基于双分支融合的图像实时语义分割方法

宋玉琴，娄 辉，张 琪，商纯良
（西安工程大学电子信息学院，西安，710600）

摘要 针对现有实时语义分割网络分割多尺度目标时存在类别错分和分割不完整的问题，提出了一种基于双分支融合的图像实时语义分割方法。提出尺度注意融合模块，融合细节分支和语义分支提取到的目标空间特征和语义信息，以提高网络对多尺度目标识别的准确率。使用边缘损失函数引导细节分支学习目标边缘轮廓，增强网络对目标边缘细节的分割性能。最后，构建全局感知模块提高网络的全局上下文感知能力。实验结果表明：文中方法在 CityScapes 和 CamVid 数据集上平均交并比（mIoU）分别为 78.1%和 76.2%，平均像素准确率（mPA）分别为 87.6%和 85.4%，对于小尺度目标边缘实现了更精准的分割，且在一个 GTX 1080Ti GPU 上推理达到实时要求，帧速率（FPS）分别达到 59.8 和 43.5。

关键词 深度学习；实时语义分割；尺度注意；特征融合；全局感知
DOI 10.3969/j.issn.2097-1915.2025.02.008
中图分类号 TP391.4 **文献标志码** A **文章编号** 2097-1915(2025)02-0062-09

A Real-Time Image Semantic Segmentation Method Based on Dual Branch Fusion

SONG Yuqin, LOU Hui, ZHANG Qi, SHANG Chunliang
(School of Electronics and Information, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract Aimed at the problems that faulty classification and incomplete segmentation are in existence in segmenting multi-scale objects to the existing real-time semantic segmentation networks, a real-time semantic image segmentation method is proposed based on dual branch fusion. The method introduces a scale attention fusion module that is able to fuse object spatial feature and semantic information extracted from the detail branch and semantic branch, thereby improving the accuracy of the network for multi-scale object recognition. The edge loss function is used to guide the detail branch into learning the object edge contour, improving the network's segmentation performance on object edge details. Finally, a global perception module is constructed to enhance the global context perception capability of the network. The experimental results demonstrate that the proposed method achieves the mean Intersection over union (mIoU) of 78.1% and 76.2% on the CityScapes and CamVid datasets respectively. Additionally, the mean pixel accuracy (mPA) is 87.6% and 85.4%, respectively. For small-scale object edges, there is a more accurate segmentation, coming up to the real-time requirements on a single GTX 1080Ti GPU, and frames per second (FPS) achieves 59.8 and 43.5 respectively.

Key words deep learning; real-time semantic segmentation; scale attention; feature fusion; global perception

收稿日期：2024-05-09
基金项目：中国纺织工业联合会科技指导性项目(2019062)；陕西省教育厅专项科研计划项目(18JK0358)
作者简介：宋玉琴(1972—)，女，陕西西安人，副教授，研究方向为机器视觉与图像处理。E-mail:81308995@qq.com
通信作者：娄 辉(1998—)，男，陕西宝鸡人，硕士生，研究方向为目标检测、图像分割。E-mail:2327186542@qq.com

引用格式：*宋玉琴，娄辉，张琪，等. 基于双分支融合的图像实时语义分割方法[J]. 空军工程大学学报，2025，26(2)：62-70. SONG Yuqin, LOU Hui, ZHANG Qi, et al. A Real-Time Image Semantic Segmentation Method Based on Dual Branch Fusion[J]. Journal of Air Force Engineering University, 2025, 26(2): 62-70.*

作为像素级的分类任务，语义分割为自动驾驶及机器人理解周围场景提供技术支撑[1-3]。由于室外场景复杂多样，算法需要实时地对采集的图像分割。随着对移动设备部署需求的不断增长，实时语义分割算法受到越来越多的关注[4-7]，而现有方法受网络层数以及逐层解码的影响，导致模型参数量和计算量指数级增长，给自动驾驶和移动机器设备的实时运行带来了艰难挑战。基于此，研究者提出基于深度学习的实时语义分割算法，主要分为基于编码-解码结构和基于多路径架构 2 种。

基于编码-解码结构的方法主要通过设计高性能编码器[8-9]或轻量化解码器[10-11]实现网络的快速性。高性能编码器采用空洞卷积[12]、深度可分离卷积[13]及非对称卷积[14]等操作取代原始卷积，在降低网络层数同时增强特征提取能力。然而这些操作会丢失局部特征信息，同时增加模型训练成本。轻量化解码器由于对网络层数和通道数进行删减，导致分割后的目标出现锯齿化现象。为解决此类问题，解码器通过构建图像金字塔有效获取图像的上下文信息。PSPNet[15]采用不同大小的池化核组建金字塔池化模块增大了感受野，减少了不同区域之间图像信息丢失。这种特征金字塔模块虽然增强了不同尺度特征之间的交互性，但分割效果仍受下采样过程中损失的空间信息影响。为了克服这一问题，Zhao 等[16]提出的 ICNet 通过多分支网络，对不同尺度输入图像分别提取空间和细节特征，最后通过级联方式融合。由于网络使用三分支子网，大大增加了模型参数量。BiSeNetv2 网络[17]通过在浅层空间分支中提取图像细节特征，在深层语义分支中获取语义信息，最后结合两分支特征强化输出特征表达。然而，为了追求实时性，语义分支解码器上采样 8 倍后直接与空间分支加和，忽略了不同层级间信息的差异性，导致分割图像存在边缘锯齿状现象。

