基于实景数据增强和双路径融合网络的实时街景语义分割算法

张志文,刘天歌,聂鹏举
(燕山大学信息科学与工程学院,河北秦皇岛 066000)

摘 要: 街景图像的分割在工业运用中具有十分重要的作用,但是街景图像具有种类繁多、光照多变等特点,此外,街景分割任务在追求准确性的同时要兼顾实时性,以上特点使得该任务具有很大的挑战性. 本文针对这一挑战性任务提出了一个由空间路径和细节路径组成的双路径网络(Dual-path Fusion Network,DFNet),其中细节路径利用高分辨率的输入得到丰富的边界信息,空间路径利用细节路径产生的高质量特征图获得足够多的语义信息;网络的开始嵌入了一个可训练的图像预处理模块(Image Preprocessing Module,IPM),该模块可以使光照不同的图像进入网络正式训练之前在RGB通道上具有方差和均值的一致性;经过预处理模块之后的特征图会分别输入到细节路径和空间路径;本文提出了一个条状注意力细化模块(Attention Refinement Module,ARM),并将其放到空间路径的最后,可以将通道级信息和局部条状信息有效结合起来;在网络的最后,利用图像融合模块(Feature Fusion Module,FFM)对两条路径的特征信息进行融合,得到最后的分割结果. 同时,本文还提出了一种基于小目标重组的"复制粘贴"数据增强方法,减弱了小目标样本数据不均衡的问题,同时扩充了数据集,该算法可以提升单个网络近2%的平均交并比(mIoU). 本文利用所提算法在CityScapes和CamVid数据集上进行了实验验证,对于CityScapes数据集来说,输入大小为1 024×2 048,其每秒处理帧数(FPS)和mIoU分别达到了98和70.1%;对于CamVid数据集来说,输入大小为720×960,其FPS和mIoU分别达到了208和65.7%. 与已有算法相比,本文算法的推理速度要优于最先进的实时街景语义分割算法,同时保持了较高的分割结果准确性,本文算法在街景图像语义分割速度和分割性能之间取得了良好的平衡.

关键词: 街景图像;语义分割;数据增强;深度卷积神经网络
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)07-1609-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210611

Real-Time Semantic Segmentation for Road Scene Based on Data Enhancement and Dual-Path Fusion Network

ZHANG Zhi-wen, LIU Tian-ge, NIE Peng-ju
(*School of Information Science and Engineering*, *Yanshan University*, *Qinhuangdao*, *Hebei* 066000, *China*)

Abstract: Semantic segmentation of road scene image plays a crucial role in industrial applications. However, challenges such as the great variety of target objects, high illumination variability in different scenes, and especially the increased requirement in speed and accuracy, make the segmentation of road scene images become difficult. To solve the above challenges, we propose an efficient convolutional neural network named dual-path fusion network(DFNet), consisting of spatial-path and detail-path. The spatial-path learns global information through low-resolution feature maps. Meanwhile, the detail-path can extract local details through high-resolution feature maps. DFNet starts with a trainable image preprocessing module(IPM), which is applied to unify the input images to have the same consistency of variance mean value on the RGB channel. Attention refinement module(ARM), which includes global pooling and strip pooling, is utilized in spatial-path to guide the feature learning while extracting the global features. After spatial-path and detail-path, a feature fusion module(FFM) is employed to effectively fuse the global and local detail features to achieve the final segmenting result. Besides the novel network DFNet, we propose a data augmentation strategy to enrich the training dataset and further solve the data imbalance issue of small objects. This straightforward "copy and paste" strategy can improve the performance of

收稿日期:2021-05-12;修回日期:2021-12-15;责任编辑:王天慧
基金项目:国家自然科学基金青年项目(No.61802335)

the same network by 2% in mIoU. We test our method on two public datasets, where it reaches FPS of 98 and mIoU of 70.1% on the CityScapes dataset(image size of 1 024×2 048), and FPS of 208 meantime mIoU of 65.7% on CamVid dataset (image size of 720×960). The experimental results show that our method achieves outperformance on speed as well as a competitive accuracy, compare to state-of-the-art methods. It also demonstrates that our approach can reach a good balance between speed and accuracy.

Key words: road scene image; semantic segmentation; data augmentation; deep convolutional neural network

1 引言

语义分割的目的是为每个图像像素分配密集标签[1],是计算机视觉的一项重要任务,在自动驾驶、视频监控、机器人传感等领域具有许多潜在的应用前景[2,3]. 随着全卷积神经网络(Fully Convolutional Networks,FCN)[4]的提出,以FCN为基础的网络[5~8]不断地提高着语义分割的性能. 然而在实际应用中,如何在保持高效推理速度的同时,保证分割结果的精确性,仍亟待解决.

现有高精度分割算法很难满足街景图像实时性的要求. 比如残差网络(Residual Networks,ResNet38)[9]和金字塔场景解析网络(Pyramid Scene Parsing Networks,PSPNet)[10]等方法,在Nvidia 1080Ti GPU上预测一张1 024×2 048分辨率图像需要大约1 s的时间. 图像级联网络(Image Cascade Neworks,ICNet)[11]设计了一种图像级联网络,使用级联图像输入(即低、中、高分辨率图像),其中低分辨率分支用来获取语义信息,中、高分辨率分支进行粗糙预测的恢复和细化. SwiftNet[12]提出了一种基于共享参数的分辨率金字塔来增加深度模型感受野的新方法,通过一个带有横向连接的轻量级编码器,有效提高了预测的准确性. 但上述两种方法的推理速度仍难以满足实时性要求. 其他一些方法[13~15]通过降低分辨率和减少特征通道,降低计算成本来满足实时性,但却导致了网络性能的下降.

另一方面,由于街景图像具有光照变化明显、目标遮挡严重、类别分布不均衡等特点,因此整体分割精度不高,小物体的分割尤为明显. 然而,小目标(例如交通灯、交通标志等)在街景图像分割任务中通常十分重要. 以上问题一定程度上阻碍了街景图像实时分割的可应用性,因此,对街景图像分割技术的进一步研究具有非常重要的现实意义.

为了解决以上问题,本文建立了一个快速的语义分割网络,该网络由细节路径和空间路径组成,网络还包括两个特有模块:图像预处理模块和条状注意力细化模块,同时,本文设计了基于小目标重组的数据增强算法来提升分割结果. 本文的主要贡献如下:

(1)提出了基于小目标重组的数据增强方法,通过在一张街景图像上选取合适的小目标,利用对应标签复制小目标,并在另一张图像中,根据语义信息选取合适位置粘贴复制的小目标,同时利用直方图匹配来解决小目标和背景光照不一致的问题,该算法可以提升分割结果2%的平均交并比(mIoU).

