计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009864] [CSTR: 32024.14.csa.009864]

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

基于深度语义引导和注意力融合的实时语义分割①

赵吴涯[1,2,3], 李顺新[1,2,3]

[1](武汉科技大学 计算机科学与技术学院, 武汉 430065)
[2](武汉科技大学 大数据科学与工程研究院, 武汉 430065)
[3](智能信息处理与实时工业系统湖北省重点实验室, 武汉 430081)
通信作者: 李顺新, E-mail: lishunxin72@163.com

摘　要: 针对现阶段实时语义分割方法模型冗余度高, 计算成本高和准确率低的问题, 本文提出了一种基于深度语义引导和注意力融合的实时语义分割方法. 采用 MobileNetV3 作为主干网络, 并在此基础上引入深度双分支并行操作, 使用语义分支指导修正空间分支中的像素点, 在不额外增加参数量的情况下增强了空间分支的细节特征. 此外, 引入注意力融合模块, 使用多尺度分支并行的子结构实现即时响应计算, 并提供一种跨空间信息聚合的方法以提高分割精度. 该方法在 Cityscapes 和 CamVid 数据集上以 81.2 f/s 和 147.6 f/s 的推理速度分别达到了 75.2% 和 77.4% 的 *MIoU*, 同时参数量仅为 3.4M. 实验结果表明, 本文方法在保持较少网络参数量的同时, 更好地权衡了语义分割的精度与速度.

关键词: 实时语义分割; 轻量级; 分支并行; 语义引导; 注意力融合

引用格式: 赵吴涯,李顺新.基于深度语义引导和注意力融合的实时语义分割.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9864.html

Real-time Semantic Segmentation Based on Deep Semantic Guidance and Attention Fusion

ZHAO Wu-Ya[1,2,3], LI Shun-Xin[1,2,3]

[1](School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)
[2](Big Data Science and Engineering Research Institute, Wuhan University of Science and Technology, Wuhan 430065, China)
[3](Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430081, China)

Abstract: To address the issues of high redundancy, high computational cost, and low accuracy in current real-time semantic segmentation methods, this study proposes a novel real-time semantic segmentation approach based on deep semantic guidance and attention fusion. The proposed architecture employs MobileNetV3 as its backbone network, enhanced with an innovative deep dual-branch parallel structure. The semantic branch guides the correction of pixel points in the spatial branch, enhancing its detail features without increasing the parameter count. Additionally, an attention fusion module with multi-scale parallel branches is employed to achieve real-time computation and to improve segmentation accuracy through cross-spatial information aggregation. The method achieves inference speeds of 81.2 f/s and 147.6 f/s on the Cityscapes and CamVid datasets, respectively, with *MIoU* scores of 75.2% and 77.4%, and a parameter count of just 3.4M. Experimental results demonstrate that the proposed method effectively balances segmentation accuracy and speed while maintaining a small parameter count.

Key words: real-time semantic segmentation; lightweight; branch parallelism; semantic guidance; attention fusion

① 基金项目: 国家自然科学基金联合基金 (U1803262)
收稿时间: 2024-11-12; 修改时间: 2024-12-07; 采用时间: 2024-12-24; csa 在线出版时间: 2025-04-28

语义分割是视觉场景解析中的重要组成部分, 其目的是将对应的语义标签分配给图像中的每一个像素, 将给定的图像划分为若干个视觉上有意义的区域, 已在自动驾驶[1], 智能监控, 医学影像分析[2]和工业缺陷检测[3]等领域中得到广泛应用. 近年来, 基于深度学习的语义分割方法[4]发展迅速, 准确率也在不断提升. 但如何在不断普及的移动终端和边缘设备上同时保持精确预测和及时响应是一个值得思考的问题.

早期的语义分割大都基于 Long 等人[5]提出的全卷积网络 (FCN), 其首次将卷积神经网络应用到语义分割任务上, 实现了端到端的训练方式, 在分割准确率和效率上都显著优于传统方法[6–8]. 但其分割结果中的类别关系不匹配和边界信息不够精细等问题严重阻碍了语义分割进一步发展. 为解决上述问题, Zhao 等人[9]提出的金字塔场景解析网络 PSPNet 使用金字塔池化模块来聚合不同区域, 不同尺度的上下文信息. Chen 等人[10]提出的 DeepLabV2 网络使用空洞空间金字塔池化模块, 在不额外增加训练参数的情况下, 扩大了感受野. 同时通过丢弃主干网络 ResNet[11]中的深层下采样部分, 将特征图的分辨率维持在一个较高的水平, 并采用条件随机场后处理技术增强模型捕捉细节的能力来提高预测结果的精细度. Fu 等人[12]提出的双重注意网络 DANet, 分别对空间和通道维度的语义相互依赖关系进行建模来自适应地结合局部特征和全局特征, 从而获取更精确的分割结果.

