Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(1), 277-287
Published Online January 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.151027

Hans汉斯

基于RGB-D联合语义分割和边界检测的跨模态融合网络

李超杰

西华大学汽车与交通学院，四川 成都

收稿日期：2025年12月9日；录用日期：2026年1月5日；发布日期：2026年1月15日

摘　要

语义分割和边界检测是自动驾驶汽车实现准确环境感知的两大关键任务，然而现有研究多将两者视为独立的任务，或者将语义和边界特征进行简单堆叠，忽略了两者之间的内在联系，缺乏对物体与边界间依赖关系的显式建模，易导致在颜色相近区域出现边界模糊、类别混淆。为此，本文提出了一个跨模态联合感知网络，通过引入深度信息Depth为RGB图像提供几何先验，并建立了一种动态边界引导机制，利用边界信息与几何结构共同指导语义分割过程。具体来说，网络采用双分支结构分别捕获RGB信息、Depth信息并提出了一个边界引导的跨模态融合模块BGCF (Boundary-Guided Cross-modality Fusion Module)，通过动态融合不同层级的RGB特征与深度特征，建立二者之间的全局依赖关系，从而获取更准确的多级融合特征。为进一步捕获多尺度全局信息，本文引用了自适应金字塔上下文模块APC (Adaptive Pyramid Context Module)。在解码阶段，采用两个独立的解码器，语义解码器通过BGCF模块输出精确的分割结果，边界解码器则采用残差结构有效融合局部细节与全局信息，提升边界检测准确性。实验结果表明，该方法在Cityscapes数据集上取得了优越的分割精度与边界检测精度。

关键词

语义分割，边界检测，联合学习网络，RGB-D

Cross-Modal Fusion Network for RGB-D Joint Semantic Segmentation and Boundary Detection

Chaojie Li

School of Automobile and Transportation, Xihua University, Chengdu Sichuan

Received: December 9, 2025; accepted: January 5, 2026; published: January 15, 2026

文章引用: 李超杰. 基于 RGB-D 联合语义分割和边界检测的跨模态融合网络[J]. 人工智能与机器人研究, 2026, 15(1): 277-287. DOI: 10.12677/airr.2026.151027

Abstract

Semantic segmentation and boundary detection are two critical tasks for autonomous vehicles to achieve precise environmental awareness. However, most existing methods treat these tasks as independent or merely stack semantic and boundary features, neglecting the intrinsic relationship between them. This oversight results in a lack of explicit modeling of the interdependence between objects and boundaries, often leading to blurry boundaries and category confusion in regions with similar colors. To address this issue, we propose a cross-modal joint perception network, which enhances RGB images by incorporating depth information as geometric priors. Additionally, the method establishes a dynamic boundary guidance mechanism that utilizes both boundary information and geometric structure to jointly steer the semantic segmentation process. Specifically, the method employs a dual-branch architecture to separately capture RGB information and Depth information while introducing a Boundary Guided Cross Modality Fusion Module (BGCF). By dynamically fusing RGB features and depth features at various levels, we establish a global dependency relationship between the two modalities to obtain more accurate multi-level fusion features. To further enhance the capture of multi-scale global information, this paper references the Adaptive Pyramid Context Module (APC). In the decoding stage, two independently designed decoders are used. One for semantic output that generates precise segmentation results through the BGCF, another for boundary detection that employs lightweight residual units to effectively integrate local details with global context, improving boundary detection accuracy. Experimental results demonstrate that the method achieves superior segmentation and boundary detection accuracy on the Cityscapes dataset.

Keywords

Semantic Segmentation, Boundary Detection, Joint Learning Network, RGB-D

1. 引言

在自动驾驶系统中，精准的语义分割与边界检测对于环境感知至关重要[1]。语义分割旨在实现逐像素分类，标识出道路、车辆、行人等关键目标的分布情况。边界检测则聚焦于精确定位物体轮廓，为感知系统提供像素级的空间信息。对于单一任务的 RGB 语义分割和边界检测，研究已取得了显著进展。一方面基于 Transformer 的模型如 Segformer [2]通过自注意力机制有效建模了全局上下文，实现精准的语义分割。另一方面 DDS [3]在解码器的各分支采用多级监督机制，强制实现跨尺度的特征响应一致性，以生成更可靠的边缘表示。然而，上述方法主要基于 RGB 模态，在物体颜色相近或光照变化剧烈等复杂场景下，其感知能力仍面临限制。

为减缓颜色相近区域边界混淆问题并增强几何鲁棒性，本研究引入深度图作为补充模态。由于深度数据对纹理变化和光照条件不敏感，能够稳定地反映物体空间轮廓，从而为语义分割任务提供几何约束，同时为边界检测任务提供结构一致的有效监督[4]。然而，RGB 与深度模态间存在明显的信息不对称性，深度数据所表达的空间关系与 RGB 图像的语义特征存在差异，需要高效的跨模态融合机制实现二者互补。SGACNet [5]提出了一种空间信息引导的自适应上下文感知网络，通过利用深度图提供的精确几何信