针对上述问题，本文提出一种基于双分支尺度注意融合和边缘增强的轻量型语义分割算法。在编码阶段，提出一种尺度注意融合（scale attention fusion，SAF）模块，以增强网络对像素点的语义类别信息获取和空间位置关联利用的能力，提高对不同特征类内和类间的信息提取能力，降低网络下采样过程中细节信息丢失的影响。解码阶段，在细节分支输出层引入边缘增强分割头，通过拉普拉斯算子卷积（Laplacian convolution，LaplacianConv）求取目标标签的边缘轮廓引导细节分支特征提取。该模块采用辅助训练策略，仅在训练时增强对图像边缘分割性能，实际推理中不增加运算量。此外，设计了一种全局感知（global perception，GP）模块，嵌入到双分支网络解码器输出末尾，进一步感知像素点之间全局关联信息，丰富上下文信息表达。

1　双分支实时语义分割网络

本文所提网络结构如图 1 所示，其中 Layer-S 表示语义分支中的编码层，Layer-D 表示细节分支中的编码层。

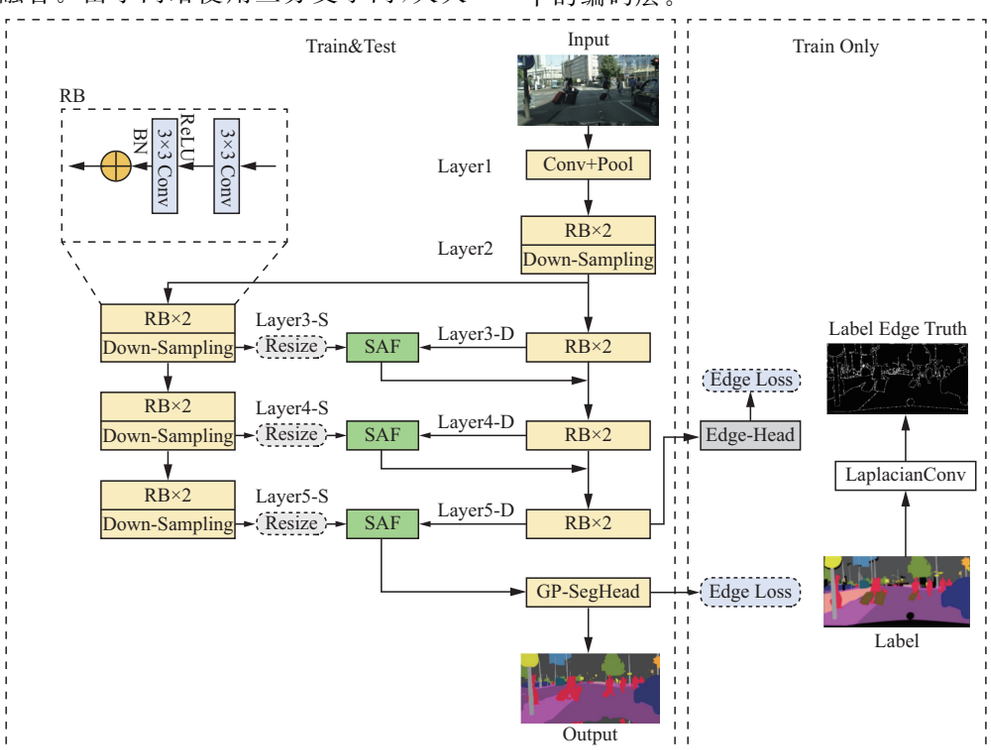


图 1　网络整体结构

Fig. 1　Network overall framework of proposed method

网络整体采用轻量型 ResNet-18 结构，使用卷积层替换最后全连接层，组建语义-细节双分支卷积网络。网络共分为 5 个阶段，第 1 阶段通过跨步卷积和池化，对输入进行 4 倍下采样操作并扩充通道数。第 2 阶段采用 ResNet-18 中的残差基础块（residual block，RB）对特征进一步下采样，以获取深层感受野。第 3～5 阶段，网络分为语义分支和细节分支 2 条支路，为了网络运行更加快速，2 个分支共享第 1、2 阶段权重。

语义分支使用 RB 对前 2 个阶段捕获的特征下采样，通过增大像素点感受野的方式让网络学习到图像语义特性。语义分支由 3 个残差卷积采样块组成，依次将特征采样至原图大小的 1/16、1/32、1/64。细节分支保持当前分辨率大小的同时进一步增大感受野，在获取像素点细节特性的同时，通过融合语义分支特征使得网络对目标语义理解得更精确。

语义分支对特征下采样提取深层语义像素信息，细节分支通过高分辨率图像获取目标边缘细节。SAF 模块融合两分支特征并输入到下一层的细节分支中，以增强对深层语义信息的感知能力。边缘增强分割头通过获取目标边缘轮廓，进一步增强细节分支的特征提取能力。最后，将两分支融合后的特征由 GP 模块组成的分割头对像素点作类别分割。

1.1 尺度注意融合模块

为了平衡像素点类内与类间的差异性，设计了尺度注意融合（scale attention fusion，SAF）模块，详细结构如图 2 所示。通过结合全局注意力与局部注意力方法，SAF 对双分支子网提取的语义特征与细节信息融合，获取像素全局上下文信息同时保留小目标细节特征。