尽管上述模型方法实现了语义分割准确率的提升, 但由于语义分割是一种密集预测任务, 对于高分辨率图像的特征提取和恢复都会消耗大量计算成本, 这对于不断普及的移动终端和边缘设备来说是相当不友好的, 也极大地限制了它们在实时场景中的应用.

近年来, 研究人员开始更加注重基础模型的效率. ShuffleNet[13]利用组卷积和通道洗牌操作来进一步减少计算量. 针对卷积层的特征映射通常包含大量冗余的问题, L-GhostNet[14]在少量普通卷积生成的特征映射上, 利用廉价的线性运算来进行通道扩充. MobileNets[15]使用深度可分离卷积代替传统卷积, 在保持良好精度的前提下无需大量内存开销. MobileNetV2[16]使用具有线性瓶颈的倒残差结构来减少特征信息丢失, 进而提高精度. MobileNetV3[17]模型不仅沿用了前两代网络硬件友好型的方法, 而且采用网络适应算法 (NAS) 自动化搜索产生参数量极小的网络, 同时在较深层引入通道注意力机制, 在准确率和推理速度上都有较好的表现.

通过不同的组合方法并结合高性能的基础模块进行拓展, 实时语义分割在处理速度上有明显的提升. 考虑到对编码器的输出进行上采样并微调细节的过程并不需要一个复杂结构, 实时网络 (ENet[18]、LinkNet[19]、LEDNet[20]) 使用非对称的编解码器架构来达到实时语义分割的速度标准 (30 f/s), 但是在场景解析和自动驾驶等实际应用上准确度较差. 最近的研究 (ICNet[21]、Lite-HRNet[22]、BiSeNet[23]、BiSeNet V2[24]) 基于多分支架构提出了许多高效率的分割模型, 能够专注于提取不同类型和不同尺度的特征. 目前大多数双分支网络都采用自主设计的特征提取层, 并且在输入高分辨率图像后直接分为细节分支和语义分支. 细节分支在丰富的通道和浅层特征中捕获空间细节, 语义分支在稀疏的通道和深层特征中捕获上下文信息, 两种类型的特征表示通过融合进一步加强. 然而, 这种设计方式致使模型在训练和推理速度方面, 受到额外分支所带来的高计算成本的限制, 进而在实际应用领域中的应用面临困难.

针对以上不足, 本文提出了一种用于实时语义分割的深度语义引导和注意力融合网络 (deep semantic guide and attention fusion network, DSGNet), 它结合了编解码器模型和双分支模型的优点, 采用 MobileNetV3 作为主干网络, 并在此基础上进行浅层特征的提取, 并在深度双分支并行分支中引入语义引导融合模块, 帮助细节分支消除噪点信息, 并引入了多尺度注意力融合方法, 高效地增强了融合后特征表示的细节.

1 DSGNet

1.1 整体结构

图 1 所示为本文提出的 DSGNet 整体结构图, 主要包括编码阶段和双分支并行阶段. 网络模型的详细参数如表 1 所示.

在编码阶段 (第 1–3 阶段), 根据特征图的下采样率划分为 3 个不同子阶段. 为在保证高精度分割结果的同时, 尽可能减少参数量, DSGNet 使用 MobileNetV3-Large 中的核心模块作为下采样单元, 即引入压缩-激励机制的倒置残差瓶颈块. 为了在下采样和细节保留之间取得平衡, 仅对原始图像下采样至 1/8 大小, 在提高模型表达能力的同时充分降低计算成本和内存占用. 同时, 保留第 3 阶段下采样后的特征图作为细节分支

的副本, 在之后的并行处理中进行特征重用.

在双分支并行阶段 (第 4、5 阶段), 将信息流分为高分辨率的细节分支和高度语义集中的语义分支, 并在信息前向传播的过程中两次使用深度语义引导模块, 让语义分支中的信息充分指导细节分支, 消除图像中的噪点信息.

