计算机科学
COMPUTER SCIENCE

面向自动驾驶的高精度实时语义分割算法架构

耿焕同，李嘉兴，蒋骏，刘振宇，范子辰

引用本文

耿焕同，李嘉兴，蒋骏，刘振宇，范子辰. 面向自动驾驶的高精度实时语义分割算法架构[J]. 计算机科学, 2024, 51(11): 174-181.

GENG Huantong, LI Jiaxing, JIANG Jun, LIU Zhenyu, FAN Zichen. High-precision Real-time Semantic Segmentation Algorithm Architecture for Autonomous Driving [J]. Computer Science, 2024, 51(11): 174-181.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000009

面向自动驾驶的高精度实时语义分割算法架构

耿焕同[1,2,3]　李嘉兴[1]　蒋　骏[1]　刘振宇[1]　范子辰[4]
1 南京信息工程大学计算机学院　南京 210044
2 中国气象局雷达气象重点开放实验室　南京 210044
3 江苏开放大学信息工程学院　南京 210036
4 南京信息工程大学软件学院　南京 210044

摘　要　PID(Proportion Integration Differentiation)语义分割架构缓解了双边架构中细节特征容易被周围的上下文信息淹没的问题(超调),同时取得了优越的性能。然而,该架构中高分辨率的边界分支严重影响了推理速度。针对此问题,提出了基于空间注意力机制和轻量辅助语义分支构建的高效 PID 架构。其中,轻量注意力融合模块用于提取精确的上下文信息并指导不同特征信息的融合,快速聚合金字塔池化模块能够快速聚合多种尺度的语义信息,并设计了一种结合 Canny 边缘检测算子的深监督训练策略以增强训练效果。与基线相比,所提模型以较小的时延代价换取了 6%的精度提升,并且在 Cityscapes,CamVid 和 KITTI 数据集上取得了准确性和速度的良好平衡,精度超越了现有同一速度区间的模型。其中,所提模型在 Cityscapes 测试集上以 120.9 frames/s 的帧率达到了 78.5%的精度。

关键词:实时语义分割;自动驾驶;超调;空间注意力机制;边缘检测
中图分类号　TP391

High-precision Real-time Semantic Segmentation Algorithm Architecture for Autonomous Driving

GENG Huantong[1,2,3],LI Jiaxing[1],JIANG Jun[1],LIU Zhenyu[1] and FAN Zichen[4]
1 School of Computer Science,Nanjing University of Information Science & Technology,Nanjing 210044,China
2 China Meteorological Administration Radar Meteorology Key Laboratory,Nanjing 210044,China
3 School of Information Technology,Jiangsu Open University,Nanjing 210036,China
4 School of Software,Nanjing University of Information Science & Technology,Nanjing 210044,China

Abstract　The proportional integration differentiation(PID) semantic segmentation architecture mitigates the problem of overshooting in the dual-branch architecture,where fine-grained features are easily overwhelmed by surrounding contextual information. However,the high-resolution boundary branch in this architecture significantly impacts the inference speed. To address this issue,an efficient PID architecture based on spatial attention mechanisms and a lightweight auxiliary semantic branch is proposed. The designed lightweight attention fusion module is used to extract precise contextual information and guide the fusion of various feature information. Additionally,a fast aggregation pyramid pooling module is introduced to rapidly aggregate semantic information across multiple scales. Finally,a deep supervision training strategy,combined with the canny edge detection operator,is designed to enhance the training effectiveness. In comparison to the baseline,the proposed model achieves a 6% increase in accuracy at the cost of a slightly increased latency. It strikes a good balance between accuracy and speed on the Cityscapes,CamVid,and KITTI datasets,outperforming existing models in the same speed range. Notably,the model achieves an accuracy of 78.5% at 120.9 frames/s on the Cityscapes test set.

Keywords　Real-time semantic segmentation,Autonomous driving,Overshoot,Spatial attention mechanism,Edge detection

1　引言

语义分割是计算机视觉领域中的一项基础任务,旨在将输入图像中的每个像素分配给特定的类别标签,实现对视觉场景的解析。随着智能需求的增加,语义分割已成为多个应用领域中的重要组件,在自动驾驶[1]、医学影像诊断[2]和遥感图像[3]等应用中起着至关重要的作用。自从 2015 年 Long 等[4]提出全卷积网络(Fully Convolutional Networks,FCN)

到稿日期:2023-10-07　返修日期:2024-04-17
基金项目:国家自然科学基金(42375145);中国气象局雷达气象重点开放实验室(2023LRM-A02)
This work was supported by the National Natural Science Foundation of China(42375145) and Open Grants of China Meteorological Administration Radar Meteorology Key Laboratory(2023LRM-A02).
通信作者:耿焕同(htgeng@nuist.edu.cn)

后,深度卷积神经网络逐渐在语义分割领域占据主导地位,随之出现了许多极具代表性的模型。为了追求更好的性能,这些模型引入了各种策略以获得更强的语义表征能力和更大的感受野。然而模型参数量和复杂性的增加,导致模型的推理速度大大下降,这限制了它们在实时应用场景尤其是自动驾驶中的广泛应用。

为了满足实时需求,设计速度更快且精度良好的实时语义分割网络成为了广受关注的研究方向。过去几年研究人员提出了许多高效的语义分割模型[5-7]。例如,Wang 等[6]提出了基于非对称卷积构建的 LEDNet,大大降低了模型的参数量和计算量;Li 等[7]提出的 SFNet 采用光流法指导网络不同阶段的特征融合。尽管这些方法展现出了一定的潜力,但始终没能做到精度和速度的良好平衡。近年来,许多文章提出了各种基于双边网络(Two-Branch Networks,TBN)的实时分割架构,通过细节分支和语义分支分别捕获细节信息和上下文信息,在速度和准确性之间取得了良好的平衡。