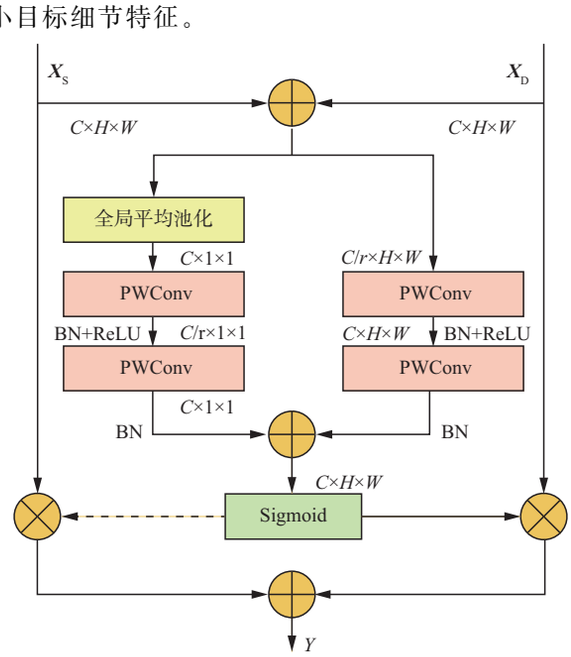


图 2 尺度注意融合模块结构

Fig. 2 Architecture of scale attention fusion module

全局注意力将每个大小为 $H\times W$ 的特征图压缩为一维向量，其计算式为：

$$g(\boldsymbol{x})=\mathrm{BN}[\boldsymbol{\omega}_2\delta(\boldsymbol{\omega}_1\mathrm{Pool}_{\mathrm{avg}}(\boldsymbol{x}))] \tag{1}$$

首先，经全局平均池化将特征空间维度大小压缩为 1×1，通过权重为 $\boldsymbol{\omega}_1$ 的卷积层进行通道维度缩放，经过中间层 ReLU 函数 δ 激活后使用权重为 $\boldsymbol{\omega}_2$ 的逐点卷积层对通道维度还原。最后，经由批归一化处理操作 BN 得到全局注意特征。

全局注意力的关键思想是通过改变空间池化核的大小，在多个尺度上实现通道信息加权学习。为了尽可能保持轻量性，只使用细节分支相对应阶段的输出特征来提取局部上下文信息。

局部注意力使用逐点卷积（point-wise convolution，PWConv）提取像素局部空间细节特征，并对每个空间位置的像素点进行通道交互来聚合通道方向上下文特征。局部特征注意力 $L(\boldsymbol{x})$ 的计算式为：

$$L(\boldsymbol{x})=\mathrm{BN}[\boldsymbol{\omega}_2\delta(\boldsymbol{\omega}_1\boldsymbol{x})] \tag{2}$$

输入特征直接通过权重为 $\boldsymbol{\omega}_1$ 的逐点卷积层对通道维度缩放为原始 $1/r$ 大小，经 ReLU 函数 δ 激活后，再由权重为 $\boldsymbol{\omega}_2$ 的逐点卷积层对通道维度大小进行还原，最后经由批归一化处理操作 BN 得到局部上下文特征。因此，对于语义分支输出特征 $\boldsymbol{X}_S$ 及细节分支特征 $\boldsymbol{X}_D$，融合后的特征 $\boldsymbol{Y}$ 计算式为：

$$\begin{aligned}\boldsymbol{s}&=\sigma[L(\boldsymbol{X}_S+\boldsymbol{X}_D)\oplus g(\boldsymbol{X}_S+\boldsymbol{X}_D)]\\ \boldsymbol{Y}&=\boldsymbol{s}\otimes\boldsymbol{X}_D+(1-\boldsymbol{s})\otimes\boldsymbol{X}_S\end{aligned} \tag{3}$$

式中：σ 为 Sigmoid 函数；$\oplus$ 为使用广播机制对不同维度特征求和；M 为双分支子网的特征经过双重注意力生成的权重系数；$\otimes$ 表示逐元素乘法。

双分支子网融合后的特征分别由全局注意力和局部注意力生成相应特征向量，使用 Sigmoid 函数生成归一化通道权重向量 $\boldsymbol{s}\in\boldsymbol{R}^{C\times1\times1}$，与细节特征 $\boldsymbol{X}_D$ 逐通道相乘后，再将其与 1 的差值作用于语义特征，最后融合得到含有语义和上下文细节信息的特征。通过对双分支子网使用权值，网络能够在语义分支和细节分支之间进行软选择或加权平均。

SAF 通过改变空间池化的大小使得特征具有全局以外的尺度。聚合不同尺度上下文信息，可以同时强调更具全局分布的大型目标，突出分布更局部的小目标，增强网络在目标极端尺度变化下识别和分割的性能。

1.2 边缘增强分割头

细节分支采用边缘损失函数训练网络对特征边缘像素的感知力。特征边缘像素大都存在于不同类别之间的过渡区域，对过渡区域的精准分割极大影响了语义分割模型的最终性能。为了进一步增强网

络对目标边缘细节的识别能力，提出边缘增强分割头，通过卷积运算将细节分支融合后的特征压缩为单通道向量，得到特征点边缘的置信度值。同时使用拉普拉斯算子获取原始标签的边缘特征，拉普拉斯算子对邻域中心像素点求取方向梯度值，锐化图像特征实现边缘提取。拉普拉斯算子公式为：

$$\boldsymbol{H}=\begin{bmatrix}-1 & -1 & -1\\ -1 & 8 & -1\\ -1 & -1 & -1\end{bmatrix} \tag{4}$$

为了增强网络对不同尺度目标分割性能，边缘增强分割头采用不同步长大小卷积获取多尺度特征，然后将细节特征映射到原始大小，并与可训练的 1×1 卷积融合进行动态重构。最后，采用大小为 0.1 的阈值将预测的信息转换为具有边信息和角信息的边缘特征图。由于提取的边缘特征图为二值图，故边缘损失函数采用二分类交叉熵函数。对于细节分支预测的 $H\times W$ 特征图，其边缘损失计算式为：

$$L_{\text{edge}}(p,y)=-(y\ln(p)+(1-y)\ln(1-p)) \tag{5}$$

式中：p 为细节分支预测的边界细节特征值；y 为从原始标签提取的边界真值。

在边缘损失函数中，y 的取值为 0 或 1，当 $y=1$ 时表示标签的真实边缘点，鼓励预测的边界值 p 接近 1；当 $y=0$ 时表示标签的非真实边缘点，此时鼓励预测的边界值 p 接近 0。边缘区域像素的准确分割是进一步提高语义分割网络性能的关键，边缘损失函数能够针对边缘区域专门优化，增强网络对边缘区域的预测能力。使用拉普拉斯算子获取原始标签的边缘特征二值图，结合细节分支预测的边界细节特征值，通过最小化边缘损失函数，使得网络更加关注类别间的边缘区域。