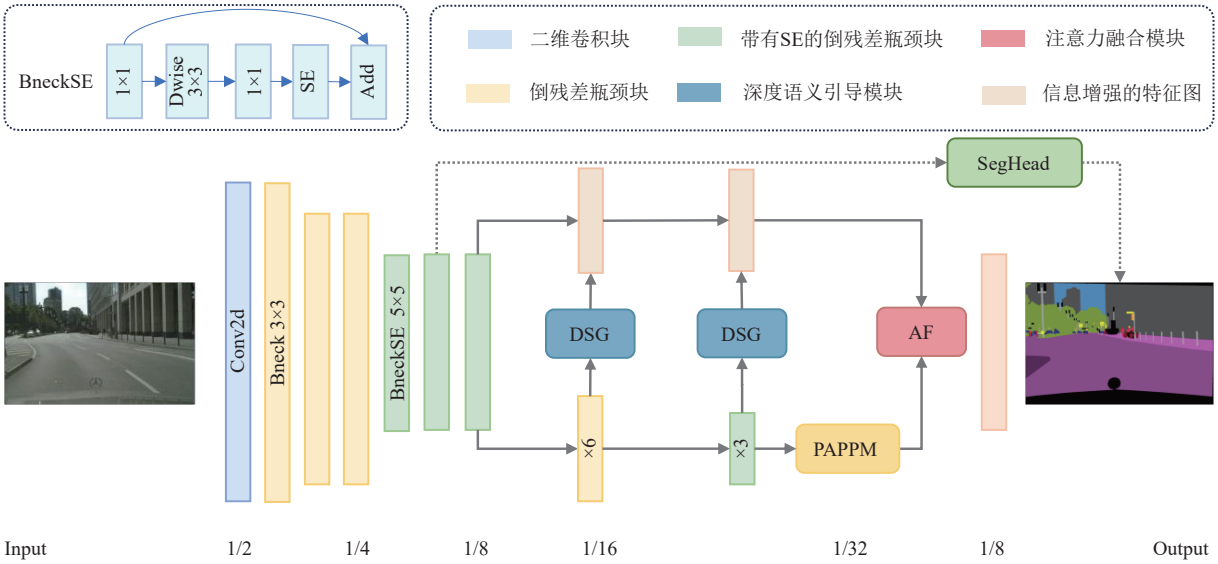


图 1 DSGNet 整体结构

表 1 DSGNet 模型参数

阶段	层号	下采样率	通道数
1	1, 2	1/2	16
2	3, 4	1/4	24
3	5–7	1/8	40
4	8–13	1/16	112
5	14–16	1/32	160

考虑到实时语义分割多应用于复杂的驾驶场景, 在同一张图像中出现的物体尺寸可能相差 10 倍以上, 需要使用跨度更大的感受野来捕获尺寸差异较大的物体, 因此, 本文引用并行聚合金字塔池化模块 (PAPPM)[25] 来高效获取不同感受野的语义信息. 由于细节分支和语义分支的特征表示具有互补性, 通过注意力融合模块实现跨维度交互, 从而捕获像素间的对应关系, 进一步聚合两个并行分支的输出特征. 为降低内存占用和减少过拟合, 在融合后直接采用 8 倍双线性插值将特征图上采样至原始输入尺寸, 最终生成预测输出图.

1.2 深度语义引导

在多分支并行处理过程中, 进行横向连接融合能增强不同尺度, 不同通道数分支之间的信息传输并提高了模型的表达能力. 早期的双向融合由于信息冗余和特点不够鲜明并不能起到很好的聚合作用. 然而, 深度语义分支保留了相对较多的通道, 专门提供丰富的上下文信息, 进而指导细节分支的小物体分割和边缘分割. 因此, 为了获取更精细的边缘细节并减少不必要的参数计算, 本文提出了深度语义引导融合模块, 只保留网络深层中高度语义集中分支到高度细节化分支的引导, 实现细节如图 2 所示.

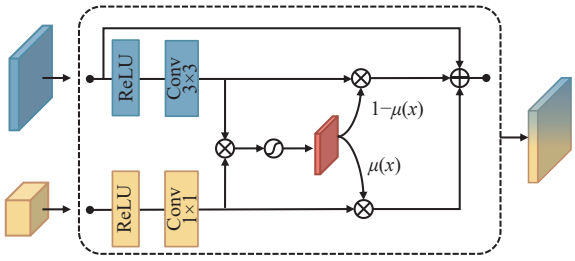


图 2 深度语义引导模块

深度语义融合模块接受两个不同尺度的输入, 语义分支 S 和细节分支 D. 根据尺度大小和信息流类别, 分别使用不同的卷积核获取相应的可训练的权重矩阵, 其中, $f_{\text{Conv1×1}}$ 和 $f_{\text{Conv3×3}}$ 分别表示应用于语义分支和细节分支的卷积操作. 随后通过 Sigmoid 函数生成门控权重, 它控制中间表示对单元整体输出的贡献. 门控权重的计算如式 (1), σ 表示 Sigmoid 函数:

$$\mu = \sigma\left[f_{\text{Conv1×1}}(S)\cdot S + f_{\text{Conv3×3}}(D)\cdot D\right] \tag{1}$$