(2)设计了一种新型的双路径实时语义分割网络,分为细节路径和空间路径,与其他双路径网络不同,本文网络的两条路径不是相互独立的,而是将细节路径产生的不同分辨率的特征图作为空间路径输入的一部分,实现不同尺度信息的交流和融合,更好地获取图像中的语义信息.

(3)设计了基于不同光照程度的图像预处理模块,并把它嵌入到网络中,通过可训练参数来调控输入图像的亮度和对比度,使不同光照的街景图像在RGB通道上分别具有亮度和对比度的一致性.

(4)设计了一个条状注意力细化模块,该模块同时包含全局池化和条状池化,其中全局池化获取全局信息,条状池化增强细长形物体的特征表示,把全局信息和局部条状信息结合起来,使用该模块,可以提升结果3%的mIoU.

2 相关工作

多尺度信息:由于图像中的物体大小不同,所以多尺度信息对分割结果起到十分重要的作用. 图像分割网络Deeplab-V3[16]利用空洞卷积来控制感受野. PSPNet[10]直接采用不同大小的池化操作来获取多尺度信息. 但是上述方法依赖于计算量,不适用于实时语义分割. Inception[17~19]系列设计了4个并行的具有不同卷积核的分支结构,最后对4个通道进行组合,但同时也增加了计算量. 深层特征聚合网络(Deep Feature Aggregation Networks,DFANet)[20]提出了一个具有多次连接结构的语义分割模块来最大化多尺度感受野. 双边网络BiSeNetV2[21]提出了一种双分支分割网络,利用细节分支获取细节信息,空间分支扩大感受野,并有效融合二者. HyperSeg[22]设计了一种嵌套的U型网络用于获取多尺度语义信息. 上述3种算法只适用于低分辨率街景图像,当输入的街景图像分辨率较大时,推理速度会明显下降.

注意力机制:注意力机制可以利用高层信息来指导前馈网络,冻结激活网络(Squeeze-and-Excitation Networks,SENet)[23]对特征图作Squeeze和Excitation操作

从而对不同通道特征图加权. 条状池化网络(Strip Pooling Networks,SPNet)[24]则利用了条状注意力模块来提高捕获远程空间依赖关系和利用通道间依赖项的能力. 全局信息和局部信息对于语义分割同等重要,但上述注意力模块均未同时考虑二者.

数据增强:利用合成数据集进行数据增强在近年来得到了广泛的关注. 在文献[25,26]中,合成图像是通过从真实世界的图像中复制对象并简单地粘贴在前者中而生成的,在目标检测方面取得了好的结果. 在文献[27]中,作者利用已知的真实标签来增加小物体到城市场景中,使用类似的"复制粘贴"策略,同时指出过多或者过少的小目标都不能最大限度地提升分割结果. 但是该策略没有考虑到图像之间的光照程度不同的问题,导致粘贴的小物体与背景不匹配,引入了大量噪声.

3　本文方法

3.1　数据增强

本文通过"复制粘贴"小目标来对数据集进行增强,本文对小目标(如CityScapes数据集中的栏杆、交通灯等)均进行数据增强,从而使训练集中的图像扩充一倍,同时可以有效增加小目标在训练集中所占的比例,从而提升了网络的分割结果准确性.

基于小目标重组算法的数据增强算法如图1所示. 其中h代表小目标最低点在图像中的纵坐标,$h+10$与$h-10$是待插入区域的最高与最低点纵坐标,图1左半部分展示了直接将小目标物体插入到待插入图像中的算法流程图,图1右半部分展示了将原图与待插入图像直方图匹配后插入小目标物体的算法流程图.

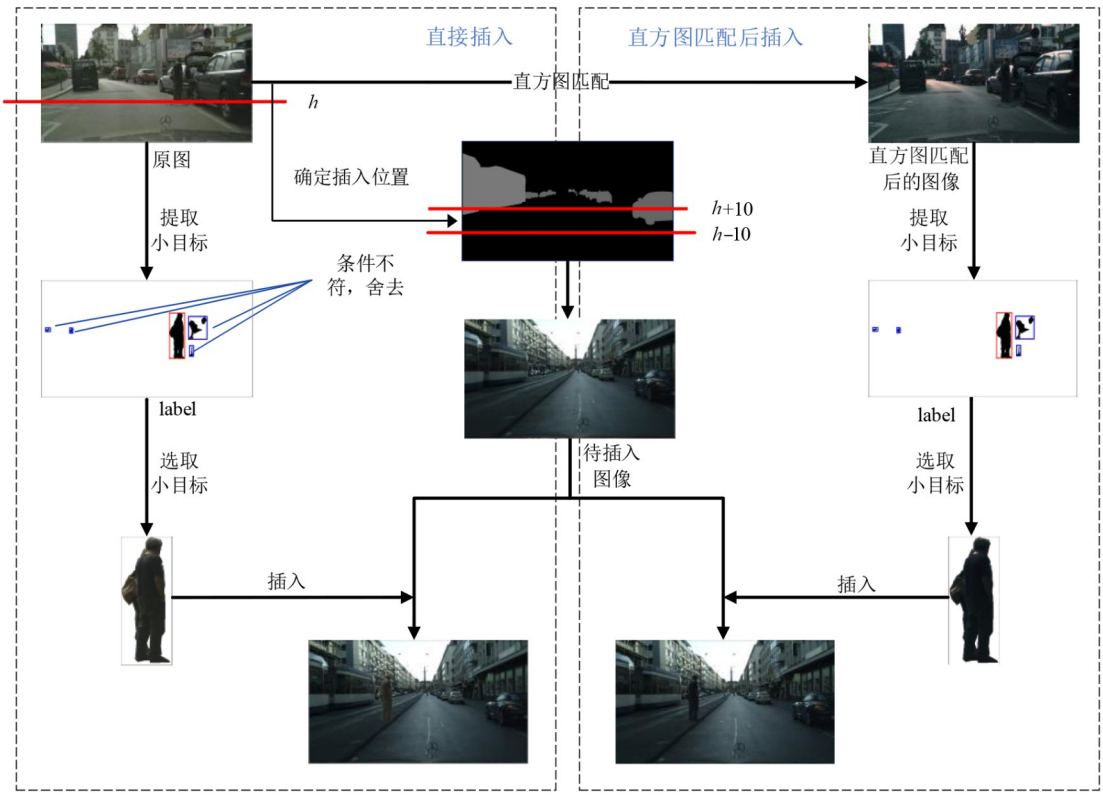


图1　基于小目标重组算法的数据增强算法示意图

本文算法将共同出现的小目标(如交通灯和电线杆等)同时提取. 同时,将小目标粘贴到合适位置,如电线杆粘贴在人行道上,确保插入后符合语义信息. 数据增强算法如算法1所示.

