DOI:10.19651/j.cnki.emt.2518767

面向城市复杂街景的实时语义分割算法

赵志兴　胡峻峰
(东北林业大学计算机与控制工程学院 哈尔滨 150040)

摘　要：实时高精度分割城市复杂街景对自动驾驶至关重要。针对现有的实时语义分割网络对高分辨率分支空间信息和细节特征捕获不足,以及高低分辨率特征融合效率低下导致信息丢失从而制约了分割精度的提升的问题,本文提出了基于多尺度部分膨胀卷积与边界协同双注意力引导融合的实时语义分割网络(MPDANet)。首先,设计多尺度部分膨胀卷积模块(MSPDC),利用并行阶梯式膨胀率的部分膨胀卷积,从不同尺度高效捕获高分辨率分支的细节特征与空间信息,解决其信息捕获不足问题。其次,构建注意力引导特征增强金字塔模块(AFPM),通过非对称池化层提取低分辨率分支的多尺度语义信息,并结合像素注意力机制进一步增强语义信息。最后,提出边界协同双注意力引导融合模块(BCDAF),通过并行通道空间注意力筛选关键语义与空间信息,抑制跨分辨率特征融合造成的信息丢失,并引入边界注意力提升目标边界分割效果。在 Cityscapes 验证集上,所提网络以 295 fps 的速度取得 78.6%的 mIoU;在 CamVid 测试集上,以 454 fps 的速度取得 77.4%的 mIoU。实验结果表明,本文所提的网络在保持实时性的同时,实现了对城市复杂街景的高精度分割。

关键词：实时语义分割;注意力机制;部分膨胀卷积;特征融合
中图分类号：TP391.41;TN0　　**文献标识码**：A　　**国家标准学科分类代码**：520.2060

Real-time semantic segmentation algorithm for complex urban street scenes

Zhao Zhixing　Hu Junfeng
(College of Computer Science and Control Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: Real-time high-precision segmentation of complex urban street scenes is crucial for autonomous driving. Aiming at the problems that existing real-time semantic segmentation networks have insufficient capture of spatial information and detailed features in high-resolution branches, as well as inefficient fusion of high and low-resolution features leading to information loss, which restricts the improvement of segmentation accuracy, this paper proposes a real-time semantic segmentation network based on multi-scale partial dilated convolution and boundary collaborative dual attention guided fusion (MPDANet). First, a Multi-Scale Partial Dilated Convolution Module (MSPDC) is designed. By using partial dilated convolutions with parallel ladder-type dilation rates, it efficiently captures detailed features andspatial information of high-resolution branches from different scales, addressing the problem of insufficient information capture. Second, an Attention-Guided Feature Pyramid Module (AFPM) is constructed. It extracts multi-scale semantic information from low-resolution branches through an asymmetric pooling layer and further enhances the semantic information by combining a pixel attention mechanism. Finally, a Boundary Collaborative Dual Attention Fusion Module (BCDAF) is proposed. It screens key semantic and spatial information through parallel channel-spatial attention, suppresses information loss caused by cross-resolution feature fusion, and introduces boundary attention to improve the segmentation effect of target boundaries. On the Cityscapes validation set, the proposed network achieves 78.6% mIoU at a speed of 295 fps; on the CamVid test set, it achieves 77.4% mIoU at a speed of 454 fps. Experimental results show that the network proposed in this paper achieves high-precision segmentation of complex urban street scenes while maintaining real-time performance.

Keywords: real-time semantic segmentation;attention mechanism;partial dilated convolution;feature fusion

0　引　　言

随着计算机视觉技术的突破性发展,语义分割作为像素级场景理解的基石任务,在自动驾驶、工业检测、医学影像分析等领域有着广泛的应用[1-2]。尽管基于深度学习的语义分割模型已经显著提升了分割精度,但其庞大的计算

收稿日期:2025-05-08

开销导致推理速度过慢的问题，制约了对推理速度敏感的领域如自动驾驶[3-4]的实际应用。

针对实时性场景的应用需求，研究者提出了实时语义分割网络的概念，其核心指标要求推理帧速率（FPS）需达到人眼视觉流畅度的最低阈值即大于 30 fps[5]。在实时语义分割网络的研究进程中，早期的探索为该领域奠定了基础。ENet[6]作为先驱之作，采用精简的单分支编码解码结构首次实现了网络的实时性，但过于精简的结构使得分割精度难以达到理想水平。BiSeNet[7]开创性地提出双分支结构，通过高分辨率分支捕获细节纹理、低分辨率分支获取语义信息的并行处理模式，经特征融合后有效提升了精度。ICNet[8]进一步采用了多分支结构将较低分辨率图像产生的特征，上采样后与较高分辨率的特征相融合。LEDNet[9]以非对称卷积替代传统卷积减少了计算开销，显著加快了推理速度。而 DABNet[10]在单分支网络架构上舍弃解码部分，对编码部分直接进行上采样实现快速推理，这些早期工作初步解决了实时性问题，但普遍存在精度与速度难以平衡的局限。

近年来国内外学者更加注重网络速度和精度的平衡，BiSeNetV2[11]使用多个小卷积核串联，代替大卷积核在增加网络推理速度的同时，获得更大的感受野显著地提高了分割精度。STDC[12]重新优化双分支结构的高分辨率分支并增设辅助分割头，在不减少推理速度的情况下提升了网络的分割精度；之后的 DDRNet[13]通过设计双分支结构各阶段高低分辨率网络的相互融合机制，并且设计了深度聚合金字塔池化模块进一步地提取了低分辨率分支的上下文语义信息显著地提升精度。但是所提出的双分支网络仍然没有解决高分辨率分支的细节和空间信息丢失的问题。Transformer 的提出为实时语义分割精度的提升提供了新的方法，RTFormer[14]将 Transformer 引入双分支网络大幅提升精度，AFFormer[15]提出高效频率自适应 Transformer 进一步的提高了网络精度，但是因为 Transformer 的复杂度，严重牺牲了网络的推理速度。

