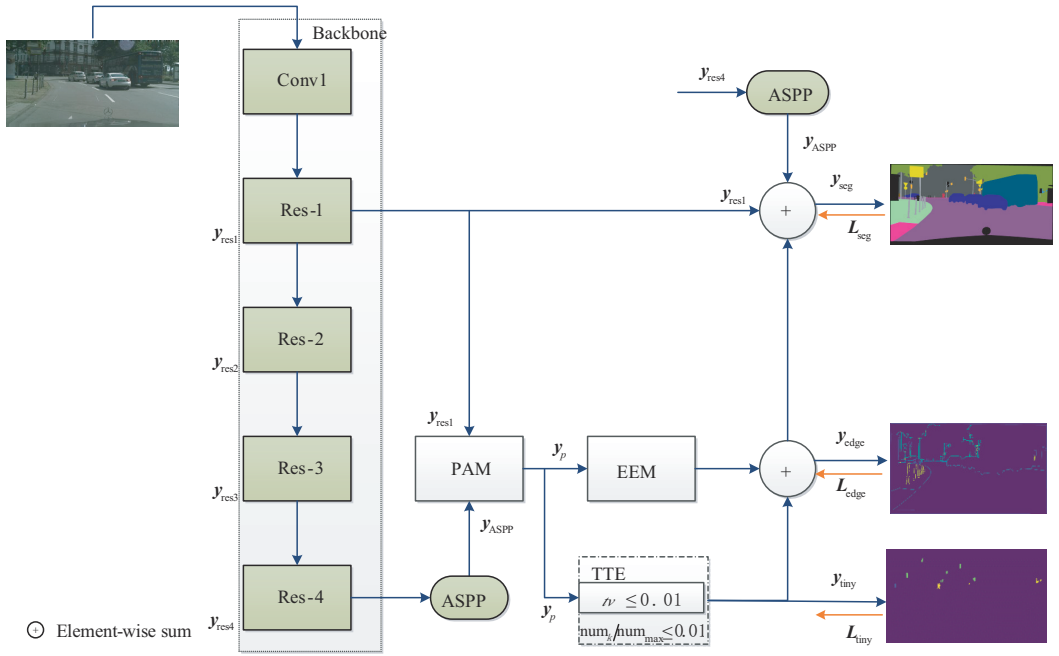


图 1 本文算法整体流程图

$$\boldsymbol{y}=\boldsymbol{y}_{\text{res1}}+\boldsymbol{y}_{\text{ASPP}}+\boldsymbol{y}_{\text{edge}}+\boldsymbol{y}_{\text{tiny}} \tag{2}$$

对特征$\boldsymbol{y}_{\text{res1}}$、$\boldsymbol{y}_{\text{ASPP}}$、$\boldsymbol{y}_{\text{edge}}$、$\boldsymbol{y}_{\text{tiny}}$都使用了1×1的卷积进行降维,使所有特征维度与低层特征$\boldsymbol{y}_{\text{res1}}$输出维度一致. 与此同时,对所有特征进行上采样,恢复到统一分辨率,再进行像素级叠加.

3.2　像素空间注意力模块(PAM)

在目前的增强特征表征能力与优化空间细节的语义分割算法中,由于边缘和小目标特征的丢失,导致小目标和边缘很难被准确分割. 为此,我们设计了一种把空间注意力和通道注意力有机结合在一起的像素空间注意力模块(PAM). 来获得具有较强语义信息的图像特征. 即通过把高层输出具有较强的语义信息的特征反馈至浅层,在PAM中高层特征指导浅层特征训练,使得浅层特征即具有更多的细节信息,又具有更多语义信息. 最终得到像素空间具有更多的语义信息,它在一定程度上解决了在模型中浅层图像特征不具有像素空间语义信息的问题. 具体原理如下.

在图2中将经过ASPP模块处理后的高层特征经过全局池化得到全局上下文信息作为浅层特征的指导信息,再经过并行avg&max轻量级池化,来加强全局类别的空间细节的注意力. 具体地说,从ASPP模块处理后的高层次特征依次经过全局池化、批量归一化(Batch Normalization,BN)和非线性变换、1×1卷积等操作生成具有全局上下文信息的特征,然后再与低层次特征相乘,获得图像的通道语义关系. 再采用avg&max并行轻量级池化加强空间注意力. 最后与高层次特征及带有全局上下文信息的通道特征融合输出. 不同于文献[28,29]中的工作,我们设计的PAM模块不仅可以处理不同大小的特征映射,还可以引导低层的特征学习更多语义信息,它输出的特征中像素空间具有较强的语义关系.

不同膨胀率ASPP输出的$\boldsymbol{y}_{\text{ASPP}}\in R^{W\times H\times C}$和主干网络的$\boldsymbol{y}_{\text{res1}}\in R^{W\times H\times C}$作为输入,$C$表示通道维数,$W\times H$表示空间分辨率,并使输入$\boldsymbol{y}_{\text{ASPP}}$与$\boldsymbol{y}_{\text{res1}}$特征分辨率一致. $\boldsymbol{y}_{\text{ASPP}}$每个通道经过全局平均池化(average pooling)和最大池化(max-pooling),然后经过两个全连接层以及多层感知结构(Multi-Layer Perception,MLP)产生通道注意力映射图. 为了减少网络参数,隐含层激活函数尺度设置为$R^{(C\times 1\times 1)/r}$. r为通道降低率,然后通过元素求和,最后合并两个输出为:

$$\begin{aligned}\boldsymbol{Y}^{C}(\boldsymbol{y}_{\text{ASPP}},\boldsymbol{y}_{\text{res1}})=\delta\big(&\text{MLP}(\text{AvgPool}(\boldsymbol{y}_{\text{ASPP}}))\\&+\text{MLP}(\text{MaxPool}(\boldsymbol{y}_{\text{ASPP}}))\big)\otimes\boldsymbol{y}_{\text{res1}}\end{aligned} \tag{3}$$