然而 TBN 架构存在细节特征容易被周围的上下文信息淹没的问题。最新的研究中,Xu 等[8]从 PID 控制器的角度审视了双边架构,分析了此问题并称其为超调(Overshoot)。他们提出三分支的 PID 架构,增设边界分支以指导特征融合,缓解了该问题,并取得了优越的性能。然而边界分支同细节分支一样需要保留高分辨率特征信息,导致推理速度大幅下降。

基于上述内容,本文提出了一种更为高效的 PID 架构,它基于轻量的辅助语义分支和空间注意力机制而不是高时延的边界分支。本文的主要贡献如下:

1)提出精确语义网络(Precisely Se-mantic Network,PSNet),其包括保留高分辨率特征图中的细节信息的细节分支(P)、负责聚合局部和全局的上下文信息以捕获远距离依赖的语义分支(I),以及为为语义分支提供低分辨率特征图中过渡语义信息的辅助语义分支(D)。模型以更少的推理速度代价缓解了超调现象,实现了精度和速度的最优平衡。

2)提出基于空间注意力机制设计的轻量注意力融合模块(Lightweight Attention Fusion Module,LAFM)来提取更精确的上下文信息并指导不同特征信息的融合。

3)提出能够快速聚合多种尺度信息的快速聚合金字塔池化模块(Fast Aggregation Py-ramid Pooling Module,FAPPM),提升了模型的多尺度表征能力。

4)结合 Canny[9]边缘检测算法设计了有效的深监督训练策略。

2 相关工作及问题分析

2.1 轻量编码器-解码器结构

Emara 等[10]将残差连接和深度可分离卷积作为骨干网络设计了 LiteSeg,并引入了空洞空间池化金字塔模块来增强多尺度表征。Nirkin 等[11]提出了一种由嵌套 UNet[12]结构组成的新型超网络 HyperSeg,用于提取更高级的上下文特征信息,该网络在保持实时性能的同时实现了高精度。Fan 等[13]提出了短期密集连接模块并设计了专用于分割任务的骨干网络 STDCSeg,其能够提取浅层细节特征,同时降低了计算成本。在此基础上,Peng 等[14]提出的 PP-LiteSeg 包括一个轻量级解码器和一个统一注意力融合模块,能更好地利用浅层细节特征,增强模型以更少的计算成本提取细节信息的能力。

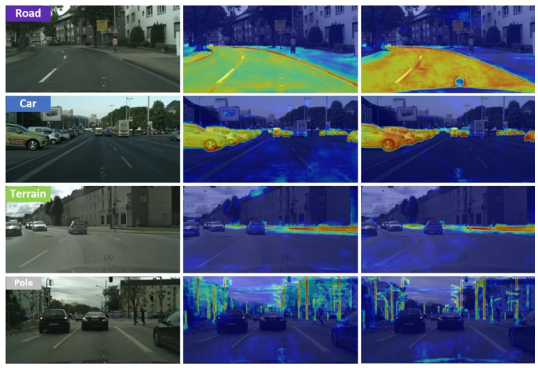
2.2 双边网络架构

由于细节信息和语义信息在语义分割中都至关重要,因此 Yu 等[15]结合了这两大因素,提出了首个双边网络(TBN)BiseNet,使用两个不同的分支分别解析细节信息和语义信息,细节分支通常保留高分辨率,而语义分支通过多次下采样步骤提取高级语义信息。在此基础上,Yu 等[16]提出了 BiseNetV2 算法,增设了用于融合两个分支之间信息的注意力特征融合模块,并引入了一种新颖的深度监督训练策略。Hong 等[17]提出的 DDRNet 算法引入了双向融合来增强两个分支之间的信息交互,他们还设计了深度聚合金字塔池化模块(Deep Aggregation Pyramid Pooling Module,DAPPM)来增强模型的全局建模能力。

2.3 PID 架构及问题分析

双边网络架构在分割任务中潜力巨大,因此成为了近几年广受关注的实时语义分割架构。然而 TBN 架构存在细节特征容易被周围的上下文信息淹没的问题(超调),这限制了其性能的进一步提升。Xu 等针对此问题进行了分析并提出了 PIDNet 算法——一个结合 PID 控制器思想的 PID 架构,其中比例分支(P)负责解析和保留高分辨率特征图中的细节信息;积分分支(I)负责聚合局部和全局的上下文信息以捕获远距离依赖;微分分支(D)负责提取高频特征以预测边界区域。该算法通过边界分支约束上下文信息从而缓解了 TBN 架构存在的超调现象,并成为了实时语义分割领域最先进的方法。

然而边界分支同细节分支一样需要保留高分辨率特征信息,在缓解超调问题的同时也严重制约了速度,因此本文旨在构建一个新颖且高效的 PID 架构以解决此问题。本文认为引起超调现象的主要原因之一在于不精确的上下文解析,如图 1 所示,从部分类别的特征激活情况可以看出,TBN 架构容易将邻近的相似类别混淆,从而引发超调现象。因此,本文以提升上下文语义解析精确度为出发点,使用简单的空间注意力机制和轻量的辅助语义分支代替边界分支,以构建更高效的 PID 三分支架构。



(a)原图 (b)DDRNet (c)Ours

图 1 热力图

Fig. 1 Heat map

3 精确语义的高效 PID 架构方法

3.1 精确语义网络(PSNet)