后续的研究如 PIDNet[16]为了解决高分辨分支的细节和空间信息丢失的问题，在双分支结构上引入第三条细节分支从而捕获细节信息，并且提出了边界注意力机制以提高了网络精度。ELANet[17]使用膨胀卷积代替双分支结构的普通卷积提高网络的推理速度，并设计了双注意力融合模块来融合不同分辨率的特性，以减弱跨分率融合造成的信息丢失从而提高网络分割精度。LCNnet[18]采用部分通道转换策略，通过三分支上下文聚合模块捕获多尺度的信息以提高网络的分割精度。SEDNet[19]设计了基于空洞卷积的通道注意力机制来扩展感受野，避免降维学习从而提高精度，并且加入了基于 Tversky 的细节损失函数以获得更多的细节特征。而 BENet[20]提出了边界提取和边界自适应模块以提高对物体边界的分割精度。

虽然实时语义分割网络已经取得了很大的进展，但是现有的网络忽略了直接对高分辨率分支的空间信息和细节特征的提取，其次高分辨率分支与低分辨率分支的跨分辨率融合，会造成信息丢失。针对上述问题本文提出了一种基于多尺度部分膨胀卷积和边界协同注意力机制的实时语义分割网络（MPDANet），本文工作如下：

1）提出了多尺度部分膨胀卷积模块（MSPDC）：使用部分膨胀卷积以阶梯式膨胀率并行计算，从不同的尺度提取细节特征和空间信息。

2）构建了注意力引导特征增强金字塔模块（AFPM）：使用非对称的池化层来代替传统的池化层高效的提取语义信息，并且通过注意力机制强化关键的语义信息。

3）设计了边界协同双注意力引导融合模块（BCDAF）：过并行通道空间注意力筛选关键的语义与空间信息，抑制跨分辨率融合造成的语义信息和空间信息丢失问题。同时加入边界注意力，聚焦边界细节提高网络分割准确率。

1 网络模型与结构

1.1 网络整体结构

在语义分割任务中，物体的空间信息、细节特征与语义信息是提升分割精度的关键要素。本文使用 DDRNet-23-Slim 作为基线网络，本文提出了 MPDANet 网络结构，结构如图 1 所示。

网络采用协同优化的双分支架构：共享高分辨率特征经 3 次下采样生成 1/8 分辨率特征图后，再划分高分辨率分支和低分辨率分支，高分辨率分支采用多尺度部分膨胀卷积模块（MSPDC），通过使用不同膨胀率的部分膨胀卷积，从不同的尺度提取高分辨率的空间信息和细节特征。低分辨率分支通过步幅为二的残差模块实现语义信息提取，两个分支沿用基线网络 DDRNet-23-Slim 在每个阶段进行互相融合的方式，并且在低分辨率分支的末端使用基于 DAPPM 改进的注意力引导特征增强金字塔模块（AFPM）来提取低分辨率分支的语义信息。并且为了更好的融合高分辨率的空间信息和低分辨率的语义信息，并且更好的分割网络的边缘，本文设计边界协同双注意力引导融合模块（BCDAF）进行高效的融合最终输出预测结果。此外，在训练阶段在保留 DDRNet 的辅助分割头，并且受到 PIDNet 启发使用和其相同的可丢弃的边界分割头和边界损失函数，从而在不增加推理计算量的前提下提升网络对复杂边界的建模能力，因此 MPDANet 最终的损失函数可以表示为如下公式：

$$Loss = L_d + L_b \tag{1}$$

式中：L_d 为 DDRNet 网络的原本得损失函数，L_b 为引入的边界损失函数。

1.2 多尺度部分膨胀卷积模块

针对实时语义分割高分辨率分支对空间信息和细节特征捕获不足的问题。本文提出了多尺度部分膨胀卷积模块（MSPDC）如图 2 所示。

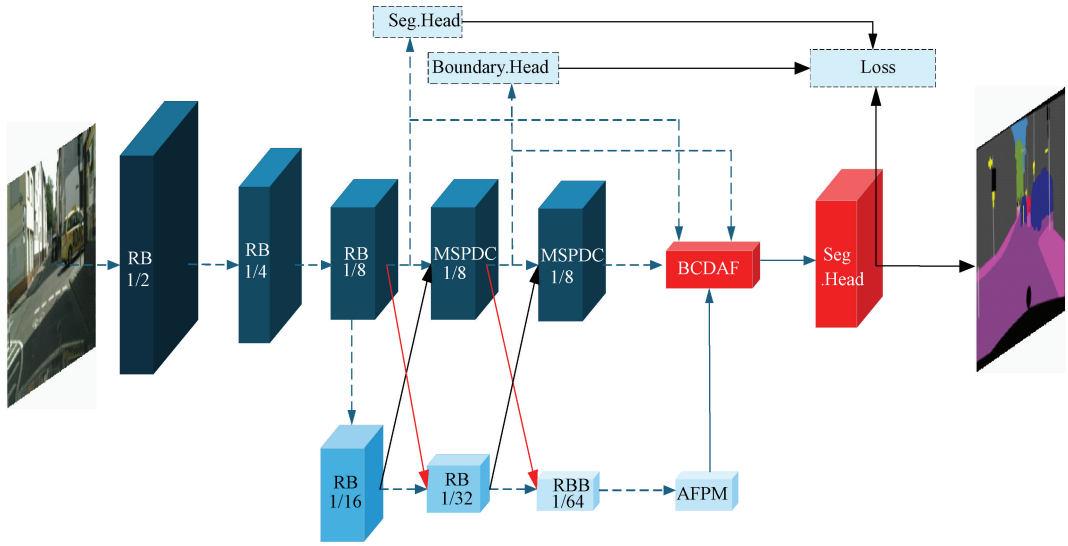


图 1　基于多尺度部分膨胀卷积与边界协同双注意力引导融合的实时语义分割网络

Fig. 1　Real-time semantic segmentation network with multi-scale partial dilated convolution and boundary collaborative dual-attention guided fusion