$\otimes$为外积运算. 为了获得特征图的空间注意力信息,对$\boldsymbol{Y}^{C}$再进行全局池化(avg&max)操作,得到2个维度的特征,分别为$\boldsymbol{Y}^{s}_{\text{avg}}\in R^{1\times H\times W}$和$\boldsymbol{Y}^{s}_{\text{max}}\in R^{1\times H\times W}$,然后经过合并,输入到单层感知网络(single Layer Perception,LP). 具体过程如下:

$$\boldsymbol{Y}^{S}=\text{LP}(\boldsymbol{Y}^{s}_{\text{avg}}\|\boldsymbol{Y}^{s}_{\text{max}}) \tag{4}$$

其中,符号‖表示卷积拼接操作. 最后对$\boldsymbol{y}_{\text{ASPP}}$、$\boldsymbol{Y}^{C}$和$\boldsymbol{Y}^{S}$进行特征融合,融合特征经过归一化BN输出.

$$\boldsymbol{y}_{p}=\text{BN}(\boldsymbol{y}_{\text{ASPP}}+\boldsymbol{Y}^{C}+\boldsymbol{Y}^{S}) \tag{5}$$

这里,符号+表示像素级相加.

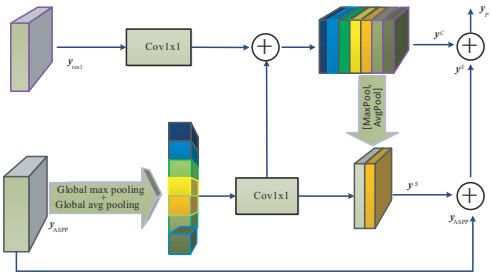


图2　像素空间注意力模块(PAM)

3.3　边缘提取模块(EEM)

为了增强网络中边缘特征和边缘语义信息,我们利用argmax对PAM模块输出的特征图$\boldsymbol{y}_{p}\in R^{W\times H\times K}$进行优化,优化后的特征输出为$[M_1,M_2,\cdots,M_k]$,然后利用梯度变换操作对优化后的特征进行处理,得到K个边缘掩膜版$[\nabla M_1,\nabla M_2,\cdots,\nabla M_k]$,经归一化和正则化处理后,与特征$\boldsymbol{y}_{p}$相乘,输出$K$个类别的边缘特征图$\boldsymbol{y}_{\text{edge}}\in R^{W\times H\times K}$,如图3所示. 由于PAM模块输出的特征具有语义关系,故得到的边缘像素含有类别信息. 由于使用了sigmoid函数对得到的边缘进行处理,本文在一定程度上缓解工作[30,31]中存在的分割边缘粗糙和稀疏的问题.

$$\left[\nabla M_1,\nabla M_2,\cdots,\nabla M_k\right]=\nabla\left[M_1,M_2,\cdots,M_k\right] \tag{6}$$

$$\nabla M=\text{dis}M+\text{dis}(1-M) \tag{7}$$

$$\boldsymbol{y}_{\text{edge}}=\boldsymbol{y}_{p}\times\delta(\nabla M) \tag{8}$$

其中δ为sigmoid函数. 同理,可以得到边缘实况图.

3.4　小目标提取模块(TTE)

如图4所示,在PAM模块经argmax优化后输出的特征M中,对每一个目标像素数num_k进行统计分析并进行排序,定义tv为目标像素数num_k与最大目标像素数num_{max}的比值,本文设置tv小于等于0.01时(可调)为特征图的小目标. 然后得到小目标掩膜版,输出的小目标掩膜版与K个通道的特征图相乘,可以得到小目标特征图$\boldsymbol{y}_{\text{tiny}}$. 由于PAM模块输出的特征中像素具有较强的语义关系,因此获得的小目标特征含有类别信息. 同理,可以得到小目标实况图.

3.5　损失函数设计

我们不仅对主干网络最后的分割特征图进行损失

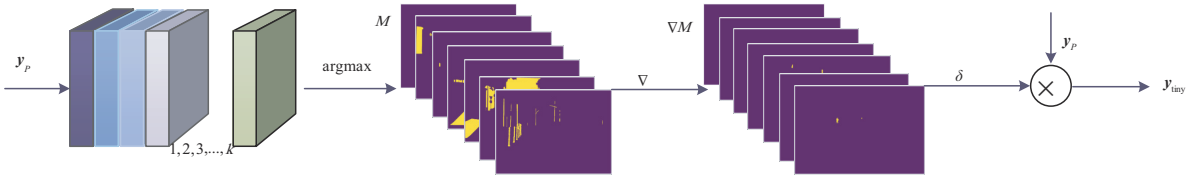


图3 边缘增强特征提取模块

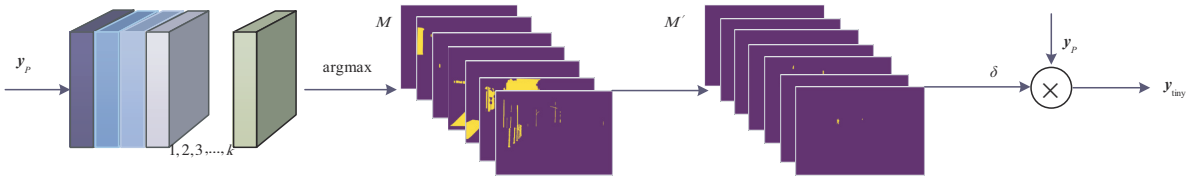


图4 小目标提取模块(TTM)