现有的 PID 架构大多使用边界分支指导特征融合,本质是使低分辨率的上下文信息更精确,从而减少上下文信息对细节特征的侵蚀(超调)。然而高分辨率的边界分支严重制约了模型推理速度,因此本文模型利用轻量的辅助语义分支取代边界分支,并结合空间注意力机制来提升低分辨率特征图的上下文信息精确度,构建了更为高效的 PID 架构——精确语义网络(PSNet)。

本文提出的 PSNet 网络结构如图 2 所示,其中细节分支(P)负责保留高分辨率特征图中的细节信息,语义分支(I)负责聚合局部和全局的上下文信息以捕获远距离依赖,辅助语义分支(D)为语义分支(I)提供低分辨率特征图中的过渡语义信息。具体来说,给定输入图像,网络在 Stem 阶段将其分辨率快速降低到初始分辨率的 1/4 以保证模型的轻量性。从 Stage-2 开始分为 3 个并行的分支,其中细节分支 P 始终保持 1/8 的高分辨率,I 分支每经过一个阶段都会进行下采样 2 倍和通道数量翻倍处理。为了使细节分支 P 能够充分获得多尺度的高级语义信息,Stage-2 至 Stage-4 每个阶段语义分支 I 和辅助语义分支 D 会通过上采样到 1/8 分辨率,对齐通道后通过 LAFM 模块融合至细节分支中,LAFM 模块的输出特征会作为下一阶段细节分支(P)的输入。Stage-5 阶段利用 FAPPM 模块获取多尺度语义信息以及全局感受野,并完成三分支的最终融合。架构末端使用一个简单的分割头(主要由两个 3×3 卷积构成)将细节分支产生的 1/8 分辨率特征的通道数减少到类别数,最后上采样到初始分辨率大小,从而完成最终的像素级预测。辅助语义分支由 Aux 模块构成,考虑到 GPU 架构的计算特点,采用了 ResNet18[18] 中高效的残差块和瓶颈块。

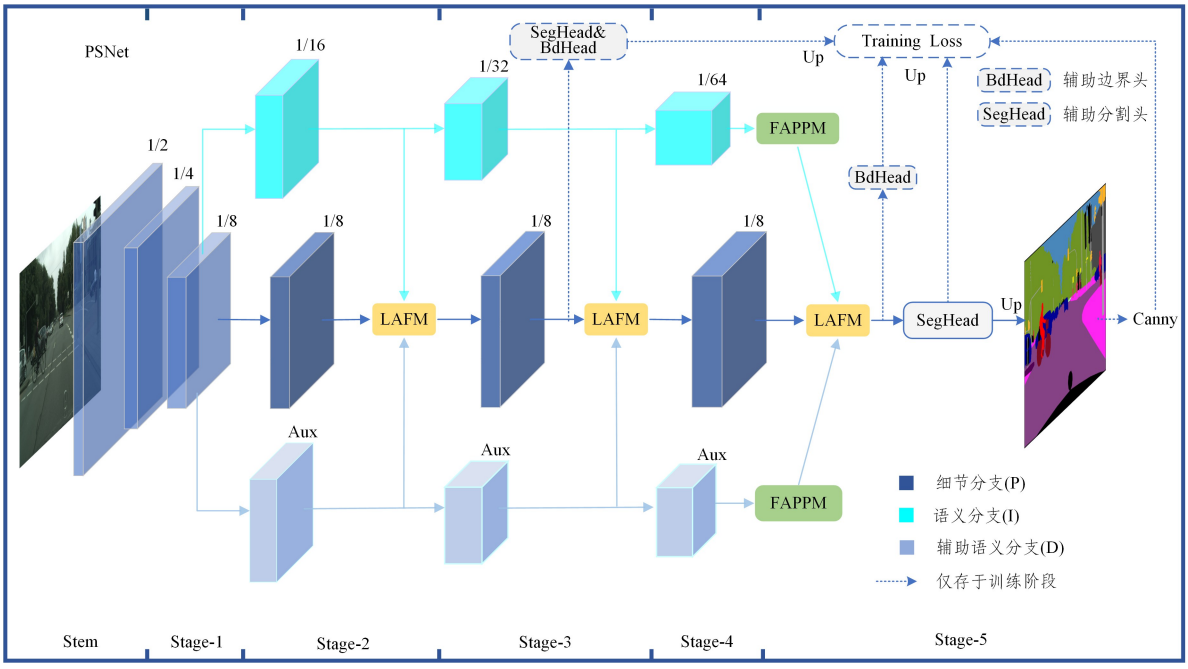


图 2 精确语义网络

Fig. 2 Network structure of PSNet

3.2 轻量注意力融合模块(LAFM)

语义分割中特征融合是必不可少的。除了逐元素求和及通道聚合外,研究人员基于不同角度提出了一些方法。Li 等[19] 提出的 DFANet 算法通过深度特征聚合结构充分利用网络的高级语义特征信息。Song 等[20] 提出的 AttaNet 算法引入了注意力机制以集中关注特征图中更相关的区域。Peng 等提出了一种统一的注意力融合模块(Union Attention Fusion Module,UAFM),它应用通道注意力和空间注意力来指导特征融合。

空间注意力机制可以使模型更关注于关键的区域,而对关键部位周围或靠近边缘的区域通常不敏感,本文将这部分区域的语义信息定义为过渡语义信息。低分辨率特征图中的过渡语义信息能够使语义分支(I)生成的低分辨率上下文信息更为精确,从而缓解超调现象,本文模型利用增设的辅助语义分支(D)来捕获这类特征信息。此外,本文基于 UAFM 的思想并结合本文架构特点设计了轻量注意力融合模块。