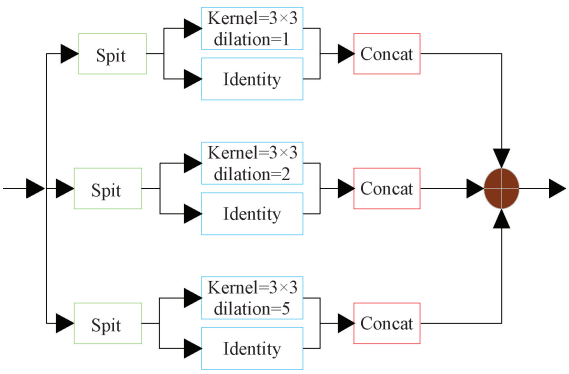


图 2　多尺度部分膨胀卷积

Fig. 2　Multi-scale partial dilated convolution

本文首先提出了通道解耦—聚合的部分膨胀卷积其核心是结合部分卷积[21]的轻量化设计和膨胀卷积的多尺度感知能力实现精度和速度平衡。首先对于输入的的特征 $\boldsymbol{X}$ 按照比例 α 解耦为膨胀卷积分支和恒等分支，其中 α 根据实验取 0.5。其中膨胀卷积分支 $\boldsymbol{X}_{conv}$ 有效通道为原通道的 α 倍，并根据膨胀率执行膨胀卷积操作来捕获不同尺度的信息。剩余的通道不做任何的处理保留为恒等分支 $\boldsymbol{X}_{id}$。然后两个分支经过通道拼接融合输出特征，部分膨胀卷积可以表示为如下公式：

$$\boldsymbol{X}_{id},\boldsymbol{X}_{conv}=\text{split}(\boldsymbol{X})=\boldsymbol{X}_{:,:\alpha\times C},\boldsymbol{X}_{:,\alpha\times C:C} \tag{2}$$

$$\boldsymbol{X}_{dconv}=\text{Conv}_{3\times3}(\boldsymbol{X}_{conv};dilation=d) \tag{3}$$

$$\text{PDconv}=\text{Concat}(\boldsymbol{X}_{dconv},\boldsymbol{X}_{id}) \tag{4}$$

式中：$\boldsymbol{X}_{id}$ 为解耦后的恒等部分，$\boldsymbol{X}_{conv}$ 为解耦后的膨胀卷积部分，$\boldsymbol{X}_{dconv}$ 为对分支进行膨胀卷积操作，PDconv 是部分膨胀卷积。

对于输入的 $\boldsymbol{X}$ 通过 3 组并行的部分膨胀卷积来从不同的尺度捕获细节特征和空间信息，并且在每个分支后使用批量化归一和 ReLU 激活函数进一步提取特征。使用膨胀率为 1 感受野为 3×3 膨胀卷积和恒等分支捕获物体的细节特征，使用膨胀率为 2 的感受野为 5×5 部分膨胀卷积捕获中程的空间信息，使用膨胀率为 5 的感受野 11×11 的部分膨胀卷积捕获全局的空间信息。然后 3 个分支相加后得到多尺度的输出 $\boldsymbol{X}_m$。上述可以表示为以下公式：

$$\boldsymbol{X}_m=\sum_{d\in\{1,2,5\}}\text{ReLU}(\text{BN}(\text{PDconv}_{3\times3}^{dilation=d}(\boldsymbol{X}))) \tag{5}$$

式中：PDconv 是部分膨胀卷积，BN 是批量化归一，d 是膨胀率。

1.3　注意力引导特征增强金字塔模块

经过多次下采样的操作后，图片含有丰富的语义信息，为了获得更多的语义信息通常使用特征金字塔来捕获多尺度的语义信息以增强图片的语义信息。本文提出了一种基于深度聚合池化金字塔模块[10]改进的注意力引导特征增强金字塔模块（AFPM），以增强低分辨率的语义信息。

如图 3 所示本文提出的 AFPM 模块，首先使用非对称池化特征金字塔用来提取多尺度的语义信息，之后使用像素注意力机制来进一步增强语义信息，首先将输入的特征 Fin 经过 5 个并行的多尺度分支来提取多尺度的语义信息，基准分支直接对 Fin 使用 1×1 的卷积提取原始的语义信息。全局上下文分支使用全局平均池化捕获全局上下文信息。

多尺度非对称分支组，采用轴向分解策略将大核平均池化核的轻量化，本文设计了 5×5、9×9、17×17 三种非对称池化捕获不同尺度的语义信息。每条分支在提取信息后使用批量归一化和 ReLU 激活函数进一步的提取全局

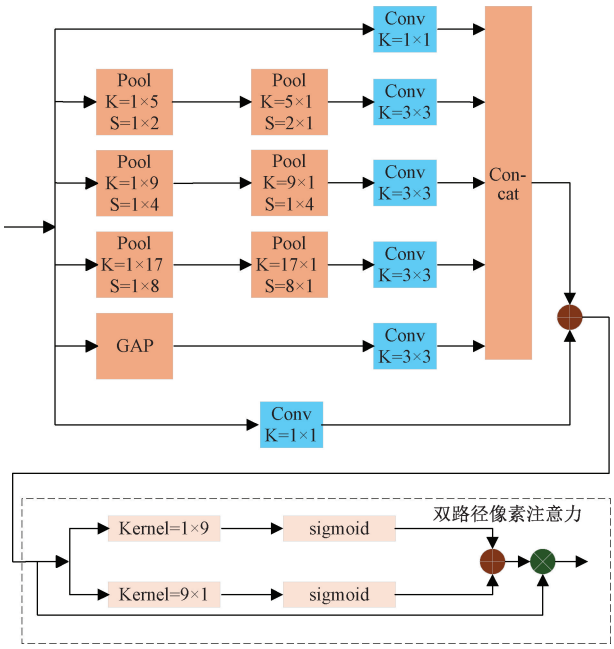


图 3 注意力引导特征增强金字塔模块

Fig. 3 Attention-guided feature pyramid module

的语义信息，之后将 5 条分支沿通道轴拼接，实现多尺度聚合。上述可以表示为以下公式：

$$\boldsymbol{F}_{base} = \mathrm{ReLU}(\mathrm{BN}(F_c)) \tag{6}$$

$$\boldsymbol{F}_{global} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathrm{GPA}(\boldsymbol{F}_c)))) \tag{7}$$

$$\boldsymbol{F}_k = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathrm{Pool}_{1\times k}(\mathrm{Pool}_{k\times1}(\boldsymbol{F}_c))))) \tag{8}$$

$$\boldsymbol{F}_{fusion} = \mathrm{Concat}(\boldsymbol{F}_{base}, \boldsymbol{F}_{global}, \boldsymbol{F}_{5\times5}, \boldsymbol{F}_{9\times9}, \boldsymbol{F}_{17\times17}) \tag{9}$$

式中：$\boldsymbol{F}_{base}$ 是基准分支，$\boldsymbol{F}_{global}$ 是全局上下文分支，$\boldsymbol{F}_k$ 是多尺度非对称分支，Concat 是沿通道拼接，$\boldsymbol{F}_{fusion}$ 是聚合后的特征，Conv 是卷积，Pool 是池化，BN 是批量化归一。