函数监督计算,且对提取的边缘和小目标特征输出进行监督计算. 为此,我们增加了边缘损失函数和小目标损失函数来监督语义边缘和语义小目标学习过程. 考虑到边缘与小目标位置像素也具有语义类别信息,为了更好地对他们进行监督,我们选择使用交叉熵损失函数对其进行监督,定义如下:

$$L_{\text{edge}}=-\frac{1}{n}\sum_{j}^{n}\left[G_j^e\ln C_\varphi\left(x_j|z\right)+\left(1-G_j^e\right)\ln\left(1-C_\varphi\left(x_j|z\right)\right)\right] \tag{9}$$

$$G_j^e=\begin{cases} y_j,\ j\text{边缘位置像素} \\ 0,\ j\text{非边缘位置像素}\end{cases} \tag{10}$$

$$L_{\text{tiny}}=-\frac{1}{n}\sum_{j}^{n}\left[G_j^t\ln C_\varphi\left(x_j|z\right)+\left(1-G_j^t\right)\ln\left(1-C_\varphi\left(x_j|z\right)\right)\right] \tag{11}$$

$$G_j^t=\begin{cases} y_j,\ j\text{小目标位置像素} \\ 0,\ j\text{非小目标位置像素}\end{cases} \tag{12}$$

其中$C_\varphi\left(x_j|z\right)$为像素$j$处的预测标签$x_j$的概率分布. y_j为实况图GT(Ground Truth map)标签,此外,主干网络分割监督损失函数表示为:

$$L_{\text{seg}}=-\frac{1}{n}\sum_{j}^{n}\left[y_j\ln C_\varphi\left(x_j|z\right)+\left(1-y_j\right)\ln\left(1-C_\varphi\left(x_j|z\right)\right)\right] \tag{13}$$

其中$C_\varphi(x_j|z)$为像素j处预测标签x_j的概率分布,y_j为GT标签. 网络建模中的总损失表示为:

$$L_{\text{total}}=\partial_1 L_{\text{seg}}+\partial_2 L_{\text{tiny}}+\partial_3 L_{\text{edge}} \tag{14}$$

其中$\partial_1,\partial_2,\partial_3$为网络超参数. 分别为分割损失、小目标损失、边缘损失的权重系数.

4 实验

首先,我们叙述了实验环境与评价标准,然后我们比较了本文算法和当前最先进的方法在Cityscapes数据集上的实验结果并进行了一系列消融实验,对结果进行了分析. 最后,又在PASCAL VOC、ADE20K和Camvid数据集上进行实验结果对比分析. 四个数据集上实验表明我们的算法不低于其他算法.

4.1 实验环境与评估标准

本实验硬件环境CPU为因特尔E5-2650V4,GPU为微星NVIDIA GeForce RTX 2080Ti. Cityscapes数据集来源于50个不同城市的街道场景,总共5 000张精细标注(精标),2 975张训练图,500张验证图和1 525张测试图. 在标注像素类别中有8个大类,每个大类中包含若干子类,共为30个小类,除去一些出现像素频率较小的类别,用19个类作为评估. 使用mIoU(mean Intersection over Union)来评估预测分割精度[32~34],其计算公式为:

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

p_{ji}为真值为j,预测结果为i的像素数,K+1是类别个数(包含背景类). p_{jj}是真实值. p_{ji}为j,被预测为i的像素数,即假正. p_{ij}则表示真实值为i,被预测为j的数量,即假负.

4.2 实验参数设置

损失函数设置:我们分别使用了多类交差熵OHEM(Online Hard Example Mining)与二进制交差熵损失函数分别对训练过程进行监督,边缘分支与小目标分支损失系数分别设置为1.

Cityscapes训练策略设置:为了进一步排除实验的偶然性,在训练过程中对所有网络进行相同设置. 优化器:为了保证训练过程中参数更新的准确率和运行时间的开销,我们选择使用SGD(Stochastic Gradient Descent)[35]作为网络训练的优化器,初始网络学习率为0.01,并采用ploy衰减策略. 训练过程中,使用4块显卡(GPU),每个GPU批尺寸设置2. 数据增强使用随机翻转,随机调整大小,随机裁剪等手段,其中随机调整大小的范围为(0.5,2.0),随机裁剪尺度为512×1 024. 此

外,验证时我们使用尺度为0.5、1.0和2.0的多尺度方案且在训练过程中未使用粗标注数据集.

PASCAL VOC、ADE20K与Camvid数据集训练策略设置:我们的训练协议参考文献[36]. 在训练过程中,我们采用多项式衰减策略,初始学习率为0.01,并使用裁剪采样作为预处理,裁剪大小512×512,批标准化参数在训练过程中进行了微调,迭代次数16万.

4.3 实验结果

为了进一步证明本文提出方法的有效性,在Cityscapes数据上我们与以下最新算法进行实验对比分析:FCN[37]、PSPNet[14]、Deeplabv3+[16]、GSCNN[21]、DSNet[38]、EAMNet[26]、PSANet[39]、DANet[40]、Maskformer[41]. 其实验结果如表1所示. 从这些分割结果可以看出,我们提出的方法在一些比较复杂的场景中能得到更好的分割效果.