将细节分支输出特征接入到边缘增强分割头，引导网络对浅层空间信息编码，通过在训练阶段求取二分类交叉熵损失引导细节分支提取图像边缘信息。此外，边缘增强分割头只在训练中采用，在推理阶段被丢弃，因此，可在高效提升分割任务准确性的同时无需花费额外推理成本。

1.3　全局感知模块

受卷积核大小限制，卷积网络对像素的全局信息利用能力有限，导致分割结果出现像素遗漏现象。期望最大化注意机制能够利用期望最大化算法迭代出简单紧凑的基，估算出像素点在全局的关联信息，而且能够有效降低计算成本和网络的整体参数量。因此，采用期望注意力机制[18]和残差卷积块搭建全局感知模块，提高网络对特征的全局信息利用能力，其结构如图 3 所示。

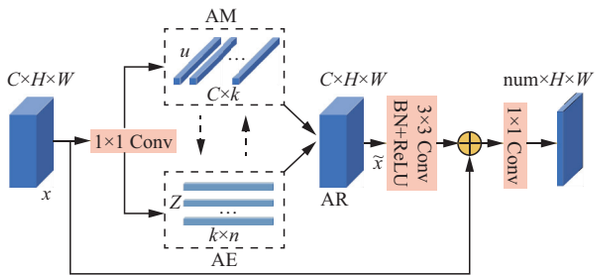


图 3　全局感知模块结构

Fig. 3　Architecture of global perception module

对于大小为 $C\times H\times W$ 的输入特征 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，其中 $n=H\times W$，每个像素点 x_i 都有对应的隐变量 z_i，再将 $\{\boldsymbol{X},\boldsymbol{Z}\}$ 组合为完整数据，其似然函数为 $\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{u})$，其中 $\boldsymbol{u}$ 为模型中所有参数的集合。注意力估计(attention estimation，AE)根据当前参数 $\boldsymbol{u}^{\text{old}}$ 计算隐变量 $\boldsymbol{Z}$ 的后验分布，并以此寻找完整数据的似然分布 $\delta(\boldsymbol{u},\boldsymbol{u}^{\text{old}})$，其表达式为：

$$\delta(\boldsymbol{u},\boldsymbol{u}^{\text{old}})=\sum\nolimits_{\boldsymbol{Z}}p(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{u}^{\text{old}})\ln p(\boldsymbol{X},\boldsymbol{Z}\mid\boldsymbol{u}) \tag{6}$$

注意力最大化(attention maximization，AM)的作用则是利用最大似然算法更新基 $\boldsymbol{u}^{\text{new}}$，即：

$$\boldsymbol{u}^{\text{new}}=\arg\max_{\boldsymbol{u}}\delta(\boldsymbol{u},\boldsymbol{u}^{\text{old}}) \tag{7}$$

AE 和 AM 交替进行，经过数次迭代得到近似收敛的基 $\boldsymbol{u}$ 和隐变量 $\boldsymbol{Z}$，进而通过注意力再估计(attention re-estimation，AR)对 $\boldsymbol{X}$ 重估得到 $\tilde{\boldsymbol{X}}$，即：

$$\tilde{\boldsymbol{X}}=\boldsymbol{u}\boldsymbol{Z} \tag{8}$$

通过 3×3 Conv-BN-ReLU 运算操作与原始输入 $\boldsymbol{x}$ 进行残差跳跃融合，最后由 1×1 卷积对特征通道维度缩放，得到通道维度为类别数目 num 大小的预测图。

2　实验设计

2.1　数据集

CityScapes 数据集[19]包含了采集自 27 个不同城市街道场景中的 5 000 张图像，总共包含 33 个类别。其中划分 2 975 张作为训练集，1 525 张作为测试集，500 张作为验证集。

CamVid 数据集[20]是剑桥大学发布的具有目标类别语义分割标签的视频序列集合。选用 701 帧图像作为图像语义分割，其中 367 帧作为训练集，233 帧作为测试集，101 帧作为验证集。

2.2　实验环境与设置

实验环境基于 Pytorch 深度学习框架，使用 1 个 GTX 1080Ti 显卡对网络训练及验证。设置 batch 大小为 8，学习率采用 Poly 策略，初始学习率

设置为 0.01，权值为 0.9，网络训练 500 轮。同时，对输入图像进行数据增强，采用随机水平翻转、平均减法和随机尺度方法，使用取值范围为[0.5，2.0]的参数对输入图像进行不同尺度上的转换。

2.3 评价指标

实验采用平均交并比(mean intersection over union，mIoU)和平均像素准确率(mean pixel accuracy，mPA)评判模型分割精度。mIoU 指类别标签中预测与真实值的交集和并集比值的平均数，mPA 为对每个标签类别准确率的平均值。计算式分别为：

$$P_{\mathrm{mIoU}}=\frac{1}{k}\sum_{i=1}^{k}\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}+F_{\mathrm{P}}} \tag{9}$$

$$P_{\mathrm{mPA}}=\frac{1}{k}\sum_{i=1}^{k}\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \tag{10}$$