本文提出了双路径像素注意力，采用 1×9 卷积和 9×1 卷积核分别沿图像宽度和高度方向建模像素的上下文关系，并且通过矩阵加法融合双轴像素注意力权重，得到像素注意力图以增强像素的语义信息，最后使用残差连接增强目标区域的语义信息。上述可以表示为以下公式：

$$\boldsymbol{A}_P = \sigma(\mathrm{Conv}_{1\times9}(\boldsymbol{F}_{fusion}) + \sigma(\mathrm{Conv}_{9\times1}(\boldsymbol{F}_{fusion}) \tag{10}$$

$$\boldsymbol{F}_{out} = \boldsymbol{A}_P \odot \boldsymbol{F}_{fusion} \tag{11}$$

式中：$\boldsymbol{A}_p$ 表示像素注意力权重矩阵，σ 是 Sigmoid 激活函数，是 $\boldsymbol{F}_{out}$ 表示输出。

1.4 边界协同双注意力引导融合模块

高分辨率保留了空间信息和细节特征，低分辨率蕴含了丰富的语义信息，传统的低效的特征融合会导致信息丢失的问题。为了解决信息丢失的问题，本文提出了边界协同双注意力引导融合模块(BCDAF)如图 4 所示，首先通过并行通道空间注意力[22-24]，在通道维度和空间维度筛选关键语义与空间信息，抑制跨分辨率融合造成的信息丢失。同时加入边界注意力，提取网络的边界特征，提高网络对边界的分割能力提高分割精度。

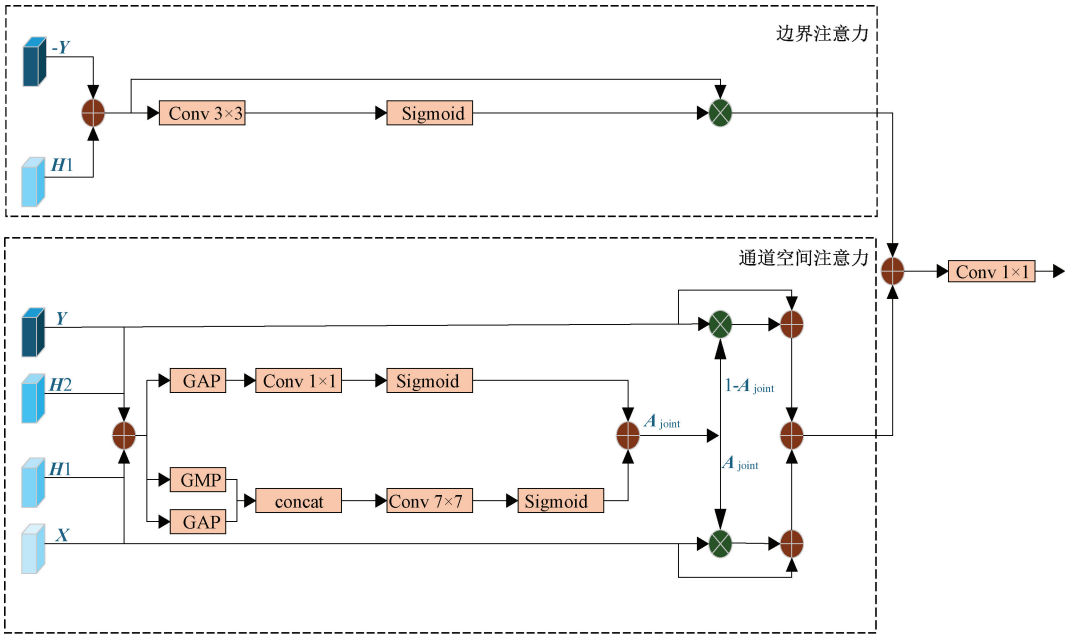


图 4 边界协同双注意力引导融合模块

Fig. 4 Boundary collaborative dual-attention guided fusion module

首先对于通道空间注意力，本文设计了一个多尺度特征聚合层，将 4 个不同阶段的特征直接相加得到聚合后的 $\boldsymbol{F}_{fusion}$，$\boldsymbol{F}_{fusion}$ 此时包含了不同阶段的浅层空间细节信息和深层语义信息，为后续注意力机制提供多特征信息，上述可以表示为以下公式：

$$\boldsymbol{F}_{fusion} = \boldsymbol{X} + \boldsymbol{Y} + \boldsymbol{H}1 + \boldsymbol{H}2 \tag{12}$$

式中：$\boldsymbol{Y}$ 是低分辨率分支的输出，$\boldsymbol{X}$ 是高分辨率分支的输出，$\boldsymbol{H}1$ 和 $\boldsymbol{H}2$ 是高分辨率的不同阶段。

对于通道注意力分支将融合特征 $\boldsymbol{F}_{fusion}$ 执行全局平均化获取通道信息，经过卷积与 Sigmoid 激活后生成通道注意力权重矩阵 $\boldsymbol{A}_C=\boldsymbol{R}^{C\times1\times1}$。

对于空间注意力分支，将 $\boldsymbol{F}_{fusion}$ 沿通道方向上执行最大池化和平均池化生成 $\boldsymbol{F}_{avg}=\boldsymbol{R}^{1\times H\times W}$ 和 $\boldsymbol{F}_{max}=\boldsymbol{R}^{1\times H\times W}$ 并在第 2 个维度上进行拼接得到 $\boldsymbol{F}_s=\boldsymbol{R}^{2\times H\times W}$ 积后通过 7×7 的卷积学习空间权重，并最终压缩为空间注意力权重矩阵 $\boldsymbol{A}_s=\boldsymbol{R}^{1\times H\times W}$。

最后将通道注意力权重矩阵和空间注意力权重矩阵进行自适应加权，生成联合注意力权重矩阵。并对高分辨率和低分辨率进行差异化增强，以强化高分辨率的空间细节信息并且补偿低分辨率的语义信息。最后将加强后的特征进行逐元素相加得到双注意力的融合特征输出。上述可以表示为以下公式：

$$\boldsymbol{A}_{joint}=\boldsymbol{A}_C+\boldsymbol{A}_s \tag{13}$$

$$\boldsymbol{F}_{high}=\boldsymbol{X}\odot\boldsymbol{A}_{joint}+\boldsymbol{X} \tag{14}$$

$$\boldsymbol{F}_{low}=\boldsymbol{Y}\odot(1-\boldsymbol{A}_{joint})+\boldsymbol{Y} \tag{15}$$

$$\boldsymbol{F}_{fuse}=\boldsymbol{F}_{high}+\boldsymbol{F}_{low} \tag{16}$$

式中：$\boldsymbol{A}_C$ 是通道注意力权重矩阵，$\boldsymbol{A}_s$ 是空间注意力权重矩阵，$\boldsymbol{A}_{joint}$ 联合注意力权重矩阵，$\boldsymbol{F}_{fuse}$ 是双注意力的融合特征输出。