从表1中可以看出,在Cityscapes验证集上,我们对Cityscape上的每一类的IoU进行了测试,每一个类别的分割性能,我们的方法几乎都略优于其他方法. 与Deeplabv3+分割结果相比,在柱子、交通灯、骑车的人、摩托车以及自行车等分割性能我们的方法分别提高2.0%、2.1%、3.9%、3.3%、1.8%. 与GSCNN比,本文算法可以在不降低其他类别(树干,摩托车等)的分割性能下,提升柱子,交通信号灯,骑车的人等小目标分割精度. 对图像中公共汽车等大目标,其精度相对FCN也有提高. 在Deeplabv3+中路面边缘我们的方法精度提升0.3%. 当我们的方法与DSNet在基线模型为Deeplabv3+,主干网络为ResNet50时,我们又进行了对比实验,DSNet分割性能只有81.5% mIoU,我们的方法是82.8% mIoU,如表1. 我们的算法着重于加强小目标与边缘的特征,而DSNet着重增强主体与边缘的特征,因此在柱子,交通灯,骑车的人,摩托车以及自行车等类别分割性能我们的方法分别提高1.3%、3.1%、2.1%、2.4%、1.7%. 在文献[40]中DSNet用8张32 GB的v-100 GPU上训练并以Wide-ResNet[42]作为主干网络可以达到83.7%的分割性能,虽然使用更深和更宽的网络可以提高分割性能,但是需要较大的计算开销.

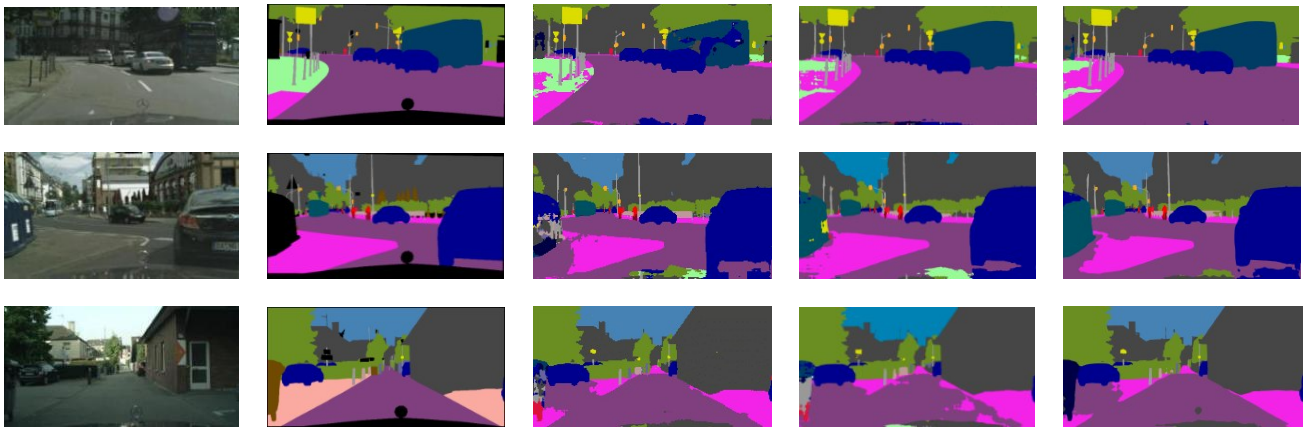
表1　在Cityscapes验证集上的各个类别分割结果

Method	IoU/%																			mIoU/%
	road	swal	build	wall	fenc	pole	tlight	sign	veg	terra	sky	person	rider	car	truch	bus	train	mobi	bike	
FCN	97.8	84.1	91.9	41.3	56.5	64.3	71.3	79.4	92.1	63.6	94.3	82.3	61.4	93.8	47.9	75.1	42.7	56.4	77.5	72.3
PSPNet	98.0	84.5	92.9	54.9	61.9	66.5	72.2	80.9	92.6	65.6	94.8	83.1	63.5	95.4	83.9	90.6	84.0	67.6	78.5	79.6
EAMANet	98.3	86.9	92.8	45.8	62.3	67.7	74.8	81.9	92.7	63.3	95.0	84.2	65.6	95.9	84.3	85.0	58.6	68.8	79.9	78.1
PSANet	98.1	84.8	92.4	46.2	58.9	66.1	73.6	80.8	92.5	64.3	94.7	83.0	62.0	95.5	78.6	88.2	80.7	67.0	79.5	78.2
DANet	98.5	87.2	93.1	55.4	61.6	67.7	74.2	80.4	92.4	64.2	95.0	83.8	64.9	95.9	**87.9**	91.4	80.9	70.1	79.5	80.2
Deeplabv3+	98.2	86.6	93.1	55.2	63.1	70.0	75.1	82.1	93.0	64.1	95.1	84.1	65.7	95.6	84.3	89.4	77.1	71.2	80.1	80.0
Gscnn	98.1	87.4	92.3	54.8	79.8	75.5	75.8	84.7	93.3	68.0	**95.7**	82.4	51.7	95.5	66.3	95.9	92.5	32.4	73.3	78.4
Maskformer	97.8	82.6	90.8	43.5	49.3	62.3	66.8	76.4	91.4	60.8	93.7	79.3	59.1	94.4	65.8	77.0	64.3	57.8	76.0	73.1
DSNet	98.3	86.5	**93.6**	**60.7**	**66.8**	70.7	73.9	81.9	931	66.1	95.2	843	67.5	95.8	86.1	92.3	85.5	72.1	80.1	81.5
Ours	**98.5**	**87.7**	93.5	54.9	65.0	**72.0**	**77.0**	**83.9**	**93.2**	**67.0**	95.5	**85.9**	**69.6**	**96.2**	86.4	**92.5**	**83.7**	**74.5**	**81.8**	**82.8**