如图 3 所示,来自 I 分支和 D 分支上采样后的特征输入沿通道维度执行均值(Mean)和最大值(Max)操作,对生成的 4 个特征图进行拼接,卷积融合后,使用 Sigmoid 函数输出权重 $\alpha \in R^{1\times H\times W}$。权重 α 使 I 分支倾向于解析关键区域的上下文信息,D 分支倾向于解析过渡语义信息,加权求和后生成更精确的语义信息。融合后将生成的精确语义特征图乘以一个比例因子 *rate* 并注入 P 分支中,其中参数 *rate* 用来控制上下文特征信息与细节特征信息的权重比,本文统一设置为 0.5。上述计算过程分别如式(1)、式(2)、式(3)所示:

$$F_{cat}=Concat(\mathrm{Mean}(I),\mathrm{Max}(I),\mathrm{Mean}(D),\mathrm{Max}(D)) \tag{1}$$

$$\alpha=\mathrm{Sigmoid}(Conv(F_{cat})) \tag{2}$$

$$P=P+(I\cdot\alpha+D\cdot(1-\alpha))\cdot rate \tag{3}$$

式(1)中的 F_{cat} 表示 I 和 D 特征输入转换为 4 个特征

图拼接后的中间特征量，式(3)中的 P 代表 P 分支当前阶段的特征量。

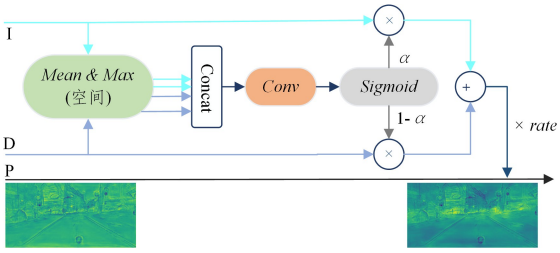


图 3 轻量注意力融合模块

Fig. 3 Lightweight attention fusion module

3.3 快速聚合金字塔池化模块(FAPPM)

在自动驾驶场景中，同一类别的目标通常会有多种尺度，这就要求模型具有从不同尺度中捕获语义信息的能力。为了提升网络的多尺度表征能力，Zhao 等[21]引入了金字塔池化模块(Pyramid Pooling Module，PPM)，该模块通过对输入特征进行多尺度的池化操作，增强了网络的上下文解析能力。Hong 等提出的深度聚合金字塔池化模块(DAPPM)进一步提高了 PPM 的上下文解析能力，并且表现出更优越的性能。然而，由于其深度较深，损失了不少速度。因此，本文结合快速空间金字塔池化(Spatial Pyramid Pooling-Fast，SPPF)[22]的思想优化了 DAPPM 模块中池化分支的连接方式，如图 4 所示。输入特征串行通过 3 个 5×5 大小的平均池化层(Average Pooling Layer，Avg)生成 3 种尺度的特征，再与全局平均池化分支和 1×1 卷积所生成的特征进行级联聚合。这个新的上下文提取模块被称为快速聚合金字塔池化模块(FAPPM)，置于架构末端。模块中的每个分支都可以捕获到不同感受野的高级语义信息，通过特征聚合，进一步提升模型的上下文解析能力。相比 DAPPM 模块，本文提出的 FAPPM 模块在保持相同感受野的情况下拥有更快的速度、更低的参数量和计算量。

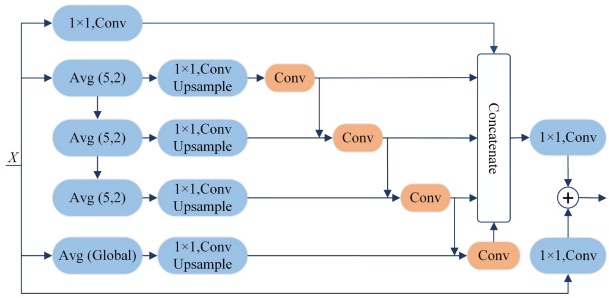


图 4 快速聚合金字塔池化模块

Fig. 4 Fast aggregate pyramid pooling module

3.4 深监督训练策略

训练阶段的额外监督可以增强卷积神经网络(DCNN)的学习。Yu 等在 BiseNetV2 中提出了一种增强训练策略，其中辅助分割头添加在每个阶段的语义分支末尾，但这也导致训练时间大大增加。考虑到训练效率和资源消耗，Hong 等提出的 DDRNet 采用了简单的深度监督训练策略，只在细节分支上添加一个辅助分割头。为获得更好的性能，在 DDRNet 深监督策略基础上，本文在细节分支的 Stage-2 和 Stage-5 处添加额外辅助边界头，并选用鲁棒性强的 Canny 算子对最终输出进行了边界提取，最后采用边界损失[23]来反向传播，如图 2 所示。训练阶段模型的损失表达式如式(4)所示：

$$Loss=L_S+\lambda_0 L_{SB}+\lambda_1 L_{auxS}+\lambda_2 L_{auxB2}+\lambda_3 L_{auxB5} \tag{4}$$

其中，L_S 表示最后预测输出的损失，L_{SB} 表示最后预测结果边界提取后的损失，L_{auxS} 表示辅助分割头的损失，L_{auxB2} 和 L_{auxB5} 分别表示 Stage-2 和 Stage-5 处的辅助边界头的损失，将它们加权求和得出训练阶段的总损失 $Loss$。借鉴先前工作的成功经验，权重值分别设为：$\lambda_0=20$，$\lambda_1=0.4$，$\lambda_2=8$，$\lambda_3=20$。

4 实验

4.1 数据集

Cityscapes[24]是一个被广泛用于城市道路场景中语义分割任务的数据集。该数据集包含了 5 000 张经过精细标注的图像，其中 2975 张用于训练，500 张用于验证，1 525 张用于测试。图像的分辨率均为 2048×1 024，这对于实时语义分割是极具挑战性的。本文实验使用了其中的 19 个类别，以便与其他方法进行公平比较。