经过多次下采样的图片边界模糊，而未经下采样的高分辨率图片保留了原始的边界特征[25]，通过高分辨率的初始特征 $\boldsymbol{H}1$ 与低分辨率特征的 $\boldsymbol{Y}$ 的差值捕获边界信息，并通过 3×3 的卷积进一步的提取边界信息，最后通过 Sigmoid 函数生成边界注意力权重，边界注意力权重在与高分辨率 $\boldsymbol{H}1$ 相乘后得到增强的边界特征，最后将边界特征以残差的形式与融合特征进行相加得到最终的输出。上述可以表示为以下公式：

$$\boldsymbol{A}_{edge}=\sigma(\mathrm{Conv}_{3\times3}(\boldsymbol{H}_1-\boldsymbol{Y})) \tag{17}$$

$$\boldsymbol{F}_{edge}=(\boldsymbol{H}_1-\boldsymbol{Y})\odot\boldsymbol{A}_{edge} \tag{18}$$

$$\boldsymbol{F}_{output}=\boldsymbol{F}_{edge}+\boldsymbol{F}_{fuse} \tag{19}$$

式中：$\boldsymbol{A}_{edge}$ 是边界注意力权重矩阵，σ 是 Sigmoid 激活函数，$\boldsymbol{F}_{output}$ 是特征输出。

2　实验结果与分析

2.1　数据集

本文采用 Cityscapes 和 CamVid 两个面向城市复杂街景的公共数据集来验证模型的有效性，Cityscapes 数据集是在自动驾驶领域常用的数据集，包含了 5 000 张 2 048×1 024 分辨率的精细标注的图像，这些图片都是从汽车驾驶员的角度对城市复杂街景进行分割的数据集。在这个数据集中，标签有 30 个类别，但是用于语义分割的类别只有 19 个，5 000 张图片中用于训练的有 2 975 张，用于测试的有 1 525 张，用于验证的有 500 张。

CamVid 数据集的获取方式与 Cityscapes 数据集相似，但是该数据集用于语义分割的类别标注仅有 11 个，并且只有 701 张 960×720 分辨率的图片，其中用于训练的图片有 367 张，用于测试的有 233 张，用于验证的有 101 张。

2.2　评价指标

评价实时语义分割任务的性能，主要是精度和速度两个方面。其中评价精度的指标是平均交并比（mIoU），它为交集与并集之比计算为如下公式：

$$\mathrm{mIoU}=\frac{1}{K+1}\sum_{i=0}^{k}\frac{\boldsymbol{p}_{ij}}{\sum_{j=0}^{k}\boldsymbol{p}_{ij}+\sum_{j=0}^{k}\boldsymbol{p}_{ji}-\boldsymbol{p}_{ii}} \tag{20}$$

式中：K 表示前景对象的个数 $\boldsymbol{P}_{ij}$ 表示原本属于第 i 类，却分类到第 j 类像素的数量。

检测速度使用每秒处理了多少张图片即帧速率（FPS），用于评价算法速度。公式如下：

$$\mathrm{FPS}=\frac{N}{\sum_{j=1}^{N}T_j} \tag{21}$$

式中：N 表示图像数量，$\boldsymbol{T}_j$ 表示算法处理第 j 幅图像的时间

2.3　实验环境

实验所用配置：操作系统：Ubuntu 22.04；GPU：NVIDIA RTX4090 显存 24 GB；CPU：i7-13700K。模型训练环境：Python3.8，Pytorch2.4。同时在 ImageNet 数据集上训练得到预训练权重来加速模型的收敛和学习过程。对于 Cityscapes 数据集实验设置具体如下：实验的批量大小为 8，最大训练 Epoch 为 500，使用 OHEM[26] 算法优化模型，且在训练阶段随机裁剪图像至 1 024×1 024，并且使用随机缩放和随机反转来增强数据，采用具有动量的随机梯度下降法，动量动态调整学习率公式如下：

$$lr=lr_{\mathrm{lrint}}\times\left(1-\frac{\mathrm{iter}}{\mathrm{max\_iter}}\right)^{\mathrm{power}} \tag{22}$$

式中：lr_{lrint} 为初始学习率设置为 0.01，iter 为当前迭代次数，max_iter 为最大迭代次数，power 为动量设置为 0.9。

在 CamVid 数据集的实验设置中不对图片在训练阶段进行裁剪，直接使用原始 960×720 分辨率的图像。其余与 Cityscapes 数据集的实验设置相同。

2.4　消融实验

为了验证各模块的有效性，本文基于 Cityscapes 数据集设计了一系列的消融实验，验证模块的有效性。

实验结果如表 1 所示，本文以 DDRNet-23-slim 作为基线网络。首先在训练阶段在原网络的基础上加入边界损失函数在不降低推理速度的同时提升了 0.3% 的 mIoU。然后将原高分辨率分支残差模块替换为 MSPDC 模块，模型在保持 354 fps 实时速度的同时，mIoU 提升 0.4% 至 77.8%。尽管推理速度下降了 39 fps 但是其速度仍远高于实时语义分割 30 fps 的要求。证明了 MSPDC 能从不同的尺度有效地捕获空间信息和细节特征。在此基础上，将

原网络的低分辨率的 DAPPM 模块替换为 AFPM 模块，通过像素注意力机制进一步挖掘低分辨率的语义信息，使精度进一步增长 0.3%至 78.1%的精度，此时的速度略微下降，实验证明 AFPM 模块能够在略微降低速度的情况下，更好的提取低分辨率分支的语义信息。最后引入 BCDAF 模块引导低分辨率分支和高分辨率分支进行高效互补融合，最终得到的 MPDANet 网络。网络能够在 295 fps 的高效推理速度下 mIoU 达到 78.6%。