其中, 该门控权重 μ 用于控制引导语义信息流向细节信

息贡献:

$$o=(1+\mu)\cdot D+(1-\mu)\cdot S \tag{2}$$

其中, o表示最终输出. 如果μ值较高, 模型会倾向于采纳细节分支的信息; 如果$1-\mu$值较高, 模型会更容易采纳语义分支的信息. 在指导融合过程中, 细节分支作为主分支, 语义分支作为引导分支, 只需微调细节特征图的像素分类值. 同时, 由于指导融合的阶段位于神经网络的深层, 训练过程中会出现性能退化问题. 因此, 仅在 D 分支上添加残差连接使模型的训练过程更加稳定.

1.3 注意力融合

在处理不同信息流的特征表示时, 有多种合并方式. 然而, 由于细节分支保留了丰富的空间轮廓信息, 主要包含低维信息, 而语义分支则提取了高度抽象化的特征, 属于高维信息, 简单地组合这两类特征, 往往忽略了它们在信息维度上的多样性, 难以充分发挥两类特征的互补优势, 使得模型在细节刻画和全局语义理解方面的平衡受到影响, 导致融合后模型性能下降, 并难以优化.

深度学习中的注意力机制模仿人类视觉机制, 选择性地关注重要特征区域, 主要应用在骨干网络中的特征提取阶段和中间层的特征融合阶段. 通道注意力自适应学习各通道的重要性并赋予不同权重从而强化重点特征. 坐标注意力[26]沿两个空间方向将精准信息嵌入到通道中, 并捕获空间上的长距离交互, 性能得到一定加强. 尽管如此, 它忽略了整个空间位置之间相互作用的重要性. 此外, 1×1 卷积核的有限感受野阻碍了局部跨通道交互建模和上下文信息的交互利用. 将这两种注意力机械地用于双分支融合的最后阶段, 会导致任一单边重要信息的丢失. 到目前为止, 仍然缺乏一种设备友好型的高效注意力融合模块.

因此, 本文提出了多尺度注意力融合模块, 其具体实现细节如图 3 所示. 首先, 在通道维度拼接不同分支的特征图. 然后, 为了使双边的信息流充分融合并方便后续处理, 对通道进行混洗分组操作. 对于输入特征图 $X\in R^{C\times H\times W}$, 将其沿通道维度方向划分为 G 个特征组:

$$X=[X_0,X_1,\cdots,X_{G-1}],\ X_i\in R^{C//G\times H\times W} \tag{3}$$

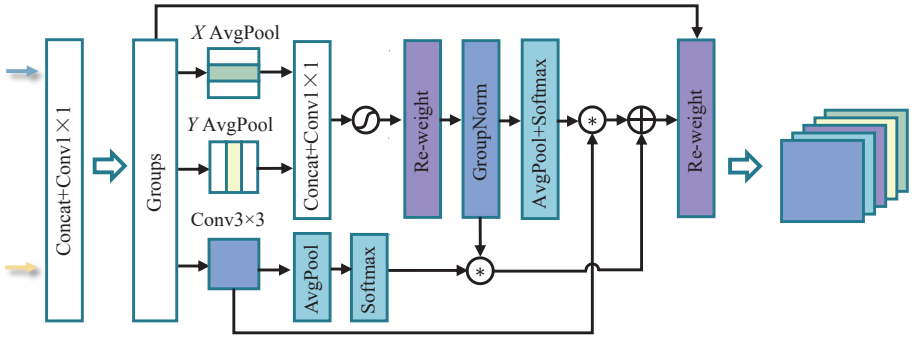


图 3 注意力融合模块

每组学习到的注意力权重将用于增强子特征中感兴趣区域的特征表示, 为上下文信息和空间信息的充分混合奠定了基础. 考虑到串行的子结构会导致网络出现大量阻塞而降低性能, 采用双分支 (B_1、B_2) 并行操作增强网络的实时性. 其中 B_1 分支分别沿水平和垂直两个方向进行一维平均池化操作. 之后使用 1×1 卷积和 Sigmoid 函数获取两个方向上的注意力权重表示 W_x 和 W_y, 并生成重新调整后的子特征图 X_1:

$$X_1=gn[X\cdot\sigma(W_x)\cdot\sigma(W_y)] \tag{4}$$

其中, gn 表示 GroupNorm 归一化, 能减少内部协变量偏移. 同时, B_2 分支使用 3×3 卷积生成空间增强后的子特征图 X_2. 接着, 将 B_1 分支和 B_2 分支的输出通过全局平均池化和 Softmax 进行处理, 生成各自不同空间尺度的注意力权重表示 Y_1、Y_2. 接着, B_1 分支和 B_2 分支分别利用对方的注意力权重来调增自身的特征图并生成最终的注意力权重表示. 最后, 通过简单的逐元素相乘增强分组后的子特征图 X_u, 并重塑回原始尺寸输出:

$$X_u=X\cdot\sigma(X_1\cdot Y_2+X_2\cdot Y_1) \tag{5}$$

在整个特征融合的过程中, 卷积核的并行化是一种更强大的结构, 可以无阻塞地处理不同分支的信息. 此外, 在并行分支中交叉融合对方注意力权重表示, 能保证融合过程的精细度, 进一步提升分割效果.

1.4 深监督

为了进一步提高模型的准确率, 本文使用了深监

督策略来增强训练过程. 深监督让中间层获得更强的监督信号, 有助于逐层优化特征表达, 使得不同层次的特征更具辨别性. 具体到图像分割任务中, 深监督能够在不同层次上学习到更细腻的边缘和轮廓信息, 从而提升模型的分割精度. 然而, 在不同分辨率下使用深监督策略的效果并不一样, 受到 BiSeNet V3[27]的启发, 本文在特征提取的第 3 阶段引入深监督策略, 并使用在线难例挖掘算法 (OHEM) 计算损失, 加速模型的收敛速度. 整体损失的计算公式如下:

$$L = L_0 + \alpha L_{\text{aux}} \tag{6}$$

其中, α表示辅助损失的权重, 经过实验对比, 当α取值为 0.4 时, 模型准确率达到最高值. L 代表模型的总损失, L_0 代表图像经过主干网络后的主损失, L_{aux}代表辅助损失.

2 实验分析

2.1 数据集

为评估本文方法的性能和泛化能力, 本文在自动驾驶领域两个知名的公开数据集 Cityscapes 和 CamVid 上进行实验. Cityscapes 数据集从驾驶者第一视角出发采集了 50 个城市不同季节的街区场景图像, 包括 5 000 张精细标注的图像和 20 000 张粗略标注的图像. 图像的分辨率高达 1024×2048, 这种全尺寸的密集输入对于实时语义分割是一个重要挑战. 在本实验中, 仅使用精细标注图像中常用的 19 个类别. 其中, 分别包含 2 975 张训练图像, 500 张验证图像和 1 525 张测试图像.

CamVid 数据集是从街道驾驶场景视频中提取的 701 张分辨率为 720×960 的图像, 其中 367 张用于训练, 101 张用于验证, 233 张用于测试. 本实验使用其所提供的 32 个类别中的 11 个类别来与其他方法进行公平比较.

2.2 评价指标

本文使用平均交并比 (mean intersection over union, *MIoU*) 和每秒处理图像的数量 (*FPS*) 以及模型总参数量来评估方法的整体效率.

MIoU 是指各个类别像素真实值与预测值的交集和并集比值的平均值, *MIoU* 越大代表分割准确率越高. 假设共有 n+1 个类别, 其中真实类别为 i 被预测为类别 j 的数量记为 p_{ij}, 预测正确的数量记为 p_{ii}, 则计算公式如下:

$$MIoU = \frac{1}{n+1}\sum_{i=0}^{n}\frac{p_{ii}}{\sum_{j=0}^{n}p_{ij}+\sum_{j=0}^{n}p_{ji}-p_{ii}} \tag{7}$$

FPS 是指在 1 s 内最多能处理图片的数量. 假设处理图片数量为 n (张), 处理时间为 t (s), 计算公式如下:

$$FPS = \frac{n}{t} \tag{8}$$

2.3 实验参数

本文所有实验均在深度学习框架 PyTorch 2.0.1 上进行. 实验主机的操作系统为 Ubuntu 20.04, 搭载了一块 RTX 3090 GPU, CUDA 版本号为 11.8.