基于小目标重组算法的数据增强算法结果如图2所示,图中分别展示了待插入图像、小目标所在图像、直方图匹配后的小目标所在图像、仅使用复制粘贴算法所生成的结果,以及使用本文算法所生成的结果. 图中红色框标识的为插入小目标,由图2可以看到,使用本文算法与仅使用复制粘贴算法[26]相比,插入小目标后的街景图像更自然.

3.2　网络结构

本文的网络整体架构如图3所示,网络骨干部分由细节路径和空间路径组成;网络中还包含图像预处理模块,注意力细化模块和特征融合模块,其中特征融合模块为BiSeNetV2[21]所设计的特征融合模块,故之后不

算法 1 基于小目标重组的数据增强算法

输入:小目标所在图像 I_a;待插入图像 I_b.

输出:插入小目标后的图像 I_o.

BEGIN

1. Hist(I_a); //获取图像 I_a 累计直方图;
2. Hist(I_b); //获取图像 I_b 累计直方图;
3. HistMatch(I_b, I_a); //I_a 直方图匹配到 I_b;
4. 在 I_a 中选取一个小目标 a;
5. WHILE (a 尺寸、比例不符合要求),
   在 I_a 中重新选取新的小目标 a;
6. 获取小目标 a 最低点在 I_a 中的纵坐标 h;
7. 在 I_b 中生成纵坐标为 $h-10$ 至 $h+10$ 待插入区域 S;
8. 在待插入区域 S 中选取一个待插入点 p;
9. WHILE (p 不符合语义信息),
   在 S 中重新选取一个待插入点 p;
10. 将小目标 a 插入到 p 点,得到 I_o;
11. 改变 I_o 对应的标签图像;

END

做详细介绍. 本文还在空间路径每个卷积块的输出位置加入了辅助损失函数.

3.2.1 光照预处理模块

由于街景图像的光照变化明显,同时训练集有限,所以网络在实际预测时结果往往很差,其部分原因是图像预处理通常使用 z-score 归一化,但训练集与预测图像的分布有一定差异. 针对这一问题,本文设计了一个预处理模块,对图像作线性和非线性变化,优化了街景图像光照不同的问题,同时,该模块随神经网络一起,实现了端到端训练. 该模块的步骤如算法 2 所示.

通过观察图 4 街景图像的直方图发现,光照程度越强的图像在直方图上数值越大,所以预处理模块首先使用伽马变换改变图像的直方图分布,γ 代表直方图的偏移量,$\gamma<1$ 时,直方图右移;$\gamma>1$ 时,直方图左移. 本文用图像在每个通道上的均值代替 γ,经过伽马变换后,不同光照的街景图像在直方图上的分布趋于一致,但图像均值和方差仍有很大差异,因此本文引入线性变化控制图像均值和方差.

(*a*) 待插入图像 (*b*) 小目标所在图像 (*c*) 直方图匹配后图像 (*d*) 复制粘贴算法[26]结果 (*e*) 本文算法结果

图 2 基于小目标重组算法的数据增强算法结果展示

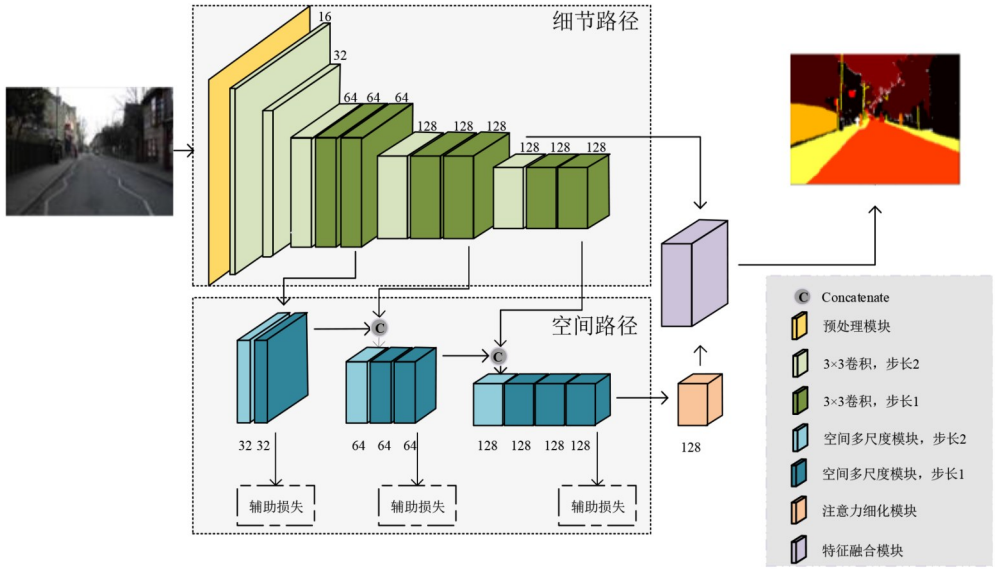


图 3 网络结构

算法2　图像预处理算法

输入:街景图像原图I_{in}.

输出:预处理后的图像RGB_{out}.

BEGIN

1. $I_{norm}=I_{in}/255$　//对I_{in}归一化,得到I_{norm};

2. $\gamma=\mathrm{mean}\left(I_{norm}\right)$　//计算I_{norm}均值,得到变换因子γ;

3. $I_{out}=I_{norm}^{\gamma}$ //对I_{norm}进行伽马变换,得到I_{out};

4. $d=\exp\left(-\mathrm{std}\left(I_{out}\right)\right)$　//计算I_{out}的方差,作用于指数函数,得到对比度因子d;

5. $l=\exp\left(-\mathrm{mean}\left(I_{out}\right)\right)$　//计算I_{out}的均值,作用于指数函数,得到亮度因子l;

6. $RGB_{in}=l\cdot I_{out}+d$　//对I_{out}进行线性变换,得到RGB_{in};

7. $RGB_{out}=\mathrm{Conv1}\times1\left(RGB_{in}\right)$　//利用1×1卷积,让RGB_{in}的三通道进行信息的交流,得到RGB_{out};

END

此外,本文引入了可训练参数α,β来控制l和d,使RGB三个通道的信息相互交流,其中α为对比度调节因子,β为亮度调节因子,如式(1)所示.

$$\begin{aligned}&\mathrm{RGB}_{out}\\&=\mathrm{RGB}_{in}\left(\sum_{i=\mathrm{R,G,B}}^{\mathrm{RGB}}\alpha_i\cdot l\cdot\mathrm{RGB}_{in}+\sum_{i=\mathrm{R,G,B}}^{\mathrm{RGB}}\beta_i\cdot d\right)\end{aligned}\tag{1}$$

图像预处理模块详细设计如图5所示.通过1×1卷积实现α,β的训练,从而对l和d起到调控作用,又可以加强RGB三通道的信息交互.不同光照的街景图像经过预处理模块后保持亮度和对比度的一致性.同时,图像预处理模块的运算量可以忽略.