表 1 消融实验

Table 1 Ablation experiment

基线网络	边界损失	MSPDC	AFPM	BCDAF	mIoU/%	FPS
√					77.1	393
√	√				77.4	393
√	√	√			77.8	354
√	√	√	√		78.1	345
√	√	√	√	√	78.6	295

为深入探究 AFPM 模块在低分辨率分支中不同位置的影响，本文设计了消融实验。在保持网络其余部分不变的前提下，分别将 AFPM 模块放置在生成 1/32 和 1/64 特征图的阶段之前，并将这两个位置对应的模块分别记为 AFPM(1)和 AFPM(2)。实验结果如表 2 所示，实验结果表明：AFPM(1)和 AFPM(2)模块由于其所处位置处理的特征图尺寸较大且网络深度较浅的限制，模块的推理精度和速度均受到显著制约。相比之下，将 AFPM 模块置于低分辨率分支的末尾时，模型取得了最佳的推理速度和精度平衡。

表 2 AFPM 模块的消融实验

Table 2 Ablation experiments of AFPM module

方法	mIoU/%	FPS
AFPM(1)	76.1	256
AFPM(2)	77.8	269
AFPM	78.6	295

为更深入地探究 BCDAF 模块的作用机制和 MSPDC 模块捕获信息的能力，本文设计了使用中间阶段的高分辨率特征图替代最终输出的高分辨率特征图，将其与低分辨率特征图进行融合。具体而言，分别选取第 1 个和第 2 个 MSPDC 模块输入前的高分辨率特征图进行融合分别记为 BCDAF(1)和 BCDAF(2)。实验结果如表 3 所示：相较于使用最终输出的高分辨率特征图，使用中间阶段的特征图进行融合在推理速度相似的同时导致精度下降。这主要源于中间特征图所包含的细节特征和空间信息不足，影响了融合效果，并且经过第一个 MSPDC 模块提取细节后使用 BCDAF(2)进行融合，网络精度显著提升。最终结果表明，MSPDC 模块能够有效的提取细节特征和空间信息。BCDAF 模块使用高分辨率分支末尾输出的高分辨率特征图进行融合，能够获得最高的精度提升。

表 3 BCDAF 模块的消融实验

Table 3 Ablation experiments of BCDAF module

方法	mIoU/%	FPS
BCDAF(1)	75.3	298
BCDAF(2)	77.9	294
BCDAF	78.6	295

另外，在 MSPDC 模块中为了探究比例系数 α 对模型性能地影响，在保持网络整体结构不变的情况下，系统测试了不同 α 值对模型精度(mIoU)和帧速率(FPS)的影响。

结果如表 4 所示，当 α 取 1 时所有的通道均执行膨胀卷积，由于此时缺少了恒等分支保留的原始尺度信息，其精度为 78.4%的 mIoU 与 227 fps 的推理速度均未达最优，表明了单纯的依靠膨胀卷积会制约计算速度和多尺度特征的表达。当 α 取 0.8 时由于引入了原始尺度其精度增长到 78.8%的 mIoU，并且速度提升为 253 fps，验证了模型多尺度特征的有效性和部分膨胀卷积的轻量化效果。并且随着 α 的降低，速度 FPS 的提升变小。并且精度在 $\alpha>0.5$ 时波动在 0.2%，当 $\alpha<0.5$ 时精度呈现加速衰减趋势。综合实时语义分割网络对于精度和速度的需求最终选取比例系数 α 取 0.5 为最优参数，该参数能够在保持较高精度 78.6%的 mIoU 的同时实现 295 的 FPS 的实时推理速度。

表 4 不同比例系数对算法性能的影响

Table 4 The impact of different scaling factors on algorithm performance

α	mIoU/%	FPS
1.0	78.4	227
0.8	78.8	253
0.6	78.6	283
0.5	78.6	295
0.4	78.2	298
0.3	77.7	303

此外为了进一步验证 AFPM 模块和其他特征金字塔模块的有效性，在 Cityscapes 数据集设置了对比实验。模块的整体网络结构不变如表 5 所示。表中 DAPPM 是 DDRNet 网络用来提取低分辨率语义信息的特征金字塔，通过不同的大核池化层来提取不同尺度的语义信息，并且通过逐级的大核相加来更进一步的提取语义信息，但是复杂的方式导致了速度降低。而 PAPPM 是 PIDNet 网络用来提取低分辨率语义信息的特征金字塔，该模块通过将 DAPPM 的大核链接方式改为并行链接，并降低通道的方

式来提高速度，但是这会降低网络的精度。MSPA[27]是一种高效的特征金字塔，通过层次幻影卷积和金字塔自校准注意力机制来提取多尺度的信息特征。AFPM 模块与上述的 3 个模块相比在取得最高精度的同时，保持了有竞争力的速度 295 fps。

表 5　不同特征金字塔对算法性能的影响

Table 5　The impact of different feature pyramids on algorithm performance

方法	mIoU/%	FPS
AFPM	78.6	295
DAPPM	78.2	303
PAPPM	77.9	307
MSPA	78.4	279

2.5　模糊与光照干扰下的网络鲁棒性分析

为验证网络的鲁棒性，本文在相同干扰下与基线网络进行了对比实验。具体为在 Cityscapes 验证集上施加高斯模糊和光照干扰，其中高斯模糊卷积核 3×3、5×5、7×7 分别代表了轻度模糊，中度模糊和高度模糊。光照为原来亮度的 1.5 倍为强光环境和光照为原来的 0.5 倍为暗光环境。

分别在上述干扰情况下测试网络的 mIoU，具体的结果如表 6 所示：验证了 MPDANet 网络在真实干扰环境中有着良好的鲁棒性。并且在光照变化场景下，其性能显著优于基线网络。即便面对中高度模糊的干扰，仍有着比基线网络更高的准确度。

表 6　不同干扰类型对算法性能的影响

Table 6　The impact of different interference types on algorithm performance

干扰类型	基线网络的 mIoU/%	MPDANet 的 mIoU/%
无干扰	77.1	78.6
轻度模糊	77.0	78.4
中度模糊	76.5	77.8
高度模糊	75.9	76.2
强光	72.6	75.2
暗光	74.2	76.2