模型训练阶段采用自适应矩估计算法 (Adam) 来更新参数, 除初始学习率外, 其他参数保持默认值. 同时使用热身策略 (WarmUp) 结合 OneCycleLR 学习率衰减策略, 使学习率在前 3 轮训练中不断快速上升到设定的最大值, 随后平滑下降到最小值. 数据增强方式包括范围为[0.5, 2.0]的随机缩放、随机裁剪、随机水平翻转和随机更改图像亮度、对比度和饱和度. 对于 Cityscapes 数据集, 初始学习率设为 0.001, 批次大小设为 12, 训练周期设为 300 轮, 裁剪分辨率设置为 1024×1024. 对于 CamVid 数据集, 初始学习率设为 0.001, 批次大小设为 4, 训练周期设为 200 轮, 裁剪分辨率设为 720×720, 并使用 Cityscapes 上的预训练权重加快训练速度, 在学习率低于 5×10^{-4} 时停止训练以避免模型的过度拟合.

2.4 结果评估

2.4.1 有效性验证

为了证明本文方法中各个模块的有效性以及整体的有效性, 清晰地观察各模块对分割结果的影响, 设计了模块之间的不同组合方式并进行消融实验. 实验在 Cityscapes 训练集上训练模型, 并在验证集上进行评估. 为了公平起见, 每次实验的超参数和随机数种子都保持一致.

实验结果如表 2 所示, 其中 Base 代表无额外分支的基准网络, 在编码后直接线性上采样 8 倍至原始尺寸. 后续方法均使用编码器和双分支并行的混合方法, Add 表示逐元素求和的融合方法; Mul 表示逐元素相乘的融合方法, DSG 表示仅使用深度语义引导模块的方法, AF 表示仅使用本文提出的注意力融合方法, AUX 表示使用深监督策略, 添加辅助损失函数的方法.

由表 2 可以看出, 加入注意力融合模块后, 尽管 *FPS* 有一定程度的下降, 但在保持参数量大体一致和满足实时语义分割时延的情况下, 网络的分割性能有显著提升. 在此基础上, 加入深度语义引导模块, 在不额外增加参数量的情况下, 进一步提升了网络的分割准确性, 证明本文提出的两种方法都是切实可行的. 在训练阶段添加辅助损失函数后, 在 *FPS* 和参数量稳定不变的情况下, 进一步提升了网络模型的精度. 相较于 Base 方法, 本文方法的 *MIoU* 提升了 5%, 而且几乎不会增加额外的参数量, 推理速度也大大超过了实时语义分割的标准.

表 2 模块消融实验结果

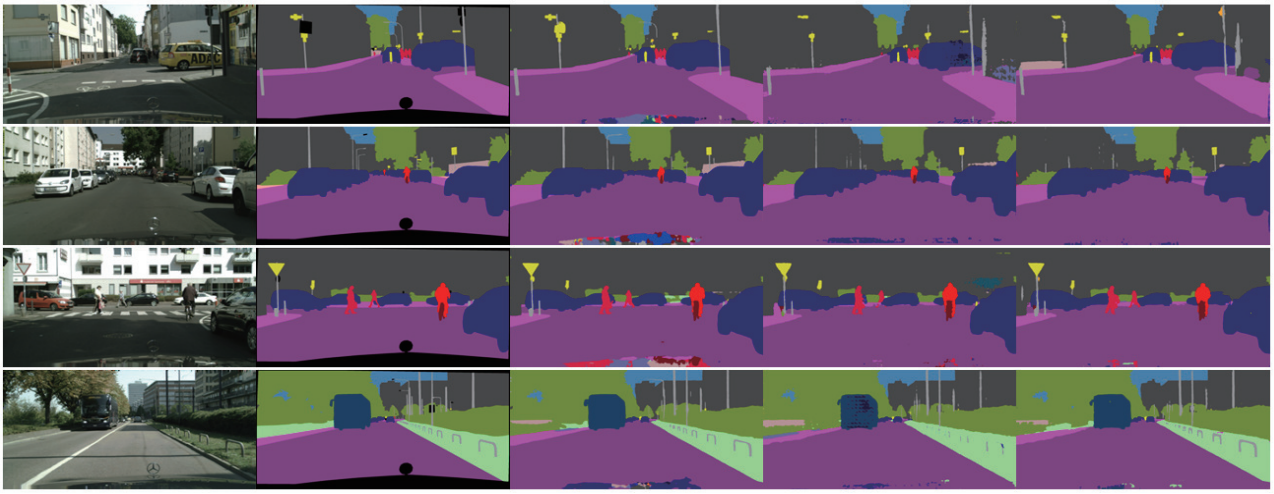
方法						*MIoU* (%)	*FPS* (f/s)	参数量 (M)
Base	Add	Mul	AF	DSG	AUX			
√	—	—	—	—	—	70.2	102.0	3.39
—	√	—	—	—	—	73.2	101.1	3.40
—	—	√	—	—	—	73.6	100.9	3.40
—	—	—	√	—	—	74.2	86.5	3.40
—	—	—	√	√	—	74.9	81.2	3.40
—	—	—	√	√	√	75.2	81.2	3.40