可视化分析:从图5的特征图的可视化结果可以看出,与FCN相比,我们平滑了大目标内部纹理,所以对公交车和汽车等大型物体的分割性能有很大改进. 与Deeplabv3+相比,我们主要改进了对场景中远处的行人等小目标的分割效果. 因为ASPP模块可以很好地对上下文聚合,从而缓解内部不一致现象. 但是ASPP模块是在网络输出端得到的语义信息来聚合上下文,它的小目标及边缘等细节信息已经残缺,所以我们添加了带有矫正的边缘增强模块,一方面缓解边缘噪声,另一方面提高对部分小目标物体的分割效果. 但是小目标与边缘所占整体像素的比例很小,所以即使提高了这些细节分割效果,但是整体分割性能也不会有太大提高. 这和我们上面实验结果一致. 从上面图6和图7可视化结果中,我们方法能很好处理FCN方法中的大目标上下存在不一致的地方,如图6中的黄色框标注的地方,我们的方法缓解了大尺寸公交车内部纹理不一致. 与此同时,如图7中红色框标注的地方,我们的算法矫正了交通信号灯以及路面边缘像素类别,抑制了非边缘位置像素类别,很好地处理了Deeplabv3+中的小目标并缓解了边缘噪声.

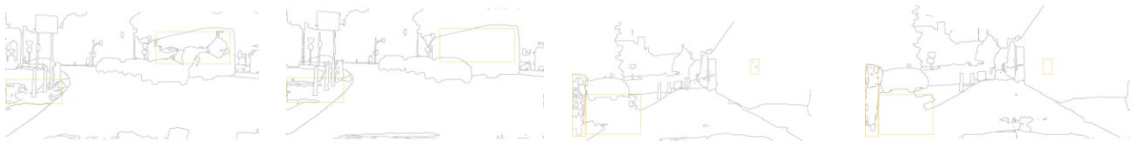
各分支可视化结果展示;为了更直观的对我们提出的模块效果进行分析,我们可视化了本文算法中各个模块输出特征,如图8中(a)为原图,(b)到(d)分别对应PAM,EEM,TTE各个模块特征图可视化结果. 可以看出,图(b)中包含了大量的空间结构信息,图(c)中含有物体轮廓信息,可以很好的增强物体边缘特征,图(d)中含交通信号灯以及远处的行人等小目标信息. 最后图(e)为融合输出特征,物体轮廓明显增强,远处物体特征也比较明显.

为了更直观的对我们提出的PAM模块进行分析,



(*a*) 原图 (*b*) 标注图 (*c*) FCN (*d*) Deeplabv3+ (*e*) 本文算法

图5 FCN、Deeplabv3+与本文分割算法可视化结果



(*a*) FCN (*b*) 本文方法 (*c*) FCN (*d*) 本文方法

图6 FCN与本文分割算法边缘可视化结果

(*a*) Deeplabv3+ (*b*) 本文方法 (*c*) Deeplabv3+ (*d*) 本文方法

图7 Deeplabv3+与本文分割算法边缘可视化结果

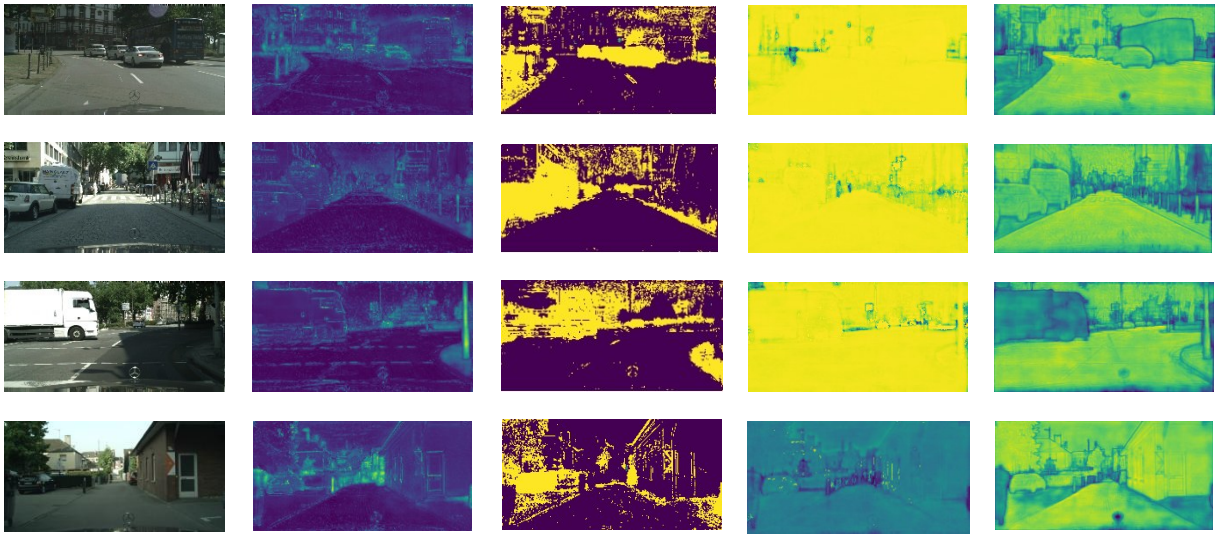
我们对PAM中高层次特征、通道特征、空间特征和最后融合输出特征进行了可视化，分别对应图9中(*b*)到(*e*)，图9(*a*)为输入图像. 我们可以看到(*b*)中包含了大量抽象的高层语义信息. 图(*c*)为高层通道相关性加到低层特征的可视化结果图，包含大量空间细节的同时又有丰富的语义信息. 图(*d*)中包含了大量空间结构信息. 图(*d*)为最后融合输出特征可视化结果.