CamVid[25]是另一个被用于语义分割任务的数据集，其中包含 701 张道路驾驶场景的图像。同大多数方法一样，367 张用于训练，101 张用于验证，233 张用于测试，且只使用标注的 32 个类别中的 11 个，图像的分辨率均为 960×720。

KITTI[26]是一个被用于自动驾驶语义分割的数据集。除了其他计算机视觉任务外，它还提供 200 个具有精细注释的图像。其类别标签方案与 Cityscapes 兼容，因此类似的训练和评估方法也适用于 KITTI。为了解决该数据集中样本数量少的问题，本文使用了迁移学习，即基于 Cityscapes 预训练权重进行微调。

4.2 实验细节

4.2.1 预训练

在微调训练之前，首先在 ImageNet[27]数据集上进行预训练，以降低模型的收敛难度。具体来说，将架构中 Stage-5 阶段的 FAPPM 去掉，并用分类头替换掉多分支融合后的分割头，进而构建最终的分类模型。模型的输入分辨率为 224×224，批量大小为 256，在单张 RTX4090 GPU 上训练 100 个轮次，同时使用随机梯度下降(SGD)作为优化器，权重衰减为 0.000 1，Nesterov 动量设为 0.9。学习率的初始值为 0.1，每 30 轮减少为原来的 1/10。图像随机裁剪为输入分辨率大小，并通过水平翻转来进行数据增强。

4.2.2 训练

本文模型采用的训练参数与先前大部分工作[8,28-30]相同，使用动量为 0.9 的随机梯度下降(Stochastic Gradient Descent，SGD)算法作为优化器，还采用了多项式学习率更新策略。同时使用简单的数据增强，包括随机裁剪、随机水平翻转以及 0.5～2.0 范围的随机缩放。三大数据集 Cityscapes、CamVid 和 KITTI 上的训练轮数、初始学习率、权重衰减、裁剪尺寸和批量大小如表 1 所列。

表 1 模型训练参数

Table 1 Model training parameters

Datasets	*Epoch*	*LR*	*WD*	*CS*	*BS*
Cityscapes	500	1×10^{-2}	5×10^{-4}	1 024×1 024	8
CamVid	200	1×10^{-3}	5×10^{-4}	960×720	8
KITTI	100	1×10^{-3}	5×10^{-4}	1 280×384	8

4.2.3 推理

本文测量推理速度的平台由单个 RTX3090，PyTorch1.8，CUDA11.7，cuDNN8.5 和 Ubantu-22.04 环境组成。遵循 STDCSeg，DDRNet，PIDNet 中采用的测速协议，测速前将 BatchNorm 合并到卷积层中，并将批量大小设置为 1。

4.3 对比实验

本文模型分别在 Cityscapes，Camvid 和 KITTI 数据集上与近几年优秀的实时语义分割模型进行了实验对比，并在不同场景对所提方法进行了可视化分析。本文方法分别使用平均交并比(mean Inter-section over Union，mIoU)、每秒处理帧数(Frames Per Second，FPS)和参数量(Parameters，Params)作为评估算法的分割精度、推理速度和模型规模的指标。

4.3.1 Cityscapes 实验结果

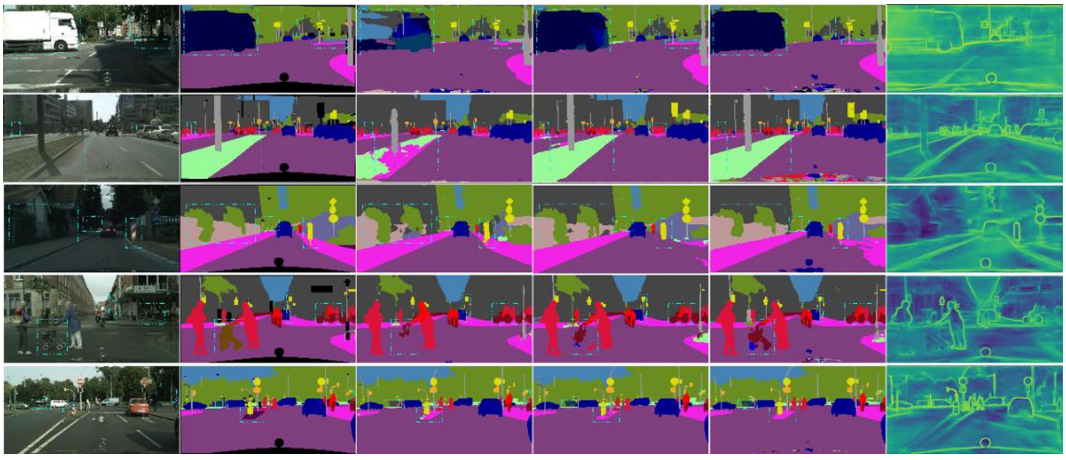
如表 2 所列，与大多数方法相比，所提算法(PSNet)在 Cityscapes 数据集上实现了更好的分割精度(其中带 * 字符的方法表示在本文实验平台上进行了精度和速度的复现)。PSNet 的 mIoU 值为 78.5%，略低于 78.9%的 SFNet，但速度快了其近 3 倍，达到 120.9 FPS。就推理速度而言，虽然本文模型不是最快的，但仍然优于大多数方法，推理速度比 FasterSeg[32] 和 Bise-NetV2 稍慢，但它们的分割精度分别低了 7% 和 5.9%。与过去两年提出的最新先进方法(包括 HyperSeg，STDCSeg 和 PP-LiteSeg)相比，所提算法不仅实现了更高的精度，而且表现出更好的速度性能。与现有基于边界分支的 PID 架构 DMRNet[33] 和 PIDNet 相比，本文提出的 PID 架构更好地平衡了精度和速度，且速度比最先进的 PIDNet 算法高出 20%以上。