3.2.2　特征融合

BiSeNetV2[21]提出的双分支分割网络,其两个分支相互独立,只通过特征融合模块融合了两个分支的最终特征.如图3所示,本文网络采用横向连接方式将细节路径的信息不断传向空间路径,并利用特征融合模块融合两条路径的信息.同时,本文在空间路径设计了空间多尺度模块,利用不同大小的卷积核获取并融合不同特征.这些设计使得网络可以更好地融合不同的特征.

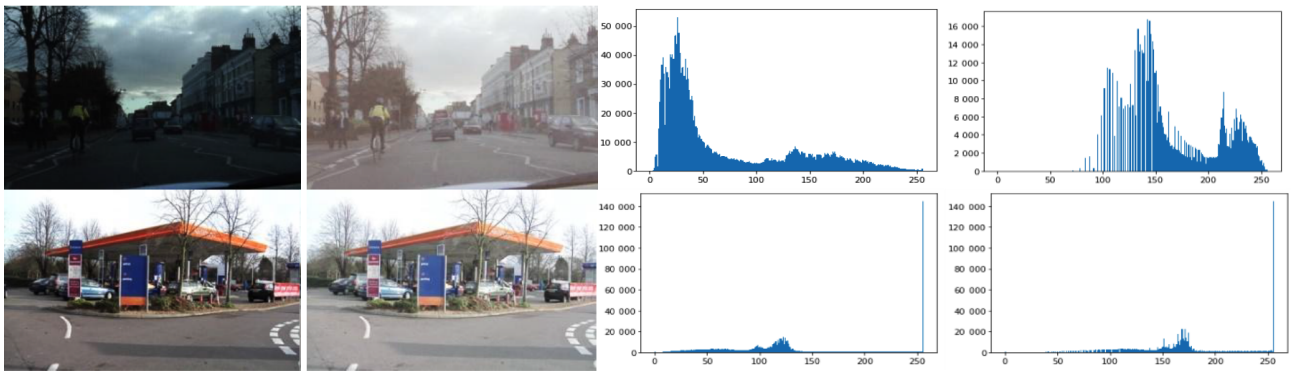
考虑到运行时间,本文的细节路径只包含3×3卷积,如表1所示,网络的细节路径包含4个阶段,其中第2~4阶段所产生的特征图将送到空间路径.本文细节路径在第一个阶段使用两个步长为2的3×3卷积降低图像分辨率,节省推理时间,同时保留一定的细节信息.细节路径的作用在于获取细节信息,以及为空间路径提供高质量特征图.

相对于细节路径,网络的空间路径设计的更为复杂,其利用细节路径提供的1/8,1/16和1/32的特征图去获取多尺度信息,以此来缓解信息丢失的问题.空间路径的详细设计如表2所示.

本文空间路径由空间多尺度模块(Space Multi-scale Module,SMM)组成,图6中详细展示了SMM模块,其中图6(*a*)和图6(*b*)是空间路径的基础模块,当步长为1时,使用图6(*a*)模块;当步长为2时,使用图6(*b*)模块.图6(*a*)和图6(*b*)中的MGConv,即图6(*c*)模块,是SMM模块的核心组成,多尺度卷积在图6(*c*)中完成,考虑到推理时间的要求,该模块中的卷积为分组卷积[28],分组数为输入特征图的数量,相当于分别在每张特征图上做卷积,而不进行通道间的信息交互.分组卷积的扩展比为6,即输出特征图是分组数的6倍.该模块首先将特征图按照1:2:1的比例分成三部分,分别用1×3、3×3和3×1卷积核进行分组卷积,之后将分组卷积得到的特征图叠加,该模块很好地获取了多尺度信息,同时,相比只使用3×3分组卷积,计算量减少了2/3.

3.2.3　注意力模块

在注意力模块中,本文利用全局池化来捕捉全局信息,同时计算一个注意力向量来对通道级的特征进行加权,指导特征学习.本文采用条形池化来获取条状信息,并将其与全局注意力进行结合,增强如栏杆等条状物体的特征表示.



(*a*)　原始图像　(*b*)　处理后的图像　(*c*)　原始图像直方图　(*d*)　处理后的直方图

图4　街景图像直方图

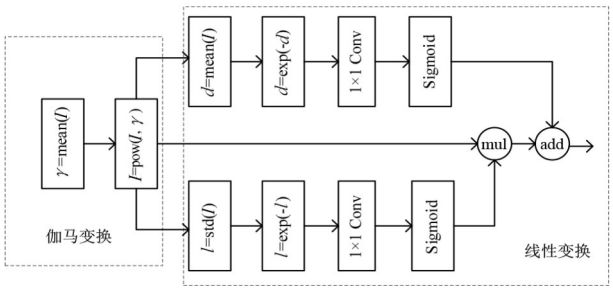


图5 光照预处理模块

表1 细节路径

阶段	细节路径			
	操作	步长	输入通道数	输出通道数
细节卷积块1	Conv 3×3	2	3	16
	Conv 3×3	2	16	32
细节卷积块2	Conv 3×3	2	32	64
	Conv 3×3	1	64	64
	Conv 3×3	1	64	64
细节卷积块3	Conv 3×3	2	64	128
	Conv 3×3	1	128	128
	Conv 3×3	1	128	128
细节卷积块4	Conv 3×3	2	128	128
	Conv 3×3	1	128	128
	Conv 3×3	1	128	128

该模块如图7所示,包含一个全局池化和两个条状池化,条状池化的大小分别为(1×W)和(H×1),其中(H, W)是输入特征图的尺寸,该模块可由式(2)表示:

$$f_{\text{out}}=f_{\text{in}}\cdot\text{sigmoid}\left(\sum_{i=1}^{N}k_i\text{mean}\left(f_{\text{in}}\right)\right) \tag{2}$$

其中,f_{in}和f_{out}代表输入和输出特征,k_i是可训练参数,对

表2 空间路径

阶段	空间路径			
	操作	步长	输入通道数	输出通道数
空间卷积块1	SMM	2	64	32
	SMM	1	32	32
C	Concat	1	160	160
空间卷积块2	SMM	2	160	64
	SMM	1	64	64
	SMM	1	64	64
C	Concat	1	192	192
空间卷积块3	SMM	2	192	128
	SMM	1	128	128
	SMM	1	128	128
	SMM	1	128	128

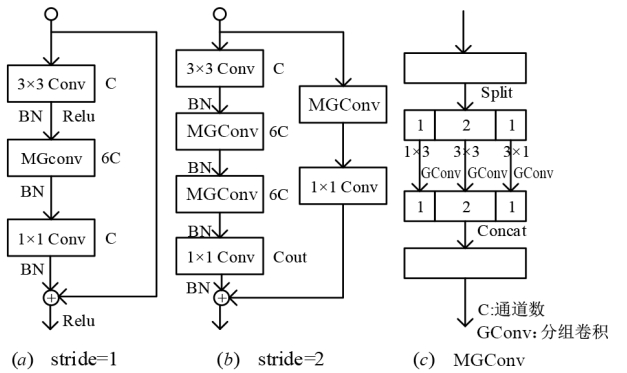


图6 空间多尺度模块(SMM)

于全局注意力来说mean()代表全局均值,而对于条状注意力来说,mean()代表条状均值,即特征图的行或者列均值. 本文提出的注意力模块实现了对通道级信息和局部条状信息的选取和组合.