式中：T_{P} 为正样本被正确预测的集合；F_{N} 为负样本被错误预测的集合；F_{P} 为正样本被错误预测的集合。

使用参数量(params)和帧速率(frames per second，FPS)衡量模型空间复杂度和推理速度。其中，参数量通过调用 Pytorch 库计算各模型参数大小，FPS 通过计算模型每秒推理图像数量得出。

3 实验结果与分析

3.1 可行性与先进性实验

本实验以 CityScapes 和 CamVid 数据集为测试基准，通过分析不同方法的分割精度验证本文方法的有效性。

如表 1 所示，在 CityScapes 数据集上，本文方法的 mIoU 和 mPA 分别达到了 78.1%和 87.6%，优于大部分方法，相较于 FANet-18，2 个指标分别提升 3.1%和 6.0%。在道路、栅栏、杆子、交通灯及行人类别上取得了最好效果，此外，在围墙和骑手类别中表现突出，优于大多数方法。

表 1 本文方法与其他方法在 CityScapes 数据集上的定量结果对比

Tab. 1 Comparisons between proposed method and state-of-the-art methods on CityScapes %

方法	mIoU	mPA	准确率						
			道路	围墙	栅栏	杆子	交通灯	行人	骑手
ENet[10]	58.5	67.3	96.3	32.2	33.2	43.4	44.0	65.5	38.4
ICNet[16]	69.5	75.4	97.1	43.3	48.9	56.5	63.4	74.6	54.1
BiSeNetV2[17]	73.4	78.8	97.7	56.4	53.4	59.3	67.1	76.9	**62.1**
FANet-18[21]	75.0	81.6	98.1	**63.4**	57.6	62.7	68.2	78.3	60.4
DFANet[22]	71.3	77.2	97.5	47.5	51.3	58.2	65.8	75.2	56.9
LAANet[23]	73.6		97.9	53.5	51.5	59.3	70.3	75.9	57.5
STDC[24]	76.8	83.4	98.0	60.9	**60.7**	63.2	76.1	80.2	57.9
LPSNet[25]	76.5	80.5	97.9	57.4	55.5	61.2	70.6	79.8	61.6
DDRNet[26]	77.4	84.8	98.1	61.3	60.3	63.0	76.6	80.5	58.5
本文方法	**78.1**	**87.6**	**98.2**	63.0	60.5	**63.5**	**77.0**	**81.0**	61.1

如图 4 所示，在 CityScapes 数据集上对不同模型的分割效果可视化，左下角矩形框内展示了所选目标的细节信息。本文方法对大目标对象，如道路、小汽车、植被等有较高的识别率，对像素较少的小目标，如杆子、信号标识、骑手等，不仅能够准确分割目标，在不同类别的边缘像素过渡处也实现了精准分割。如图 4(b)～图 4(e)第 1 列图片所示，本文方法对杆子分割更精确，证明网络通过多尺度融合后对不同尺度目标分割具有较好的性能。如图 4(b)～图 4(e)第 2 列图片所示，本文方法较好地分割出了骑手的腿和脚，效果仅次于 BiSeNetV2，对车轮的边缘和内部分割展示了比 BiSeNetV2 更好的效果，由此证明，细节分支对网络学习目标边缘信息能力具有增强效果。如图 4(b)～图 4(e)第 3 列图片所示，本文方法在杆子和信号标识类别分割中，可视化效果对比其他方法有显著性优势，没有出现其他方法分割不连续和漏检的现象。