2.6　不同算法在 Cityscapes 数据集上的对比

为验证 MPDANet 的有效性，本文在 Cityscapes 数据集上，从分割精度和分割速度这两个方面，与一些代表性和优秀的算法进行比。特别地，针对近三年发表的优秀分割模型，在本文的实验条件下进行重新验证，确保比较实验的公平性。实验结果如表 7 所示。相比于经典的网络 BiSeNetV2 和 HyperSeg-S[28]，MPDANet 网络的分割精度分别高出 3.9%和 0.3%mIoU，虽然 AFFormer 网络的精准度与本网络相差不大，但是因为其引入了 Transformer 导致其在性能更强的平台获得了远低于本网络的推理速度。相比于近几年的优秀模型，MPDANet 在测试集上，达到了 78.4%的 mIoU 优于 PIDNet-S 和 DDRNet-23-Slim。仅低于 PIDNet-M 和 DDRNet-23 但相比这两个网络 MPDANet 的推理速度分别高出 175 fps 和 138 fps。同时

表 7　不同算法在 Cityscapes 数据集上的对比分析

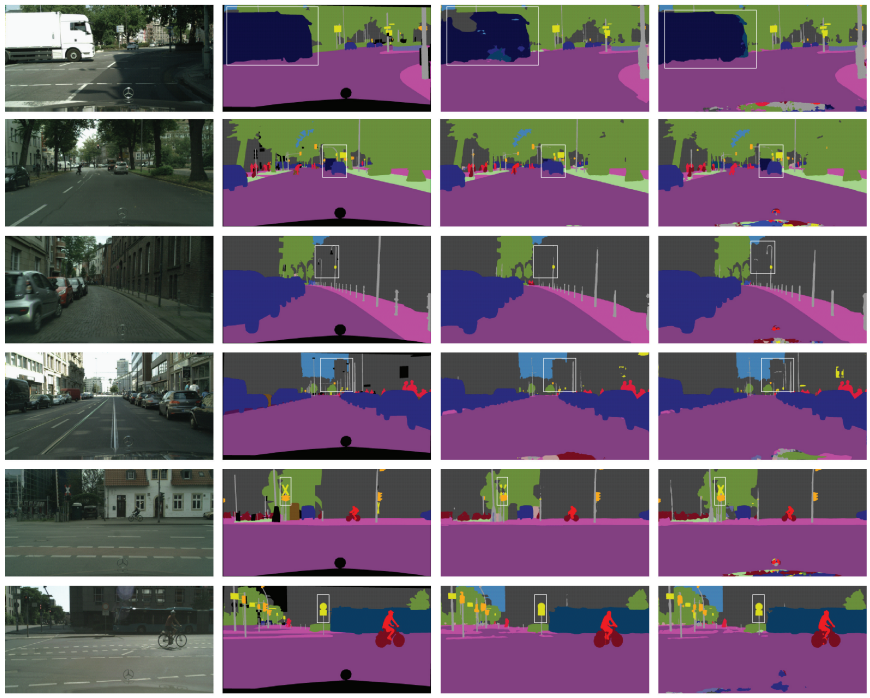
Table 7　Comparative analysis of different algorithms on the Cityscapes dataset

方法	输入图像尺寸	参数量/MB	GPU	mIoU/%		FPS
				验证集	测试集	
ICNet	2 048×1 024	26.50	TitanX	—	69.5	30.0
BiSeNet	1 536×786	5.80	GTX1080Ti	69.0	68.4	105
BiSeNetV2	1 024×512	—	GTX1080Ti	73.4	72.6	156
HyperSeg-S	1 536×786	10.20	RTX 1080Ti	78.2	78.1	16.2
STDC1-Seg75	1 536×786	14.20	GTX 1080Ti	75.8	75.3	126
STDC2-Seg75	1 536×786	22.20	GTX 1080Ti	77.0	76.8	97
RTFormer-S	2 048×1 024	4.80	RTX 2080Ti	76.3	75.4	110
AFFormer-base	2 048×1 024	3.00	V100	78.7	—	22
ELANet	1 024×512	0.76	RTX 4090	75.6	75.4	298
LCNet	1 024×512	0.51	RTX 4090	73.0	73.3	431
PIDNet-S	2 048×1 024	7.60	RTX 4090	78.4	78.2	307
PIDNet-M	2 048×1 024	34.40	RTX 4090	80.1	80.1	120
DDRNet-23-Slim	2 048×1 024	5.71	RTX 4090	77.1	77.2	393
DDRNet-23	2 048×1 024	20.10	RTX 4090	79.2	79.4	157
MPDANet	2 048×1 024	8.05	RTX 4090	78.6	78.4	295

注：“—”表示原文章未给出相关数据

以 5.6%的 mIoU 的优势领先采用图像裁剪加速策略的 LCNet。实验表明 MPDANet 网络在分割精度优于大部分的网络同时，其推理速度也高于大部分网络。在精度和速度之间实现了良好的平衡。

本文对基线网络 DDRNet-23-Slim 和本文设计的网络 MPDANet 进行了可视化结果分析，以更好的展现本文的优越性。可视化结果如图 5 所示，第 1、2 行显示在卡车等大型的目标上 DDRNet-23-Slim 因为空间信息的丢失导致车身部分，分割为轿车和细柱的错误，而 MPDANet 通过对高分辨的空间信息的捕获，更好的捕获了卡车的轮廓，实现了更加良好的分割。第 2、3 行展示了对于较远的细杆和路灯 DDRNet-23-Slim 网络受限于细节特征捕获的不足而未能检测出来，而 MPDANet 凭借其对细节的捕获，定位了细杆的位置并进行分割。剩余几行显示了 DDRNet-23-Slim 交通指示牌的边缘分割存在模糊与锯齿，而相对应的 MPDANet 由于对边界信息的增强，实现了对边界的良好分割。可视化结果直观地验证了 MPDANet 网络的优越性。



(a) 输入图片 (a) Input image　(b) 真实标签 (b) Ground truth label　(c) 基线网络 (c) Baseline network　(d) 本文网络 (d) Proposed network

图 5　可视化结果图

Fig. 5　Visualization result plots

2.7　不同算法在 CamVid 数据集上的对比

本文在 CamVid 数据集上进行对比，其中为了更加公平的进行对比，本文对 DDRNet-23-Slim 网络在本文实验平台进行了复现，结果如表 8 所示。

表 8　不同算法在 CamVid 数据集上的对比分析

Table 8　Comparative analysis of different algorithms on the CamVid dataset

方法	GPU	mIoU/%	FPS
ENet	TitanX	68.3	61
BiSeNet	GTX1080Ti	68.7	116
BiSeNetV2	GTX1080Ti	72.4	124
MSFNet	RTX 2080Ti	75.4	91
STDC1-Seg75	GTX1080Ti	73.0	197
STDC2-Seg75	GTX1080Ti	73.9	152
DDRNet-23-Slim	RTX 4090	74.7	600
MPDANet	RTX 4090	77.4	454