表 2 Cityscapes 对比结果

Table 2 Comparision results on Cityscapes

Model	*mIoU*/%		*FPS*	GPU	Resolution	Params	Year
	val	test					
FasterSeg[32]	73.1	71.5	**163.9**	GTX 1080Ti	2 048×1 024	4.4×10^6	2019
SwiftNet[31]	75.5	75.4	39.9	GTX 1080Ti	2 048×1 024	11.8×10^6	2019
BiseNetV2	75.8	75.3	47.3	GTX 1080Ti	1 024×512	—	2020
DMRNet[33]	78.2	77.6	68.7	RTX 2080Ti	2 048×1 024	6.9×10^6	2023
CABiNet[34]	76.6	75.4	76.5	RTX 2080Ti	2 048×1 024	2.6×10^6	2021
HyperSeg *	76.2	75.8	59.1	RTX 3090	1 024×512	10.1×10^6	2020
SFNet *[7]	—	**78.9**	30.4	RTX 3090	2 048×1 024	12.9×10^6	2020
STDCSeg *	77.0	76.8	58.2	RTX 3090	1 536×768	—	2021
PP-LiteSeg *	78.2	77.5	68.2	RTX 3090	1536×768	—	2022
DDRNet *	77.8	77.4	140.4	RTX 3090	2 048×1 024	5.7×10^6	2022
PIDNet *	78.5	78.2	98.7	RTX 3090	2 048×1 024	7.6×10^6	2023
PSNet(Ours)	**78.5**	78.5	120.9	RTX 3090	2 048×1 024	9.9×10^6	—

图 5 展示了不同模型输出的可视化结果，包括 Cityscapes 测试集上辅助边界头的输出。在此示例中，青色虚线框表示不同模型对某些类别的预测的差异。从图中可看出，双边架构存在一定的超调现象，尤其体现在被“侵蚀”的电线杆和公交车上。除此之外，其对一些颜色相近的类别存在不同程度的混淆问题，而本文模型几乎不存在类似问题，且具有更精确的上下文解析。结合表 1 中的实验数据，可以得出所提模型在精度和速度之间实现了新的最优平衡的结论。



(a)原图 (b)Ground Truth (c)BiseNetV2 (d)DDRNet (e)Ours (f)Staage-5 边界特征图

图 5 Cityscapes 预测结果可视化(电子版为彩图)

Fig. 5 Visualization of prediction results on Cityscapes

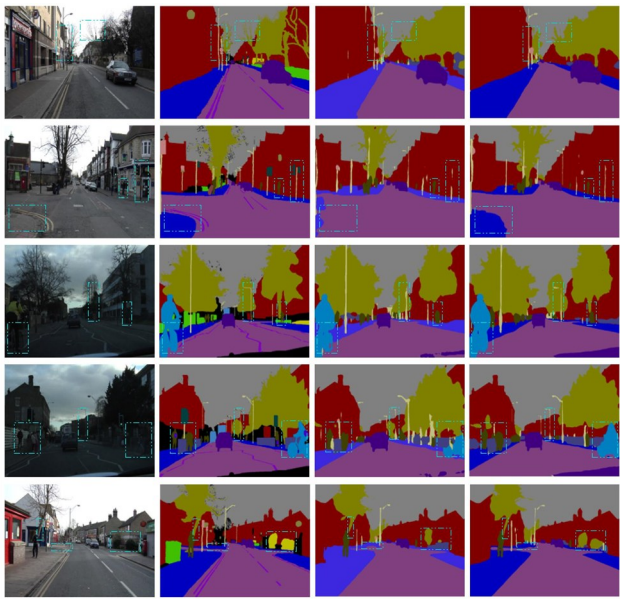
4.3.2 CamVid 实验结果

如表 3 所列，所提模型（PSNet）的推理速度和精度超过了表中大多数模型。PSNet 的 mIoU 值为 79.6%，超过了之前最先进的模型 DDRNet，且时延增加不到 1 ms。图 6 显示了本文算法在 CamVid 测试集上与 DDRNet 的可视化对比，可以看出，本文方法不仅对“道路”“树”和“杆”等类别的预测表现更好，而且在带有遮挡的挑战性场景中也表现出更好的鲁棒性。

表 3 CamVid 对比结果

Table 3 Comparision results on CamVid

Model	*mIoU*/%	*FPS*
STDC2Seg	73.9	152.6
PP-LiteSeg	75.0	154.8
DMRNet	75.8	96.1
BiseNetV2	76.7	124.0
HyperSeg	78.4	38.0
DDRNet	78.6	230.0
PSNet(Ours)	**79.6**	255.1



(a)原图 (b)Ground Truth (c)DDRNet (d)Ours

图 6 CamVid 上的预测结果可视化

Fig. 6 Visualization of prediction results on CamVid

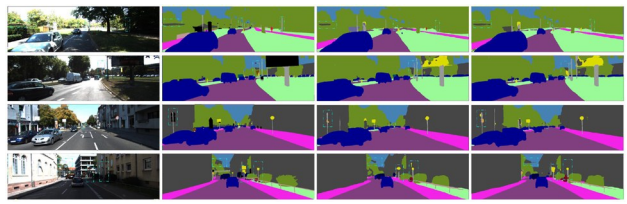
4.3.3 KITTI 实验结果

在相同实验环境下复现了两大先进算法的精度与速度，并将本文算法与之对比。表 4 中的结果表明，双边架构的 DDRNet 明显拥有更快的速度，但精度却远不如基于 PID 架构的算法，分割精度相比本文方法低了 5.6%。相比 PIDNet 网络，本文模型在精度和速度方面都有更好的表现，尤其在速度方面超过其 30FPS 以上。图 7 给出了两种算法在 KITTI 验证集上的可视化效果，可以看出，尽管所提算法没有边界分支，但在小目标或遮挡目标区域上 PSNet 具有相似或更好的预测效果。