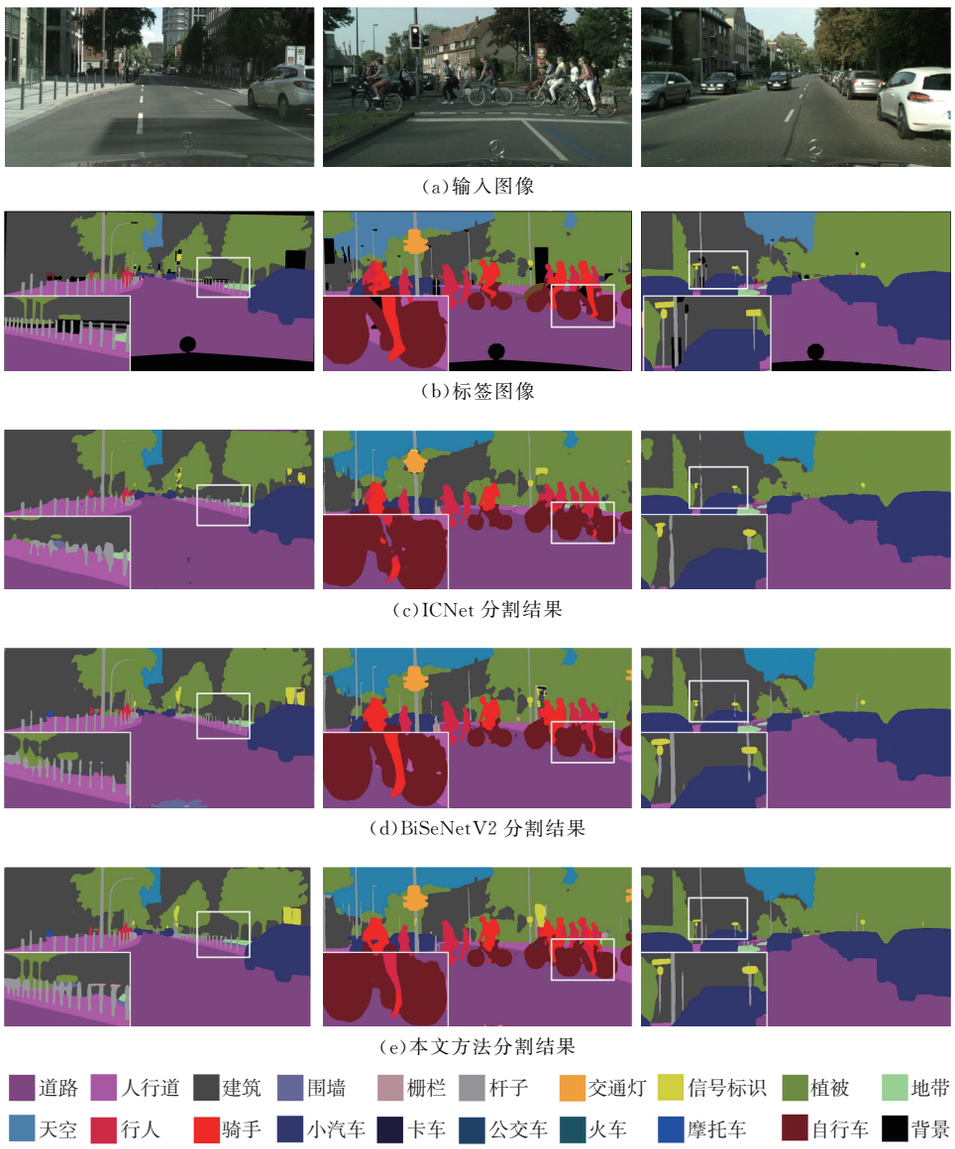


图 4　CityScapes 数据集上的可视化结果对比

Fig. 4　Visualization of semantic segmentation results on CityScapes

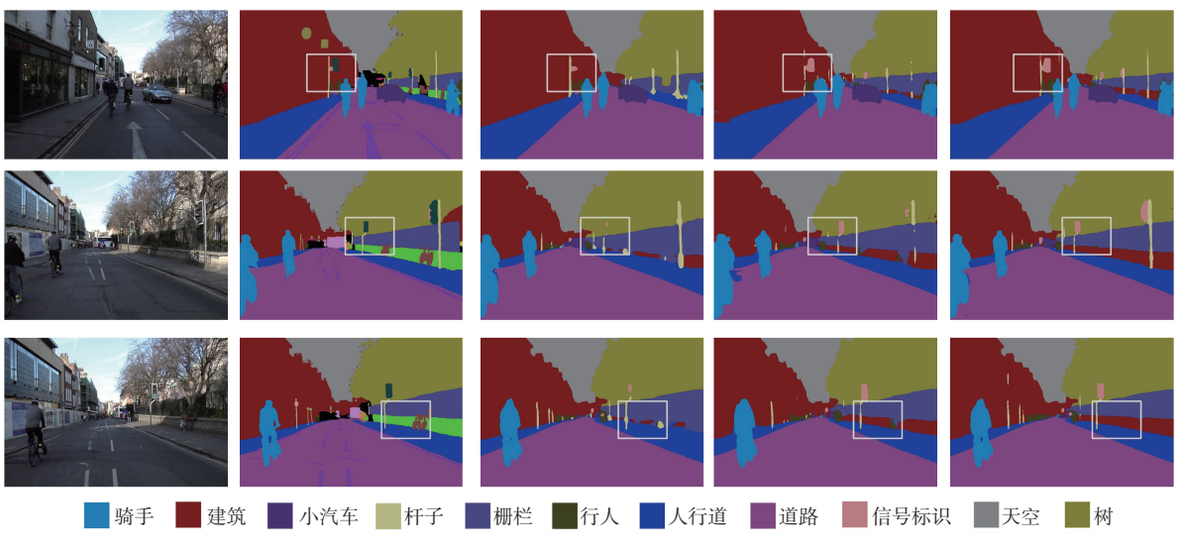
如表 2 所示，在 CamVid 数据集上，本文方法取得了 mIoU 为 76.2%和 mPA 为 85.4%的结果，相较于 DFANet，2 个指标分别提升 5.7%和 11.1%。在小汽车、栅栏、人行道以及信号标识类别的分割中，本文方法取得了最优结果，在骑手、杆子和行人类别分割中也表现突出，优于大多数方法。

表 2　本文方法与其他方法在 CamVid 数据集上的定量结果对比

Tab. 2　Comparisons between proposed method and state-of-the-art methods on CamVid　　%

方法	mIoU	mPA	准确率						
			自行车	小汽车	杆子	栅栏	行人	人行道	信号标识
ENet[10]	51.3	64.8	57.3	72.2	34.1	55.4	47.4	77.4	45.1
ICNet[16]	67.1	72.1	61.1	83.6	43.7	53.5	58.8	84.5	46.3
BiSeNetV2[17]	72.4	75.4	76.7	86.2	50.8	63.4	69.1	85.0	44.3
FANet-18[21]	74.7	79.2	**78.1**	90.1	49.3	65.4	75.4	87.3	45.4
DFANet[22]	70.5	74.3	73.5	87.5	45.3	59.2	65.3	84.5	40.6
LAANet[23]	67.9		68.7	83.7	36.6	59.2	62.9	80.7	**51.3**
STDC[24]	73.9	78.4	76.7	86.2	**51.8**	57.4	69.1	85.0	44.3
LPSNet[25]	75.1	83.2	74.2	90.6	49.5	65.5	71.4	86.1	42.2
DDRNet[26]	75.6	85.2	76.9	**93.4**	50.1	66.3	72.6	87.3	44.3
本文方法	**76.2**	**85.4**	77.2	93.1	50.5	**68.3**	**73.8**	**89.4**	46.7

CamVid 数据集截取自视频序列，图像为连续帧图像，实验选取了 3 帧图像作为可视化效果展示，如图 5 所示。从第 1 行图片可以看出，得益于边缘增强分割头设计，本文方法对骑手和行人的边缘可视化效果更为突出。第 2 行图片中，像素占比较小且类别分布不均衡的目标，如杆子，本文方法的可视化效果比其他方法有显著提高。第 3 行图片中，本文方法对于道路、建筑和树等类别可视化效果最明显。