4.4 消融实验

主干网络上的提升：我们选择应用全卷积FCN主干网络分别使用ResNet50和ResNet101作为主干网络，设计了消融实验. 如表2所示，当使用ResNet50，作为骨干网络时，原FCN的mIoU为71.4%，带有ASPP模块的FCN精度为76.6%，当嵌入我们的模型时，分割精度分别提升3.5%. 当以ResNet101为我们的骨干网时，分割精度分别提升3.3%. 基于ResNet50的模型比ResNet101的模型仅高出了0.7%，说明网络达到一定层数时，其性能的提升和网络层的深度未成正比.

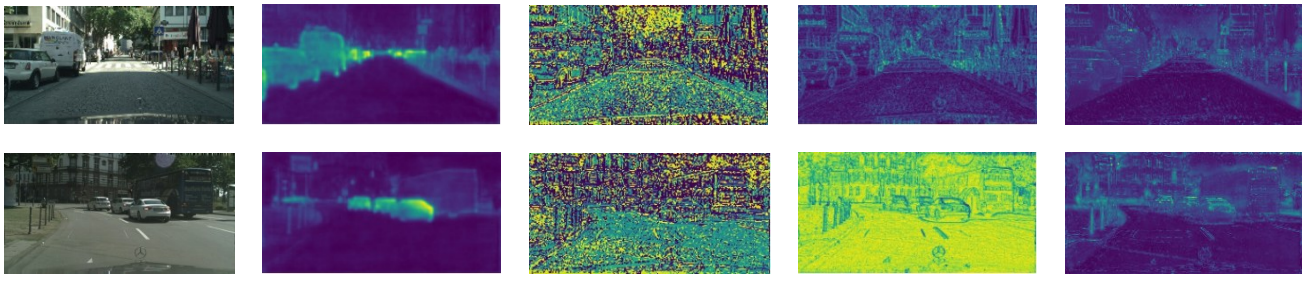
与相近方法比较：表3为我们的方法与当前最相近方法的性能比较. 我们选取了在近期工作中与我们方法最相近的四个方法包括：DCN[43]、GSCNN[23]、DSNet[38]、STLNet[44]. 上述实验结果表明，与以上前三种最相近方法相比，本文方法的增益分别为2.6%、3.0%、0.7%，我们的模块性能最优. 即使与国际最新工作STLNet相比，本文的算法性能也具有可比性.

监督消融实验：在表4中对本文方法的损失函数进行了消融实验. 如果仅用边缘损失函数对基线网络进行监督，分割性能提升0.2%，边缘增强可以去除目标边缘噪声，但是目标边缘像素占目标比例极少，所以仅对边界进行监督其分割性能提升极小. 但是对边缘和分



(a) 原图　(b) PAM模块输出特征　(c) EEM模块输出特征　(d) TTE模块输出特征　(e) 融合输出特征

图8　本文算法网络中各个模块可视化结果



(a) 原图　(b) 高层次特征　(c) 通道特征　(d) 空间特征　(e) 输出融合特征

图9　PAM中各个特征可视化结果

表2　以FCN作为基线Cityscape验证集上的消融实验

Backbone	Method	mIoU/%	Δ/%
ResNet50	FCN	71.4	—
	ASPP	76.6	—
	Ours	80.1	3.5↑
ResNet101	ASPP	77.5	—
	Ours	80.8	3.3↑

表3　与最相近方法的消融实验

Method	mIoU/%	Δ/%
FCN+ASPP	77.5	—
DCN	78.2	0.7↑
GSCNN	77.8	0.3↑
DSNet	80.1	2.6↑
STLNet	80.9	3.4↑
Ours	80.8	3.3↑

割主体同时进行监督,分别用二进制损失函数和OHEM,分割精度提升1.0%,3.5%. 说明综合损失函数能更好的挖掘基于边界形状位置的像素类别信息,且边缘与主体部分存在正交性.

表4　以FCN为基线关于损失函数监督消融实验

Method	L_{seg}	$L_{edge\text{-}bce}$	$L_{edge\text{-}ohem}$	mIoU/%	Δ/%
FCN+ASPP	√	—	—	76.6	—
	—	√	—	76.8	0.2↑
	√	√	—	77.6	1.0↑
	√	—	√	80.1	3.5↑

各部分消融实验:表5为各个模块的消融实验. 为了验证我们提出的算法对网络性能的影响,分别去掉TTE和EEM模块. 如果不使用我们提出的TTE模块,引入EEM,mIoU提高到79.0%.同时,使用TTE和EEM后,我们的mIoU分别从77.5%提高到80.8%.