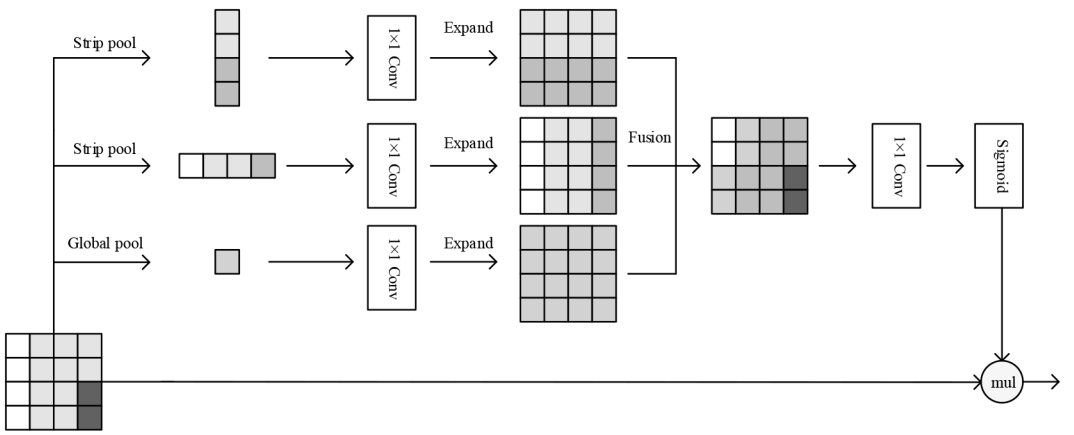


图7 注意力细化模块

3.2.4 损失函数

除了网络最终的预测损失,本文还引入多个辅助损失来监督网络训练,增强网络的特征表示,如图3所示,本文在每个空间路径卷积块之后插入辅助损失.

如式(3)所示,网络的损失函数是最终的分割损失和辅助损失的总和. 其中p和q分别是辅助损失和最后

损失的系数,本文在训练网络时,将p和q设为1.

$$\text{loss}=p\text{loss}_1+q\text{loss}_2 \tag{3}$$

其中,loss_1为分割损失,loss_2为辅助损失,二者计算方法相同,均如式(4)所示,为交叉熵损失与mIoU损失之和. 该公式的前半部分为交叉熵损失函数,解决mIoU损失的训练不稳定的问题;后半部分为mIoU损失,一定程度上解决正负样本不平衡问题.

$$\begin{aligned}\text{loss}=&-\sum_N y_{\text{true}}\log\left(y_{\text{pred}}\right)\\&+\frac{1}{N}\left(1-\frac{y_{\text{true}}\cdot y_{\text{pred}}}{\sum y_{\text{true}}+\sum y_{\text{pred}}-y_{\text{true}}\cdot y_{\text{pred}}}\right)\end{aligned} \tag{4}$$

其中,N代表类别数,辅助损失函数中,y_{true}表示对应分辨率标签,y_{pred}表示空间路径的预测概率值;最终预测损失函数中,y_{true}为真实标签,y_{pred}为网络最后的预测概率值.

4　实验结果

4.1　实现细节

本文分析了不同优化器之间的优缺点,采用文献[29]提出的方法,即随机梯度下降法(Stochastic Gradient Descent,SGD)和Adam配合使用,前期使用Adam快速收敛,后期切换到SGD,避免Adam所产生的振荡问题.

对于CityScapes[30]数据集的前40 000次迭代和CamVid[31]数据集的前20 000次迭代,本文使用Adam优化器,初始学习率为1e-3,bata_1为0.9,beta_2为0.999. 对于CityScapes数据集的最后10 000次迭代和CamVid数据集的最后5 000次迭代,使用SGD优化器,动量为0.9,权重衰减为1e-5,使用“poly”学习率策略,初始学习率设为1e-2. 对于CityScapes数据集,批处理大小为10,对于CamVid数据集,批处理大小为16. 在数据增强方面,本文使用随机水平翻转、随机缩放和随机裁剪策略,随机缩放尺度在[0.75,2]范围内. CityScapes裁剪分辨率大小为1 024×2 048,CamVid裁剪分辨率大小为720×960.

本文不采用任何可以提升分割准确率的评估技巧,比如图像翻转和多尺度测试. 对于1 024×2 048分辨率大小的图像,本文只使用一张1080Ti卡来测量推理时间,并重复1 000次迭代,以减小误差. 同时,本文采用mIoU来衡量分割精度.

4.2　模块有效性

在这一小节中,本文详细研究了框架中每个模块的影响. 在下面的实验中,使用相同的网络结构,且网络中的训练策略完全一致,分别对添加各个模块前后的网络在CityScapes[30]和CamVid[31]数据集上进行实验和评估,对于CityScapes数据集,本文用验证集进行评估.

4.2.1　数据增强

为了验证数据增强算法的有效性,本文在CityScapes数据集上做了相关实验,统计了每个类别的分割结果,实验结果如表3所示.

由表3可以看出,与不使用数据增强算法相比,使用本文数据增强算法的8类小目标的IoU全部得到了提升,同时,由于本文算法可以有效扩充训练数据,其他未使用数据增强的类别结果也得到了提升. 本文算法的mIoU比不使用数据增强的算法的mIoU提升了2%,比只“复制粘贴”算法[26]的mIoU提升了1.1%. 本文的数据增强算法同样适用于其他街景图像分割网络.