表 4 KITTI 对比结果

Table 4 Comparision results on KITTI

Model	*mIoU*/%	*FPS*
DDRNet*	59.3	**436.8**
PIDNet*	64.5	290.2
PSNet(Ours)	**64.9**	324.2



(a)原图 (b)Ground Truth (c)PIDNet (d)Ours

图 7 KITTI 的预测结果可视化

Fig. 7 Visualization of prediction results on KITTI

4.3.4 不同场景预测效果分析

为了进一步验证本文算法的泛化性和普适性，从 Cityscapes 测试集和 KITTI 测试集（均无真实标签）中选出 4 对自动驾驶常见场景样例进行预测并分析，其中场景包括街道道路场景、城市道路场景、郊区道路场景和高速公路场景。

如图 8(a)所示，本文模型针对街道场景的分割整体达到了精细分割水平，其中对于易于混淆的植被与建筑、人与自行车，模型能够准确地区分并划分出完整轮廓；图 8(b)展示了城市道路场景的分割效果，本文模型能够精确且光滑地分割出道路、植被、人群等重要类别，甚至能够检测出道路上的抛洒物；图 8(c)为两个郊区道路场景样例，第一个样例由于涉及的类别较为常见，因此分割得较好，第二个样例相对复杂，对于训练标签中不存在的铁路类别，模型将其识别为了颜色相近的建筑物类别；图 8(d)的第二个样例的预测存在相同情况，模型将护栏识别为了人行道，除此特殊情况之外，模型对于这两个场景的分割效果尤其是道路和车表现优异。综上所述，本文模型能够适应不同场景，具有较好的泛化性和鲁棒性。

(a)街道 (b)城市 (c)郊区 (d)高速公路

图 8 不同场景预测结果可视化

Fig. 8 Visualization of prediction results in different scenarios

4.4 消融实验

为验证本文方法的有效性，基于 Cityscapes 的验证集对算法各个组件进行了消融实验。

4.4.1 精确语义网络（PSNet）和轻量注意力融合模块（LAFM）的有效性

为了验证所提出的 PSNet 架构和 LAFM 的有效性，首先从基线（简单的双边网络，只有一个细节分支（P）和一个语义分支（I））开始，然后逐个添加组件。表 5 列出了该过程的实验结果。实验结果表明，PID 架构中的 D 分支采用边界分支会直接导致推理速度降低近 30 FPS，尽管有 0.9%的精度提升，但这显然不是一个好的方案。而本文方法采用轻量的辅助语义分支作为 D 分支，在精度相似的情况下速度直接提升

超过10FPS。在此基础上添加本文提出的轻量注意力融合模块(LAFM),精度最高提升了1.7%。由实验结果可以得出,ImageNet预训练(IM)也很有效,对模型有1.1%的精度增益。

表5 骨干消融实验

Table 5 Backbone ablation experiment

IM	P+I分支	边界分支(D)	辅助语义分支(D)	低分辨率-高分辨率融合			*mIoU*/%	*FPS*
				逐点相加	LAFM(rate=1)	LAFM(rate=0.5)		
	√						69.6	**135.1**
	√	√					70.5	106.4
	√		√				70.3	117.1
	√		√	√			70.8	112.2
	√		√		√		71.0	106.3
	√		√			√	71.3	106.3
√						√	**72.4**	106.3

4.4.2 快速聚合金字塔池化模块(FAPPM)的有效性

在实时语义分割模型中,提取上下文信息的模块应该在提高准确性的同时平衡推理速度。因此,本文提出的FAPPM基于DAPPM实现了精度、速度和参数量的优化。这一小节的实验以无多尺度聚合模块的PSNet作为基线。根据表6的结果可以得出,加入FAPPM可以使mIoU值提升4.3%。与DAPPM相比,FAPPM不仅在精度方面表现更好,而且速度更快、参数量更少。

表6 FAPPM消融实验

Table 6 FAPPM ablation experiment

Module	*mIoU*/%	*FPS*	*Params*
None	72.4	106.3	7.9×10^6
DAPPM	76.0	99.1	10.0×10^6
FAPPM	**76.7**	99.9	9.9×10^6

4.4.3 辅助损失函数的有效性

本小节比较了边界损失、OHEM(Online Hard Example Mining)交叉熵损失[35]和辅助损失对模型的影响。基线是仅使用二元交叉熵损失函数进行训练的PSNet。表7中,当OHEM与辅助损失结合使用时,效益达到最大,mIoU增加1%以上。边界损失也很有效,最终模型的mIoU值达到78.5%。

表7 辅助损失函数消融实验

Table 7 Auxiliary loss function ablation experiment

OHEM	辅助损失	边界损失	*mIoU*/%
			76.7
	√		77.5
√			76.2
√	√		77.9
√	√	√	**78.5**

结束语 针对现有PID架构中边界分支影响推理速度的问题,本文探索了一种基于空间注意力机制和轻量辅助语义分支的高效PID架构,该架构能提升推理速度并缓解TBN架构中存在的超调现象。同时,本文提出的轻量级注意力融合模块利用空间注意力机制来增强语义解析能力并指导不同分支的特征融合。此外,本文设计的FAPPM模块简化了传统金字塔池化模块的连接方式,提高了运行效率。在三大公共自动驾驶数据集Cityscapes,CamVid和KITTI上的实验结果充分表明了所提模型的有效性。虽然所提模型在GPU上能以良好的速度实现高精度分割,但当脱离GPU或用于移动平台时,其性能可能会差强人意。因此探索移动端高精度语义分割模型会是未来的研究方向之一。

参 考 文 献

[1] FENG D, HAASE-SCHÜTZ C, ROSENBAUM L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3): 1341-1360.