4.5　其他数据实验

为了进一步验证我们提出算法的通用性,我们还在VOC2012、ADE20K和Camvid其他场景分割数据上进行了本算法实验验证. VOC的训练集有2 913张图片

表5 以FCN为基线我们方法各部分的消融实验

Method	mIoU/%	Δ/%
FCN+ASPP	77.5	—
+ EEM	79.0	1.5↑
+ TTE	79.8	2.3↑
+ EEM+TTE	80.8	3.3↑

共6 929个物体,20个类(不含背景)用来作为评估标准. 本文分别以ResNet50和ResNet101为骨干网,分割性能提高了1.2%和1.9%左右. ADE20K数据集中,训练集包含20 210张图像,测试集3 489张图像,验证集2 000张图像,其中我们用150个类别作为评估. Camvid也是城市街景数据,在该数据集中包含802张精标图像,其中选择32个语义类别作为评估. 从表6、表7和表8中可以看出,在其它几个分割数据上,本算法都有性能提升.

表6 VOC 2012数据集实验结果(输入图片大小512×512)

BackBone	Method	mIoU/%	Δ/%	#GFLOPs
ResNet50	PSPnet	73.5	—	178.4
	Deeplabv3+	74.5	—	179.5
	Ours	75.7	1.2↑	177.1
ResNet101	PSPnet	74.6	—	256.1
	Deeplabv3+	74.9	—	275.2
	Ours	76.4	1.9↑	256.8

表7 ADE20K数据集实验结果(输入图片大小512×512)

Backbone	Method	mIoU/%	Δ/%	#GFLOPs
ResNet50	PSPnet	42.7	—	179.0
	Deeplabv3+	43.9	—	198.6
	Ours	44.8	2.0↑	180.8
ResNet101	PSPnet	43.2	—	256.9
	PSAnet	43.7	—	267.6
	Deeplabv3+	45.4	—	276.3
	Maskformer	45.5	—	73.0
	Ours	45.8	2.5↑	258.4

表8 Camvid数据集实验结果(输入图片大小512×512)

Backbone	Method	mIoU/%	Δ/%	#GFLOPs
ResNet101	FCN	47.4	—	275.4
	PSPnet	48.9	—	256.1
	Deeplabv3+	67.0	—	254.1
	Ours	69.1	2.0↑	254.8

5 总结

从以上实验结果来看,与Deeplabv3+等方法对比,本文方法在一定程度上提高了对小目标图像的分割精度. 比如,从图5和图8的可视化结果来看,远处的行人细节信息有明显增加,网络输出特征中包含了大量空间细节和丰富的语义信息. 与相近方法[21,22]相比,由于本文提取地边缘及小目标具有语义类别信息,且对边缘及小目标像素类别又进行了训练校正,所以它们能与主网络图像特征更好地交互融合. 这不仅提高了小目标的分辨率,改善了对边缘的分割效果,同时也使大目标轮廓更加清晰,缓解了边缘附近的毛躁与混淆现象,提高了大目标分割精度.

本文算法与以往方法的不同之处主要存在以下三个方面. 首先,我们设计了一个新的轻量级注意力模块PAM,该模块使带有丰富细节的低层获得了高层语义信息;然后分别对该模块输出特征进行边缘与小目标建模,提取小目标及边缘特征. 最后对建模提取结果分别设置相应的损失函数进行监督训练. 由于是在网络底层PAM模块中提取得小目标及边缘特征,因此其具有丰富细节和语义类别信息. 训练后的特征与ASPP输出的特征、主干网络第一层特征融合,使得小目标特征、边缘特征、主网络图像特征三者之间进行交互. 在增强了小目标与边缘特征的同时,也矫正了图像像素的类别标签,提高了图像的分割精度.

参考文献

[1] FARABET C, COUPRIE C, NAJMAN L, et al. Learning hierarchical features for scene labeling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(8): 1915-1929.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2022-06-15]. https://arxiv.org/abs/1409.1556.

[3] Al-QIZWINI M, BARJASTEH I, Al-QASSAB H, et al. Deep learning algorithm for autonomous driving using googlenet[C]//IEEE Intelligent Vehicles Symposium. Los Angeles: IEEE, 2017: 89-96.

[4] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[5] HUANG G, LIU Z, MAATEN L VAN DER, et al. Densely connected convolutional networks[C]//Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4700-4708.

[6] LIU S, De MELLO S, GU J, et al. Learning affinity via spatial propagation networks[C]//Neural Information Processing Systems. Long Beach: MIT Press,2017: 1520-1530.

[7] 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(4): 834-848.

[8] KRAHENBUHL P, KOLTUN V. Efficient inference in fully connected crfs with Gaussian edge potentials[J]. Advances in Neural Information Processing Systems, 2011, 24: 109-117.

[9] POHLEN T, HERMANS A, MATHIAS M, et al. Full-resolution residual networks for semantic segmentation in street scenes[C]//Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4151-4160.

[10] GUO D, ZHU L, LU Y, et al. Tiny object sensitive segmentation of urban street scene with spatial adjacency between object classes[J]. IEEE Transactions on Image Processing, 2018, 28(6): 2643-2653.

[11] YANG Z, YU H, FENG M, et al. Tiny object augmentation of urban scenes for real-time semantic segmentation [J]. IEEE Transactions on Image Processing, 2020, 29: 5175-5190.

[12] CHANDRA S, KOKKINOS I. Fast exact and multi-scale inference for semantic image segmentation with deep Gaussian CRFs[C]//European Conference on Computer Vision. Netherlands: Springer, 2016: 402-418.