表3　数据增强实验结果

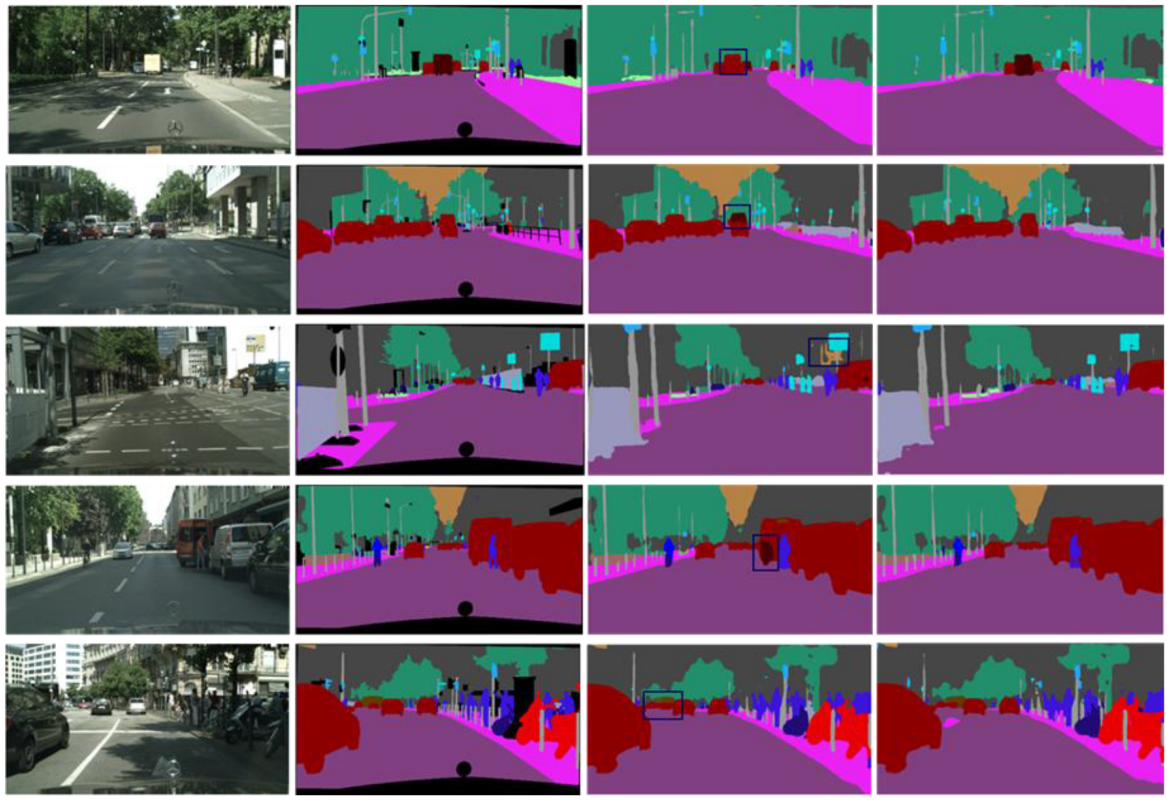
类别	小目标	IoU/%		
		原始数据	复制粘贴	本文算法
车道	否	97.3	97.4	**97.5**
人行道	否	79.1	80.0	**80.1**
建筑	否	89.0	**89.7**	89.5
墙	否	**56.6**	55.1	54.3
栅栏	否	53.4	**53.8**	52.8
支撑杆	是	47.2	**48.5**	48.0
交通灯	是	57.2	58.6	**59.6**
标志	是	65.9	67.3	**67.5**
植物	否	89.7	**90.1**	89.8
地势	否	59.3	**59.7**	56.0
天空	否	**92.6**	92.5	92.5
行人	是	71.8	72.4	**72.7**
骑行者	是	56.0	53.4	**56.5**
小汽车	否	91.5	91.7	**92.1**
卡车	否	68.7	61.9	**69.6**
公交车	否	74.6	73.1	**79.0**
列车	否	50.2	60.9	**70.0**
自行车	是	50.2	**60.2**	58.2
摩托车	是	68.5	69.2	**69.9**
mIoU/%	—	69.4	70.3	**71.4**

图8展示了CityScapes数据集的可视化分割结果,其中第一列为街景图像原图,第二列为图像标签,第三列为不使用数据增强算法所得到的结果,第四列为使用本文算法所得到的结果. 图中红框标识的为不使用数据增强算法所产生的错分区域,可以看出,错分区域得到了明显改善.

4.2.2　预处理模块

为了验证图像预处理模块的有效性,本文在CamVid数据集和CityScapes数据集做了相关实验,如表4所示.

预处理模块能够有效提升CamVid数据集分割效



(*a*) 原始图像 (*b*) 标签图像可视化 (*c*) 不使用数据增强算法结果 (*d*) 使用数据增强算法结果

图8 数据增强结果展示

表4 预处理模块结果

方法	mIoU/%	
	CityScapes	CamVid
无预处理模块	70.4	64.2
有预处理模块	**70.6**	**65.7**

果,但对于CityScapes数据集基本没有提升. 其原因在于CamVid数据集包含图像较少,且光线变化强烈. 相比之下,CityScapes数据集只包含白天的街景图像,而且训练集足够大. 因此本文只对CamVid数据集使用图像预处理模块.

4.2.3 注意力模块

为了验证本文条状注意力细化模块的作用,本文在CityScapes数据集上进行了以下实验. 实验结果如表5所示.

图9展示了3个实验在CityScapes数据集上的可视化结果,其中第一列为街景图像原图,第二列为图像标签,第三列为不使用注意力模块所得到的分割结果,第四列为使用注意力模块所得到的分割结果. 图中红框标识的为不使用数据增强算法所产生的错分区域,使用本文的注意力模块后,错分区域得到了明显改善.

表5 注意力模块结果

注意力类型	mIoU/%	
	CamVid	CityScapes
无注意力	64.2	67.8
全局注意力	64.9	69.0
条状注意力	65.2	69.9
全局+条状注意力	**65.7**	**70.4**

4.2.4 空间多尺度模块

本文在空间路径中设计了空间多尺度模块,本文在CityScapes数据集和CamVid数据集上进行了实验,统计了每秒处理帧数(FPS)和mIoU,实验结果如表6所示.

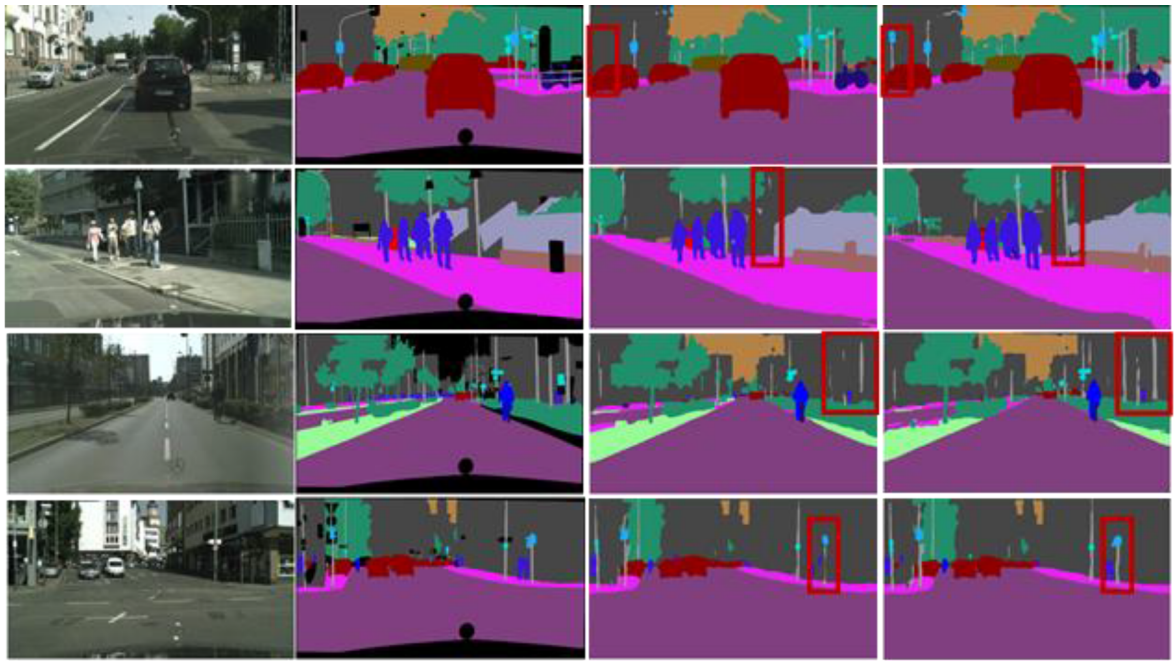
在推理速度方面,空间多尺度模块对输入图像尺寸敏感,随着输入图像尺寸增大,其对推理速度影响减小. 在推理准确度方面,空间多尺度模块具有比较稳定的提升,该模块可提升同一数据集mIoU约1%.