2.4.2 对比实验

本文对所提出的模型在 Cityscapes 数据集上与近几年的先进模型进行了公平对比, 并将部分模型所预测的分割结果进行可视化分析. 结果表明 DSGNet 在准确性、推理速度和参数量之间实现了较好的权衡. 其中, 标记为*的模型是在本实验平台下计算的推理速度. 由表 3 可知, DSGNet 获得了最高的分割精度 75.2%. 同时以全分辨率图像 (1024×2048) 作为推理输入测得的处理速度远高于表 3 中大部分的实时语义分割方法, 证实了本文方法的有效性. 其中, PP-LiteSeg-T2[28]采用专为语义分割设计的骨干网络 STDC, 主要通过短期密集连接, 融合不同感受野和多尺度信息. 在其他模块中, 也多次使用更简单的计算方式来实现不同子分支间的信息融合. 尽管 PP-LiteSeg-T2 的 *MIoU* 接近 DSGNet, 其 *FPS* 是 DSGNet 的 1.7 倍, 但其参数量却是 DSGNet 的 1.9 倍. 分析图 4 可以看出, 对于大尺寸的目标, 如公交车, DSGNet 正确地分割出其外形轮廓且无伪影信息; 对于小尺寸的目标, 如远处的路灯、交通信号灯、近处的栏杆等, DSGNet 清晰地分割出它们的细节信息. 同时从图 5 不难看出, 本文方法在参数量和预测准确率方面实现了最有效的权衡, 大大提升了实时语义分割方法在边缘设备应用的可能性.

表 3 Cityscapes 数据集分割性能对比

模型	*MIoU* (%)	*FPS* (f/s)	GPU	分辨率	参数量 (M)
ENet[18]	58.3	21.6	Titan X	1080×1920	0.4
ICNet[21]	69.5	30.3	Titan X	1024×2048	26.5
SFNet (DF1)[29]	74.5	74.0	GTX 1080Ti	1024×2048	9.0
BiSeNet[23]	74.7	65.5	Titan XP	1024×2048	49.0
DFANet A*[30]	71.3	46.9	RTX 3090	1024×2048	7.8
BiSeNet V2*[24]	72.6	81.3	RTX 3090	1024×2048	2.3
FANet*[31]	74.4	49.9	RTX 3090	1024×2048	12.3
CFPNet*[32]	70.1	51.8	RTX 3090	1024×2048	0.6
Lite-HRNet*[22]	72.8	28.1	RTX 3090	1024×2048	1.1
PP-LiteSeg-T2*[28]	74.9	142.8	RTX 3090	1024×2048	6.3
DSGNet	**75.2**	**81.2**	RTX 3090	1024×2048	3.4



(a) 原始图 (b) 真实标签 (c) 本文方法 (d) CFPNet (e) ICNet

图 4 对比实验的可视化结果

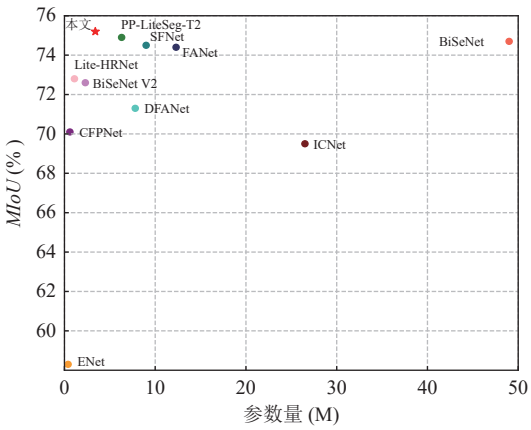


图 5 网络模型性能对比

为证明 DSGNet 的鲁棒性和泛化性, 本文在 CamVid 数据集上与其他优秀模型进行精度和推理速度的对比. 由表 4 可知, 本文方法在测试集上获得了 77.4% 平均交并比的准确度和 147.6 f/s 的推理速度, 相比于其他方法取得了较好的分割效果和较高的推理速度. 相较于传统的双分支方法 BiSeNet V2, 精度提高了 0.7% 的同时 *FPS* 也提升了 18%.