[13] JAMPANI V, KIEFEL M, GEHLER P V. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks[C]//Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 4452-4461.

[14] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2881-2890.

[15] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[C]//International Conference on Learning Representations. San Diego: OpenReview. net, 2015: 1-14.

[16] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//European Conference on Computer Vision. Munich: Springer, 2018: 801-818.

[17] BERTASIUS G, SHI J, TORRESANI L. Semantic segmentation with boundary neural fields[C]//Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3602-3610.

[18] CHENG D, MENG G, XIANG S, et al. Fusionnet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(12): 5769-5783.

[19] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1925-1934.

[20] PENG C, ZHANG X, YU G, et al. Large kernel matters - improve semantic segmentation by global convolutional network[C]//Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4353-4361.

[21] TAKIKAWA T, ACUNA D, JAMPANI V, et al. Gated-SCNN: Gated shape cnns for semantic segmentation[C]// International Conference on Computer Vision. South Korea: IEEE, 2019: 5229-5238.

[22] DING H, JIANG X, LIU A, et al. Boundary-aware feature propagation for scene segmentation[C]//International Conference on Computer Vision. South Korea: IEEE, 2019: 6819-6829.

[23] HOLSCHNEIDER M. A real-time algorithm for signal analysis with the help of the wavelet transform[J]. Wavelets, 1988, 1: 286-297.

[24] VAIDYANATHAN P P. Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial [J]. Proc IEEE, 1990, 78(1): 56-93.

[25] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//International Conference on Computer Vision. South Korea: IEEE, 2019: 603-612.

[26] LI X, ZHONG Z, WU J, et al. Expectation-maximization attention networks for semantic segmentation[C]//International Conference on Computer Vision. South Korea: IEEE, 2019: 9167-9176.

[27] ZHONG Z, LIN Z Q, BIDART R, et al. Squeeze-and-attention networks for semantic segmentation[C]//Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 13065-13074.

[28] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Computer Vision and Pattern Recognition. Washington: IEEE, 2018: 7132-7141.

[29] CHEN L, ZHANG H, XIAO J, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning[C]//Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5659-5667.

[30] ZHIDING Y, CHEN F, LIU M, et al. Casenet: Deep category-aware semantic edge detection[C]//Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1761-1770.

[31] ACUNA D, KAR A, Fidler S. Devil is in the edges: Learning semantic boundaries from noisy annotations[C]//

Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 11075-11083.

[32] PERAZZI F, PONT-TUSET J, McWilliams B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]//Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 724-732.

[33] 梁新宇, 林洗坤, 权冀川, 肖铠鸿. 基于深度学习的图像实例分割技术研究进展[J]. 电子学报, 2020, 48(12): 2476-2486.
LIANG X, LIN X, QUAN Y, et al. Research on the progress of image instance segmentation based on deep learning[J]. Acta Electronica Sinica, 2020, 48(12): 2476-2486. (in Chinese)

[34] 蔡超丽, 李纯纯, 黄琳, 杨铁军. ED-NAS: 基于神经网络架构搜索的陶瓷晶粒SEM图像分割方法[J]. 电子学报, 2022, 50(2): 461-469.
CAI C, LI C, HUANG L, et al. ED-NAS: Ceramic grain segmentation based on neural architecture search using SEM images[J].Acta Electronica Sinica, 2022, 50(2): 461-469. (in Chinese)

[35] MISHRA P, SARAWADEKAR K. Polynomial learning rate policy with warm restart for deep neural network[C]// IEEE Region 10 Conference. India: IEEE, 2019: 2087-2092.

[36] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. Lille: PMLR, 2015: 448-456.

[37] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431-3440.

[38] LI X, LI X, ZHANG L, et al. Improving semantic segmentation via decoupled body and edge supervision[C]// European Conference on Computer Vision. Glasgow: Springer, 2020:1-14.

[39] ZHAO H, ZHANG Y, LIU S, et al. PSANET: Point-wise spatial attention network for scene parsing[C]//European Conference on Computer Vision. Munich: Springer, 2018: 267-283.

[40] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//Computer Vision and Pattern Recognition. New York: IEEE, 2019: 3146-3154.

[41] BOWEN C, ALEX S, ALEXANDER K. Per-pixel classification is not all you need for semantic segmentation[C]// Neural Information Processing Systems. Virtual Conference: MIT, 2021:1-12.

[42] ZAGORUYKO S, KOMODAKIS N. Wide residual networks(EB/OL). (2016-03-23) [2022-06-15]. https://arxiv.org/abs/1605.07146.

[43] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]//International Conference on Computer Vision. Venice: IEEE, 2017: 764-773.

[44] ZHU L, JI D, ZHU S, et al. Learning statistical texture for semantic segmentation[C]//Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 12537-12546.

作者简介

任莎莎 女, 1992 年 3 月出生, 安徽淮北人. 现为华南理工大学软件学院在读博士研究生. 主要研究方向为信号处理、图像理解与分割等方向.
E-mail: 201910107240@mail.scut.edu.cn

刘 琼(通讯作者) 女, 1959 年 3 月出生, 云南昆明人. 现为华南理工大学软件学院教授、博士生导师. 承担、参加国家自然科学基金项目、国家863、973及地方政府项目10余项, 国内外学术刊物和会议发表论文50余篇, 中国专利5件. 研究方向为计算机网络、计算机视觉、模式识别等.
E-mail: liuqiong@scut.edu.cn