4.3 速度与准确度比较

在本节中,本文将上述最佳模型与2个基准数据集的其他先进的方法进行比较.

CitysSapes: CityScapes数据集的输入尺寸分为512×1 024[20~22,32~36]和1 024×2 048[4,10~12,37~43]两组,本文网络针对高分辨率图像,故主要与其余输入为1 024×2 048的模型进行对比. 如表7所示.

由表7可以看出,对于高分辨率的街景图像,本文



(*a*) 原始图像　(*b*) 标签图像可视化　(*c*) 不使用注意力细化模块结果 (*d*) 使用注意力细化模块结果

图9　注意力细化模块结果展示

表6　空间多尺度聚合模块

数据集	输入大小	卷积核	FPS	mIoU/%
CamVid	720×960	3×3	208	65.7
		1×3	157	66.5
		3×1,3×3		
CityScapes	1 024×2 048	3×3	101	70.4
		1×3	98	71.4
		3×1,3×3		

所提出方法的推理速度明显优于最先进方法,同时保持了较高的分割精度准确性. 对于CityScapes数据集,对比1 024×2 048的输入,本文方法同样达到了最快的推理速度,比其余方法中最快的SwiftNet[12]要快58 FPS,是其速度的2倍以上. 综上所述,本文的方法在分割速度上取得了现有领先水平,同时保持了高精度的分割.

CamVid: CamVid数据集的输入尺寸为720×960,本文与其他实时分割网络[11,15,20,21,31,40]进行比较,比较的指标包括FPS和mIoU. 如表8所示.

本文方法在CamVid数据集上达到了65.7%的mIoU和208 FPS,在对比算法中是唯一一个超过200 FPS的,与ENet[15]、DFANetA[20]、DFANetB[20]、RefineNet[40]、BiSeNetV1[36]相比,本文方法在mIoU和FPS上都取得了领先地位;与ICNet[11]、SwiftNet[12]等网络相比,本文方法虽然在mIoU上相对落后,但推理速度比ICNet[11]快180 FPS,比BiSeNetV1-L[36]快92 FPS,比BiSeNetV2[21]快83.5 FPS,比BiSeNetV2-L[21]快175 FPS. 在CamVid数据集上与现有方法相比,本文方法在推理速度上取得了极大的领先,同时保持了较高的分割精度. 同时本文的参数数量只有3.4 M,只比ENet[15]的参数量多.

表7　CityScapes实验结果

方法	预训练	输入大小	mIoU/%		FPS
			val	test	
ESPNet[33]	ImageNet	512×1 024	—	60.3	112.9
ESPNetV2[34]	ImageNet	512×1 024	66.4	66.2	—
ERFNet[35]	No	512×1 024	70.0	68.0	41.7
DFANet A[20]	ImageNet	512×1 024	—	70.3	**160**
GUN[32]	ImageNet	512×1 024	69.6	70.4	33.3
BisenetV2[21]	No	512×1 024	73.4	72.6	156
HyperSeg[22]	ImageNet	512×1 024	**76.2**	**75.8**	36.9
Dilation10[37]	ImageNet	1 024×2 048	68.7	67.1	0.25
LRR[39]	ImageNet	1 024×2 048	70.0	69.7	—
Deeplabv2[38]	ImageNet	1 024×2 048	71.4	70.4	—
RefineNet[40]	ImageNet	1 024×2 048	—	73.6	—
DUC[43]	ImageNet	1 024×2 048	76.7	76.1	—
PSPNet[10]	ImageNet	1 024×2 048	—	78.4	0.78
SQ[41]	ImageNet	1 024×2 048	—	59.8	16.7
ICnet[11]	ImageNet	1 024×2 048	—	69.5	30.3
DABNet[42]	No	1 024×2 048	—	70.1	27.7
SwiftNet[12]	ImageNet	1 024×2 048	**75.4**	**75.5**	39.9
Ours	No	1 024×2 048	71.4	70.1	**98**

表8 CamVid实验结果

方法	参考	预训练	参数	mIoU/%	FPS
ENet[15]	CVPR2016	ImageNet	0.4 M	51.3	61.2
DFANet B[20]	CVPR2019	ImageNet	7.8 M	59.3	160
DFANet A[20]	CVPR2019	ImageNet	4.8 M	64.7	120
ICNet[11]	ECCV2018	ImageNet	26.5 M	67.1	27.8
RefineNet[40]	CVPR2017	ImageNet	—	63.3	—
SwiftNet[12]	CVPR2019	ImageNet	11.8 M	72.8	—
BiSeNetV1[36]	ECCV2018	ImageNet	5.8 M	65.6	175
BiSeNetV1[36]	ECCV2018	ImageNet	49 M	68.7	116
BiSeNetV2[21]	arXiv2020	No	—	72.4	124.5
BiSeNetV2[21]	arXiv2020	No	—	**73.2**	32.7
Ours		No	3.4 M	65.7	208

5 结论

本文提出了一种新的双路径分割网络,极大提升了街景图像实时语义分割的性能. 本文设计了光照预处理模块来处理不同光照的图像,提升了不同光照图像的分割精度;本文设计了特殊的注意力模块,同时对全局和局部特征进行组合;本文设计了基于小目标重组的图像增强算法,有效扩充数据集,使mIoU提升了1.8%. 最后,本文验证了所提算法在CityScapes和CamVid基准数据集上的有效性. 对于输入尺寸为1 024×2 048的CityScapes数据集,FPS和mIoU分别达到了98和70.1%;对于输入尺寸为720×960的CamVid数据集,FPS和mIoU分别达到了208和65.7%. 结果显示本文方法在速度上取得极大提升的同时保持了分割精度.