[2] ASGARI T S, ABHISHEK K, COHEN J P, et al. Deep semantic segmentation of natural and medical images: a review[J]. Artificial Intelligence Review, 2021, 54: 137-178.

[3] YUAN X, SHI J, GU L. A review of deep learning methods for semantic segmentation of remote sensing imagery[J]. Expert Systems with Applications, 2021, 169: 114417.

[4] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[5] CHEN Q S, ZHANG Y, PU L, et al. Multi-path Semantic Segmentation Based on Edge Optimization and Global Modeling[J]. Computer Science, 2023, 50(S1): 2207137.

[6] WANG Y, ZHOU Q, LIU J, et al. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation [C]// 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 1860-1864.

[7] LI X, YOU A, ZHU Z, et al. Semantic flow for fast and accurate scene parsing[C]// Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23—28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020: 775-793.

[8] XU J, XIONG Z, BHATTACHARYYA S P. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19529-19539.

[9] BEZDEK J C. A convergence theorem for the fuzzy ISODATA clustering algorithms[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1980(1): 1-8.

[10] EMARA T, MUNIM H E, ABBAS H M, et al. LiteSeg: A Novel Lightweight ConvNet for Semantic Segmentation[J]. arXiv: 1912.06683, 2019.

[11] NIRKIN Y,WOLF L,HASSNER T. Hyperseg:Patchwise hypernetwork for real-time semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:4061-4070.

[12] RONNEBERGER O,FISCHER P,BROX T. U-net:Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015:18th International Conference,Munich,Germany,October 5 — 9, 2015, Proceedings, Part III 18. Springer International Publishing,2015:234-241.

[13] FAN M,LAI S,HUANG J,et al. Rethinking bisenet for real-time semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:9716-9725.

[14] PENG J,LIU Y,TANG S,et al. Pp-liteseg:A superior real-time semantic segmentation model[J]. arXiv:2204. 02681,2022.

[15] YU C,WANG J,PENG C,et al. Bisenet:Bilateral segmentation network for real-time semantic segmentation[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:325-341.

[16] YU C,GAO C,WANG J,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.

[17] HONG Y D,PAN H H,SUN W C,et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes[J]. arXiv:2101. 06085,2021.

[18] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[19] LI H,XIONG P,FAN H,et al. Dfanet:Deep feature aggregation for real-time semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:9522-9531.

[20] SONG Q,MEI K,HUANG R. AttaNet:Attention-augmented network for fast and accurate scene parsing[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021:2567-2575.

[21] ZHAO H,SHI J,QI X,et al. Pyramid scene parsing network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2881-2890.

[22] JOCHER G. YOLOv5 by Ultralytics(Version 7. 0) [EB/OL]. https://doi. org/10. 5281/zenodo. 3908559.

[23] TAKIKAWA T,ACUNA D,JAMPANI V,et al. Gated-scnn:Gated shape cnns for semantic segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:5229-5238.

[24] CORDTS M,OMRAN M,RAMOS S,et al. The cityscapes dataset for semantic urban scene understanding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:3213-3223.

[25] 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.

[26] ABU ALHAIJA H,MUSTIKOVELA S K,MESCHEDER L,et al. Augmented reality meets computer vision:Efficient data generation for urban driving scenes[J]. International Journal of Computer Vision,2018,126:961-972.

[27] RUSSAKOVSKY O,DENG J,SU H,et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision,2015,115:211-252.

[28] GENG H,JIANG J,SHEN J,et al. Cascading Alignment for Unsupervised Domain-Adaptive DETR with Improved DeNoising Anchor Boxes[J]. Sensors,2022,22(24):9629.

[29] GU Y H,HAO J,CHEN B. Semi-supervised Semantic Segmentation for High-resolution Remote Sensing Images Based on DataFusion[J]. Computer Science,2023,50(S1):22050001-6.

[30] CHEN L,XU G,FU N N,et al. Research on 3D Point Cloud Semantic Segmentation Method Fused with Edge Detection[J]. Journal of Chongqing Technology and Business University(Natural Science Edition),2022,39(5):1-9.

[31] 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]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:12607-12616.

[32] CHEN W,GONG X,LIU X,et al. Fasterseg:Searching for faster real-time semantic segmentation [J]. arXiv: 1912. 10917, 2019.

[33] WANG Y,CHEN S,BIAN H,et al. Deep Multi-Resolution Network for Real-Time Semantic Segmentation in Street Scenes [C]// 2023 International Joint Conference on Neural Networks (IJCNN). IEEE,2023:1-8.

[34] KUMAAR S,LYU Y,NEX F,et al. Cabinet:Efficient context aggregation network for low-latency semantic segmentation [C]// 2021 IEEE International Conference on Robotics and Automation(ICRA). IEEE,2021:13517-13524.

[35] SHRIVASTAVA A,GUPTA A,GIRSHICK R. Training region-based object detectors with online hard example mining [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:761-769.

GENG Huantong, born in 1973, professor, Ph.D supervisor, is a senior member of CCF(No. 12356S). His main research interests include multi-objective optimization and deep learning.

(责任编辑:何杨)