MPDANet 以 77.4%的 mIoU 的分割精度和 454 fps 的推理速度，展现了卓越的分割精度和速度的平衡。相较于网络 MSFNet[29] 和 STDC1-Seg75 分割精度分别高 2.0%和 3.5% mIoU。虽然推理速度低于基线网络 DDRNet-23-Slim，但是分割精度提高 2.7%mIoU。体现了本文网络能够在维持高精度的同时，仍能保持有竞争力的推理速度

3　结　　论

针对现有的实时语义分割网络在高分辨率分支对空间信息和细节特征提取不足和跨分辨率融合低下的问题，本文提出了一种面向复杂街景的实时语义分割网络。通过在高分辨率分支引入多尺度部分膨胀卷积模块捕获高分辨率的空间信息和细节特征。同时引入注意力引导特征增强金字塔模块更好的从低分辨率分支聚合多尺度的

语义信息。并且最后使用边界协同双注意力引导融合模块,高效的融合低分辨率分支和高分辨率分支以提高网络的分割精度。

通过在在 Cityscapes 和 CamVid 数据集上的实验表明。本文方法实现了实时语义分割对分割精度和速度的综合最优。本文提出的网络在实现分割精度上升的同时,推理速度仍有优化的空间。未来针对该问题可以进一步探究更加轻量化的注意力机制和高分辨率空间信息和细节特征提取的方法以提高网络的速度。

参考文献

[1] 郑凯,李建胜.基于深度神经网络的图像语义分割综述[J].测绘与空间地理信息,2020,43(10):119-125.
ZHENG K, LI J SH. A review on image semantic segmentation based on deep neural networks [J]. Surveying and Mapping and Spatial Geographic Information, 2020,43(10):119-125.

[2] 李利荣,丁江,梅冰,等.基于像素注意力特征融合的城市街景语义分割算法研究[J].电子测量技术,2023,46(20):184-190.
LI L R, DING J, MEI B, et al. A study on a pixel attention feature fusion-based urban street view semantic segmentation algorithm [J]. Electronic Measurement Technology,2023,46(20):184-190.

[3] 徐晓龙,俞晓春,何晓佳,等.基于改进 U-Net 的街景图像语义分割方法[J].电子测量技术,2023,46(9):117-123.
XU X L, YU X CH, HE X J, et al. A semantic segmentation method for street view images based on improved U-Net [J]. Electronic Measurement Technology, 2023,46(9):117-123.

[4] 周勇,刘泓滨,侯亚东.复杂城市交通场景下的自动驾驶语义分割方法[J].电子测量与仪器学报,2024,38(4):241-247.
ZHOU Y, LIU H B, HOU Y D. A semantic segmentation method for autonomous driving in complex urban traffic scenarios [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(4):241-247.

[5] 高常鑫,徐正泽,吴东岳,等.深度学习实时语义分割综述[J].中国图象图形学报,2024,29(5):1119-1145.
GAO CH X, XU ZH Z, WU D Y, et al. Deep learning-based real-time semantic segmentation: A survey[J]. Journal of Chinese Image and Graphics, 2024,29(5):1119-1145.

[6] PASZKE A, CHAURAIA A, KIM S, et al. Enet: A deep neural network architecture for real-time semantic segmentation [J]. ArXiv preprint arXiv: 1606.02147, 2016.

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

[8] ZHAO H, QI X, SHEN X, et al. Icnet for real-time semantic segmentation on high-resolution images[C]. European Conference on Computer Vision (ECCV), 2018: 405-420.

[9] 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.

[10] LI G, YUN I, KIM J, et al. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation [J]. ArXiv preprint arXiv: 1907.11357, 2019.

[11] 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: 3051-3068.

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

[13] PAN H, HONG Y, SUN W, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 24(3): 3448-3460.

[14] WANG J, GOU C, WU Q, et al. Rtformer: Efficient design for real-time semantic segmentation with transformer [J]. Advances in Neural Information Processing Systems, 2022, 35: 7423-7436.

[15] DONG B, WANG P, WANG F. Head-free lightweight semantic segmentation with linear transformer [C]. AAAI Conference on Artificial Intelligence,2023, 37(1): 516-524.

[16] XU J, XIONG Z, BHATTACHARYYA S P. PIDNet: A real-time semantic segmentation network inspired by PID controllers[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 19529-19539.

[17] YI Q, DAI G, SHI M, et al. Elanet: Effective lightweight attention-guided network for real-time semantic segmentation[J]. Neural Processing Letters, 2023, 55(5): 6425-6442.

[18] SHI M, LIN S, YI Q, et al. Lightweight context-

aware network using partial-channel transformation for real-time semantic segmentation [J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(7): 7401-7416.

[19] MA S, ZHAO Z, HOU Z, et al. SEDNet: Real-time semantic segmentation algorithm based on STDC [J]. International Journal of Intelligent Systems, 2025, 2025(1): 8243407.

[20] LEI X, CHEN Z, YU Z, et al. BENet: Boundary-enhanced network for real-time semantic segmentation[J]. The Visual Computer, 2025, 41(1): 229-241.

[21] CHEN J, KAO S, HE H, et al. Run, don't walk: Chasing higher FLOPS for faster neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12021-12031.

[22] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713-13722.

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

[24] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.

[25] GAO S, ZHANG P, YAN T, et al. Multi-scale and detail-enhanced segment anything model for salient object detection [C]. 32nd ACM International Conference on Multimedia, 2024: 9894-9903.

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

[27] YU Y, ZHANG Y, CHENG Z, et al. Multi-scale spatial pyramid attention mechanism for image recognition: An effective approach[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108261.

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

[29] SI H, ZHANG Z, LYU F, et al. Real-time semantic segmentation via multiply spatial fusion network[J]. ArXiv preprint arXiv:1911.07217, 2019.

作者简介

赵志兴，硕士研究生，主要研究方向为计算机视觉与图像处理。

E-mail:15716978513@163.com

胡峻峰(通信作者)，博士，副教授，硕士生导师，主要研究方向为机器视觉、图形处理、模式识别与智能控制。

E-mail:nefuhjf@126.com