参考文献

[1] 罗会兰, 张云. 基于深度网络的图像语义分割综述[J]. 电子学报, 2019, 47(10): 2211-2220.
LUO H L, ZHANG Y. A survey of image semantic segmentation based on deep network[J]. Acta Electronica Sinica, 2019, 47(10): 2211-2220. (in Chinese)

[2] 徐频捷, 陈逸杰, 李之南. 基于事件驱动的车道线识别算法研究[J]. 电子学报, 2021, 49(7): 1379-1385.
XU P J, CHEN Y J, LI Z N, et al. Research on event-driven lane recognition algorithms[J]. Acta Electronica Sinica, 2021, 49(7): 1379-1385. (in Chinese)

[3] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI: IEEE, 2012: 3354-3361.

[4] LONG J, SHELHAMER E, DARRELL T. Fully Convolutional Networks for Semantic Segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 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, 2018, 40(4):834-848.

[6] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for Image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[7] NOH, H, HONG, S, HAN B. Learning deconvolution network for semantic segmentation[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1520-1528.

[8] WU Z F, SHEN C H, HENGEL A VAN DEN. Wider or deeper: Revisiting the ResNet model for visual recognition [J]. Pattern Recognition, 2019, 90: 119-133.

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

[10] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 6230-6239.

[11] ZHAO H S, QI X J, SHEN X Y, et al. ICNet for real-time semantic segmentation on high-resolution images[C].//European Conference on Computer Vision. Munich: Springer, 2018: 418-434.

[12] ORSIC M, KRESO I, BEVANDIC P, et al. In defense of pre-trained ImageNet architectures for real-time semantic segmentation of road-driving images[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA: IEEE, 2019: 12599-12608.

[13] WU Z F, SHEN C H, HENGEL A V. Real-time semantic image segmentation via spatial sparsity[EB/OL]. (2017-10-01)[2021]. https://arxiv.org/pdf/1712.00213.pdf.

[14] 孟琭, 徐磊, 郭嘉阳. 一种基于改进的MobileNetV2网络语义分割算法[J]. 电子学报, 2020, 48(9): 1769-1776.
MENG L, XU L, GUO J Y. Semantic segmentation algorithm based on improved MobileNetV2[J]. Acta Electronica Sinica, 2020, 48(9): 1769-1776. (in Chinese)

[15] PASZKE A, CHAURASIA A, KIM S, et al. ENet: A deep neural network architecture for real-time semantic segmentation[EB/OL]. (2016-06-07) [2021]. https://arxiv.org/pdf/1606.02147.pdf.

[16] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-10-05) [2021]. https://arxiv. org/pdf/1706.05587.pdf.

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

[18] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB/OL]. (2015-03-02) [2021]. https://arxiv. org/pdf/1502.03167.pdf.

[19] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning[C]//AAAI' 17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 4278-4284.

[20] LI H C, XIONG P F, FAN H Q, et al. DFANet: Deep feature aggregation for real-time semantic segmentation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, CA: IEEE, 2019: 9514-9523.

[21] YU C Q, GAO C X, WANG J B, 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.

[22] NIRKIN Y, WOLF L, HASS NE R T. HyperSeg: Patch-wise hypernetwork for real-time semantic segmentation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, TN: IEEE, 2021: 4061-4070.

[23] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Recognition and Machine Intelligence, 2020, 42(8): 2011-2023.

[24] HOU Q B, ZHANG L, CHENG M M, et al. Strip pooling: Rethinking spatial pooling for scene parsing[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle, WA: IEEE, 2020: 4002-4011.

[25] DWIBEDI D, MISRA I, Cut HEBERT M., paste and learn: Surprisingly easy synthesis for instance detection [C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE , 2017: 1310-1319.

[26] DVORNIK N, MAIRAL J, SCHMID C. Modeling visual context is key to augmenting object detection datasets [M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 375-391.

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

[28] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[29] KESKAR N S, SOCHER R. Improving generalization performance by switching from Adam to SGD[EB/OL]. (2017-10-20)[2021]. https://arxiv.org/pdf/1712.07628.pdf.

[30] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, NV: IEEE, 2016: 3213-3223.

[31] BROSTOW G J, FAUQUEUR J, CIPOLLA R. Semantic object classes in video: A high-definition ground truth database[J]. Pattern Recognition Letters, 2009, 30(2): 88-97.

[32] MAZZINI D. Guided upsampling network for real-time semantic segmentation[EB/OL]. (2018-07-19) [2021]. https://arxiv.org/pdf/1807.07466.pdf.

[33] MEHTA S, RASTEGARI M, CASPI A, et al. ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation[C]//European Conference on Computer Vision. Munich: Springer, 2018: 3567-3578.

[34] MEHTA S, RASTEGARI M, SHAPIRO L, et al. ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA: IEEE, 2019: 9182-9192.

[35] ROMERA E, ALVAREZ 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(1): 263-272.

[36] YU C Q, WANG J B, PENG C, et al. BiSeNet: Bilateral segmentation network for real-time semantic segmentation[C]//European Conference on Computer Vision. Munich: Springer, 2018: 334-349.

[37] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2016-04-30)[2021]. https://arxiv.org/pdf/1511.07122.pdf.

[38] CHEN L C, PAPANDREOUS 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, 2018, 40(4): 834-848.

[39] GHIASI G, FOWLKES C C. Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation[C]// European Conference on Computer Vision. Amsterdam: Springer, 2016: 519-534.

[40] LIN G S, MILAN A, SHEN C H, et al. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 5168-5177.

[41] TREML M, ARJONA-MEDINA J, UNTERTHINER T, et al. Speeding up semantic segmentation for autonomous driving[C]//NIPS 2016-29th Conference on Neural Information Processing Systems. Barcelona: NIPS Workshop-mlits, 2016,123:312-331.

[42] LI G, YUN I, KIM J, et al. DABNet: depth-wise asymmetric bottleneck for real-time semantic segmentation[J]. CoRR, 2019, abs/1907.11357. https://arxiv. org/pdf/1907.11357.pdf.

[43] WANG P Q, CHEN P F, YUAN Y, et al. Understanding convolution for semantic segmentation[J]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, NV: IEEE, 2018: 1451-1460.

作者简介

张志文 男,出生于1996年. 硕士研究生. 主要研究方向为机器学习、计算机视觉.
E-mail: zhangzhiwen_ysu@yeah.net

刘天歌(通讯作者) 男,出生于1988年. 博士. 副教授. 主要研究方向为机器学习、计算机视觉.
E-mail: liutiange@ysu.edu.cn

聂鹏举 男,出生于1995年. 硕士研究生. 主要研究方向为机器学习、计算机视觉.
E-mail: nie2764@stumail.ysu.edu.cn