(a)输入图像 (b)标签图像 (c)ICNet 分割结果 (d)BiSeNetV2 分割结果 (e)本文方法分割结果

图 5 CamVid 数据集上的可视化结果对比

Fig. 5 Visualization of semantic segmentation results on CamVid dataset

为了验证网络的实时性能，对所有方法测试模型参数量和帧速率。所有的测试基于 1 个 GTX 1080Ti GPU，测试图像选用 512×1 024 像素的 CityScapes 和 CamVid 数据集，以保证实验结果公正可靠。对于各算法模型参数量指标，实验使用 Thop 库函数计算，FPS 通过模型处理相同数量图像的时间转换求得。

由于移动设备运算速率低、内存小，所以算法的高帧率和低参数量是满足实际应用中实时需求的必要条件。如表 3 所示，本文方法在 CityScapes 数据集上能够以 59.8 帧/s 的速度处理图像，速度略低于 STDC，但是在分割精度上超出 STDC 1.3%，达到了 78.1% mIoU，模型参数也仅有 4.4×10^6，远远小于 STDC 的 12.5×10^6。

表 3 不同方法在 CityScapes 数据集上的性能对比

Tab. 3 Comparison of the performance of different methods on CityScapes dataset

方法	mIoU/%	params/10^6	FPS/(帧·s^{-1})
ICNet[16]	69.5	26.5	38.5
BiSeNetV2[17]	73.4	**3.4**	57.4
STDC[24]	76.8	12.5	**63.3**
DDRNet[26]	77.4	7.3	47.8
本文方法	**78.1**	4.4	59.8

如表 4 所示，本文方法在 CamVid 数据集上以 43.5 帧/s 的速度处理图像，且模型参数仅为 6.8×10^6。本文方法的 FPS 略低于 BiSeNetV2，然而在分割精度上达到了 78.1% mIoU，超出 BiSeNetV2 4.7%。综上所述，本文方法在分割精度和实际应用部署 2 个方面达到了平衡，具有良好的应用价值。

表 4 不同方法在 CamVid 数据集上的性能对比

Tab. 4 Comparison of the performance of different on CamVid dataset

方法	mIoU/%	params/10^6	FPS/(帧·s^{-1})
ENet[10]	51.3	**0.36**	41.6
ICNet[16]	67.1	26.50	27.7
BiSeNetV2[17]	72.4	6.40	**52.6**
FANet-18[22]	74.7	9.60	41.6
本文方法	**76.2**	6.80	43.5

3.2 消融实验

如表 5 所示，为了进一步验证本文方法的有效性，在 CityScapes 数据集上分别对各个模块实验，测试其对分割精度的影响。结果显示，仅使用 ResNet-18 作为双分支网络时，mIoU 和 mPA 分别为 74.6%和 76.3%。增加 SAF 模块后，mIoU 提高了 1.2%，mPA 提高了 4.4%。在此基础上使用边缘增强分割头(edge head)后，mIoU 和 mPA 分别提高了 1.6%和 4.8%。最后，在前 2 个模块的基础上添加 GP 模块，mIoU 和 mPA 分别提高了 0.7%和 2.1%，达到了 78.1%和 87.6%。

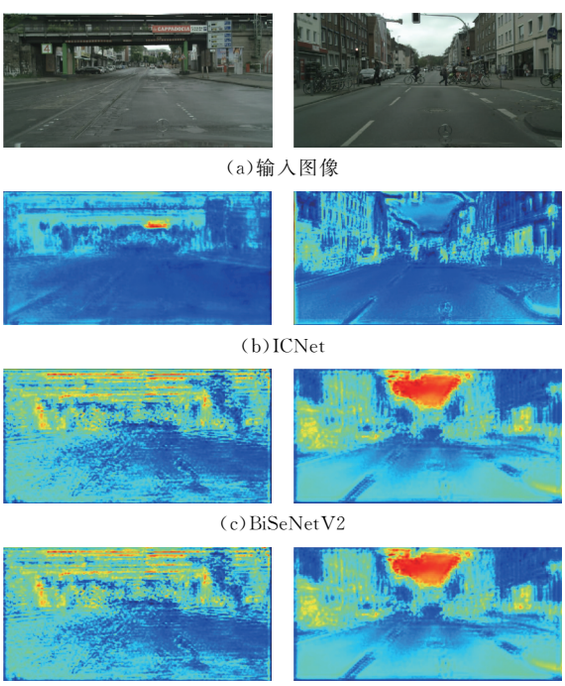
表 5　不同模块在 CityScapes 数据集上的性能分析

Tab. 5　Ablation studies of three modules on CityScapes dataset

SAF	EdgeHead	GP	mIoU/%	mPA/%
—	—	—	74.6	76.3
√	—	—	75.8	80.7
√	√	—	77.4	85.5
√	√	√	78.1	87.6

注："—"表示不添加该模块。

图 6 为不同方法在 CityScapes 数据集上的类别激活热力图，颜色由蓝到红表示关注度逐渐加深。相较于其他方法，本文方法在建筑、行人及小汽车等类别都有一定关注度，而不是仅关注像素占比较大的类别信息。结合图 4 和图 5 的可视化效果可以看出，本文提出的语义和细节双分支网络实现了更为精准的分割效果。



(a)输入图像　(b)ICNet　(c)BiSeNetV2　(d)本文方法

图 6　不同方法在 CityScapes 数据集上的类别激活热力图

Fig. 6　Category activate heatmap of proposed method with state-of-the-art methods on CityScapes dataset