表 4 CamVid 数据集分割性能对比

模型	*MIoU* (%)	*FPS* (f/s)
DFANet A[30]	64.7	120.0
BiSeNet[23]	68.7	116.3
SFNet (DF2)[29]	70.4	134.1
BiSeNet V2[24]	76.7	124.5
PP-LiteSeg-B[28]	75.3	154.8
DSGNet	**77.4**	**147.6**

3 结论与展望

本文提出了一种基于深度语义引导和注意力融合的实时分割网络, 解决了语义分割中准确率低下, 响应速度慢和模型冗余度高的问题. 从编解码器和多分支的角度重新设计了一种深度双分支的网络架构, 以满足实时语义分割的速度要求. 在分割精度与模型参数量方面, 设计了一种新颖的语义引导融合模块, 用于修正特征图中被错误分类的像素点. 提出了注意力融合模块, 从不同的尺度充分融合信息, 在不增加额外参数量的情况下实现更精细的预测输出. 在 Cityscapes 和 CamVid 数据集上的实验结果表明, 本文方法在分割精度、推理速度和模型参数量之间实现了良好的平衡. 未来的研究将致力于进一步优化深度语义引导和注意力融合模块, 以提升模型的效率和精度, 并探索其在实际场景中的应用.

参考文献

1 王海, 李建国, 蔡英凤, 等. 基于激光雷达点云的动态驾驶场景多任务分割网络. 汽车工程, 2024, 46(9): 1608–1616.

2 文思佳, 张栋, 赵伟强, 等. 融合 CNN-Transformer 的医学图像分割网络. 计算机与数字工程, 2024, 52(8): 2452–2456. [doi: 10.3969/j.issn.1672-9722.2024.08.036]

3 董永峰, 孙松毅, 王振, 等. 融合注意力机制与联合优化的表面缺陷检测. 计算机辅助设计与图形学学报, 2024, 36(1): 102–111.

4 史文婕, 孔亚男, 刘建, 等. 深度语义分割算法综述. 工程机械, 2024, 55(10): 190–197. [doi: 10.3969/j.issn.1000-1212.2024.10.038]

5 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.

6 Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62–66. [doi: 10.1109/TSMC.1979.4310076]

7 Adams R, Bischof L. Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(6): 641–647. [doi: 10.1109/34.295913]

8 Maini R, Aggarwal H. Study and comparison of various image edge detection techniques. International Journal of Image Processing (IJIP), 2009, 3(1): 1–11.

9 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. 6230–6239.

10 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]

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

12 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. 3141–3149.

13 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An

extremely efficient convolutional neural network for mobile devices. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856.

14 Chi J, Guo SH, Zhang HP, *et al*. L-GhostNet: Extract better quality features. IEEE Access, 2023, 11: 2361–2374. [doi: 10.1109/ACCESS.2023.3234108]

15 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.

16 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520.

17 Howard A, Sandler M, Chen B, *et al*. Searching for MobileNetV3. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1314–1324.

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

19 Chaurasia A, Culurciello E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP). St. Petersburg: IEEE, 2017. 1–4.

20 Wang Y, Zhou Q, Liu J, *et al*. LEDNet: A lightweight encoder-decoder network for real-time semantic segmentation. Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP). Taipei: IEEE, 2019. 1860–1864.

21 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 (ECCV). Munich: Springer, 2018. 418–434.

22 Yu CQ, Xiao B, Gao CX, *et al*. Lite-HRNet: A lightweight high-resolution network. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 10435–10445.

23 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 (ECCV). Munich: Springer, 2018. 334–349.

24 Yu CQ, Gao CX, Wang JB, *et al*. BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision, 2021, 129(11): 3051–3068. [doi: 10.1007/s11263-021-01515-2]

25 Xu JC, Xiong ZX, Bhattacharyya SP. PIDNet: A real-time semantic segmentation network inspired by PID controllers. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 19529–19539.

26 Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13708–13717.

27 Tsai TH, Tseng YW. BiSeNet V3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. Neurocomputing, 2023, 532: 33–42. [doi: 10.1016/j.neucom.2023.02.025]

28 Peng JC, Liu Y, Tang SY, *et al*. PP-LiteSeg: A superior real-time semantic segmentation model. arXiv:2204.02681, 2022.

29 Lee J, Kim D, Ponce J, *et al*. SFNet: Learning object-aware semantic correspondence. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2273–2282.

30 Li H, Xiong P, Fan H, *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. 9522–9531.

31 Singha T, Pham DS, Krishna A. FANet: Feature aggregation network for semantic segmentation. Proceedings of the 2020 Digital Image Computing: Techniques and Applications (DICTA). Melbourne: IEEE, 2020. 1–8.

32 Lou AG, Loew M. CFPNet: Channel-wise feature pyramid for real-time semantic segmentation. Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP). Anchorage: IEEE, 2021. 1894–1898.

(校对责编: 王欣欣)