4　结语

本文提出的基于双分支融合的图像实时语义分割方法，使用尺度注意融合模块和边缘增强分割头，提高了网络对目标边缘轮廓提取和捕获小目标的能力，改善了实时语义分割网络存在的边缘像素缺失问题。使用尺度注意融合模块对语义分支和细节分支不同层级融合，增强语义信息与细节特征交互性，同时提高网络对多尺度图像信息利用能力。使用边缘损失函数对细节分支引导学习，增强目标边缘特征提取能力。全局感知模块通过对特征重构后求最大似然解，增强了像素点之间的关联性。在 CityScapes 和 CamVid 数据集上的实验表明，本文方法在 mIoU 和 mPA 指标上都有显著优势，且模型参数和推理速度优于大部分方法。下一步将对模型参数量进行优化研究，以便在移动端部署应用中发挥更好的效果。

参考文献

[1] LV Q X，SUN X，CHEN C R，et al. Parallel Complement Network for Real-Time Semantic Segmentation of Road Scenes[J]. IEEE Transactions on Intelligent Transportation Systems，2022，23(5)：4432-4444.

[2] ASGARI TAGHANAKI S，ABHISHEK K，COHEN J P，et al. Deep Semantic Segmentation of Natural and Medical Images：A Review[J]. Artificial Intelligence Review，2021，54(1)：137-178.

[3] 孟俊熙，张莉，曹洋，等. 基于 Deeplab v3＋的图像语义分割算法优化研究[J]. 激光与光电子学进展，2022，59(16)：161-170.
MENG J X，ZHANG L，CAO Y，et al. Optimization of Image Semantic Segmentation Algorithms Based on Deeplab v3＋[J]. Laser & Optoelectronics Progress，2022，59(16)：161-170. (in Chinese)

[4] ORŠIC M，KREŠO 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.

[5] FALASCHETTI L，MANONI L，TURCHETTI C. A Low-Rank CNN Architecture for Real-Time Semantic Segmentation in Visual SLAM Applications[J]. IEEE Open Journal of Circuits and Systems，2022，3：115-133.

[6] 张哲晗，方薇，杜丽丽，等. 基于编码-解码卷积神经网络的遥感图像语义分割[J]. 光学学报，2020，40(3)：0310001.
ZHANG Z H，FANG W，DU L L，et al. Semantic Segmentation of Remote Sensing Image Based on Encoder-Decoder Convolutional Neural Network[J]. Acta Optica Sinica，2020，40(3)：0310001. (in Chinese)

[7] 朱磊，冯达，朱奇伟，等. 结合非对称卷积与特征蒸馏的图像超分辨率重建网络[J]. 西安工程大学学报，2024，38 (2)：93-100.
ZHU L，FENG D，ZHU Q W，et al. Image Super-Resolution Reconstruction Network Combining Asymmetric Convolution and Feature Distillation[J]. Journal of Xi'an Polytechnic University，2024，38(2)：93-100. (in Chinese)

[8] ZHUANG M X, ZHONG X Y, GU D B, et al. LRDNet: A Lightweight and Efficient Network with Refined Dual Attention Decorder for Real-Time Semantic Segmentation [J]. Neurocomputing, 2021, 459: 349-360.

[9] ZHANG R, ZHU F, LIU J Y, et al. Depth-Wise Separable Convolutions and Multi-Level Pooling for an Efficient Spatial CNN-Based Steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1138-1150.

[10] PASZKE A, CHAURASIA A, KIM S, et al. ENet: A Deep Neural Network Architecture for Real-time Semantic Segmentation[EB/OL]. (2016-06-07)[2023-09-10]. https://arxiv.org/abs/1606.02147.

[11] MEHTA S, RASTEGARI M, CASPI A, et al. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 561-580.

[12] KIM J, HEO Y S. Efficient Semantic Segmentation Using Spatio-Channel Dilated Convolutions [J]. IEEE Access, 2019, 7: 154239-154252.

[13] CHOLLET F. Xception: Deep Learning with Depthwise Separable Convolutions [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017: 1800-1807.

[14] 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). Taiwan, China: IEEE, 2019: 1860-1864.

[15] 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 (CVPR). Honolulu, HI: IEEE, 2017: 6230-6239.

[16] ZHAO H S, QI X J, SHEN X Y, et al. ICNet for Real-Time Semantic Segmentation on High-Resolution Images [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 418-434.

[17] 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.

[18] LI X, ZHONG Z S, WU J L, et al. Expectation-Maximization Attention Networks for Semantic Segmentation [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019: 9166-9175.

[19] 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.

[20] BROSTOW G J, SHOTTON J, FAUQUEUR J, et al. Segmentation and Recognition Using Structure from Motion Point Clouds [M]//Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 44-57.

[21] HU P, PERAZZI F, HEILBRON F C, et al. Real-Time Semantic Segmentation with Fast Attention[J]. IEEE Robotics and Automation Letters, 2021, 6(1): 263-270.

[22] 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.

[23] ZHANG X L, DU B C, WU Z Y, et al. LAANet: Lightweight Attention-Guided Asymmetric Network for Real-Time Semantic Segmentation [J]. Neural Computing and Applications, 2022, 34 (5): 3573-3587.

[24] FAN M Y, LAI S Q, HUANG J S, et al. Rethinking BiSeNet for Real-Time Semantic Segmentation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN: IEEE, 2021: 9711-9720.

[25] ZHANG Y H, YAO T, QIU Z F, et al. Lightweight and Progressively-Scalable Networks for Semantic Segmentation[J]. International Journal of Computer Vision, 2023, 131(8): 2153-2171.

[26] PAN H H, HONG Y D, SUN W C, et al. Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Traffic Scenes [J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(3): 3448-3460.

(编辑:徐楠楠)