息来动态引导 RGB 特征的多尺度上下文聚合，从而在保持高效性的同时，显著提升 RGB-D 语义分割的精度。

准确的语义分割结果有助于提升边界检测的可靠性，而精确的边界定位则影响语义分割结果的精细程度，二者协同工作，共同提升自动驾驶系统的环境感知能力。现有联合学习框架通常将双任务视为“并行任务”，仅在解码器阶段进行简单的特征堆叠或共享编码器，语义特征与边界特征的长距离依赖未被显式建模，当相邻物体颜色相近时，边界像素极易被错误分类。现有的方法已证明，语义分割和边界检测是相辅相成的，准确的语义标签有助于精确的边界描绘，而定义良好的边界通过提供几何约束来提高分割精度[1]。Gated-SCNN [6]在传统分割网络中引入了具有边界流和语义流的双分支架构，通过门控机制，实现了几何特征和语义特征的动态融合，提高了分割精度和对象边界的结构连贯性。然而，这种方法缺乏对语义边界依赖关系的明确建模，使得在全局上下文理解与细粒度之间难以取得平衡。

语义分割与边界检测并非简单的“并行任务”，而是存在可建模的条件依赖，边界概率应在语义类别发生跃迁的区域显著升高，而语义预测在边界附近应抑制跨区域的信息泄漏并保持区域内一致性。若仅采用共享主干或特征拼接，网络往往只能进行弱耦合，难以显式学习“语义变化与边界响应”的对应关系，进而导致边界处的粘连、轮廓过平滑或伪边缘增强等现象。

基于此，本文提出“显式建模语义—边界依赖关系”的核心思路，通过可学习的跨分支交互机制，利用边界线索对语义特征施加空间约束，同时以语义先验对边界特征进行语义筛选，从而在特征层形成双向协同优化。

因此，本文构建了一种跨模态多任务联合感知网络，以实现精确的 RGB-D 语义分割与边界检测。该网络包含两个编码器，分别用来提取 RGB 特征和 Depth 特征，此外还设计了边界引导的跨模态融合模块 BGCF (Boundary-Guided Cross-modality Fusion Module)，通过动态权重学习实现语义与边界特征的自适应融合，并结合多尺度上下文模块 APC (Adaptive Pyramid Context Module) [5]以聚合不同尺度的上下文特征。此外，进一步构建包含交叉任务一致性约束的联合损失函数。该损失以互监督方式联动两分支，利用语义预测的置信度自适应调整边界学习，同时利用边界预测的清晰度锐化语义分割的边缘响应，从而驱动两任务功能互补与性能协同提升。

主要贡献总结如下：

1) 面向自动驾驶场景，提出联合语义分割与边界检测框架，在 RGB-D 输入下系统挖掘两任务的相关性与互补线索，实现性能协同提升。

2) 提出了边界引导的跨模态融合模块 BGCF 动态学习语义–边界特征的通道关联，自适应分配权重，有效提升检测与分割精度。引入自适应金字塔上下文模块 APC，获取多层级特征间的长距离依赖关系，增强模型对多尺度目标的感知能力。

2. 相关工作

2.1. 语义分割

语义分割作为一项密集像素级分类任务，旨在为图像中的每个像素分配相应的语义类别标签。自全卷积网络(FCN) [7]奠定 RGB 图像语义分割基础以来，该领域持续涌现出多种改进方法。U-Net [8]通过编码器–解码器结构和跳跃连接，有效融合深层语义信息与浅层细节特征。DeepLab [9]系列引入空洞卷积扩大感受野，并通过空间金字塔池化(ASPP)模块捕获多尺度上下文信息。随着 Transformer 架构的广泛应用，MaskFormer [10]通过将语义分割重新定义为掩码分类问题，即预测 N 个二元掩码及其类别，统一了语义分割与实例分割任务范式，开创了分割研究新方向。AFFormer [11]通过自适应频率模块，将图像从空间域转换到频率域，以动态滤除冗余高频细节并保留关键低频结构，从而在保持高精度的同时显著提

升模型效率。

尽管基于 RGB 图像的语义分割方法已取得显著进展，其在复杂场景下的性能仍受限于纹理相似、光照变化及遮挡等因素。近年来，随着深度传感技术的普及，RGB-D 图像逐渐成为研究热点。相较于 RGB 图像，深度图像提供了额外的几何结构信息，能够在一定程度上缓解上述问题，从而提升分割精度。然而，由于深度模态与 RGB 模态在特征表示上存在差异，将深度信息有效整合到 RGB 分割架构中仍具有一定挑战性。早期的 RGB-D 语义分割方法利用 FCN [7]将 RGB-D 信息视为单一输入，并使用同一骨干网络进行处理。然而，后续研究逐渐认识到需分别处理 RGB 和深度信息。AsymFormer [12]提出了一种用于实时 RGB-D 语义分割的不对称跨模态表示学习方法，针对不同模态的计算特性进行不对称设计，对深度信息使用轻量级解码器，对 RGB 信息使用复杂解码器，并通过跨模态注意力模块进行有效融合。ACNet [13]通过一个并行的双流编解码结构，并设计了一个注意力互补融合模块，以自适应地挖掘并融合来自 RGB 图像的外观特征和来自深度图的几何特征，从而提升分割精度。

2.2. 边界检测

边界检测的核心目标是准确识别图像中不同对象或语义区域之间的边界轮廓，为后续的视觉理解任务提供结构化的几何和拓扑先验。传统的检测方法通常依赖于手动设计的低级特征算子(如 Canny、Sobel 或结构化边缘 SE)来捕捉局部亮度、颜色或纹理的突变，并将其与梯度幅度阈值或图像分割算法相结合，以实现像素级边缘分类。随着深度学习的发展，多数方法倾向于将边界检测嵌入到神经网络中，利用其与语义分割任务的互补关系来实现相互促进。

动态特征融合(DFF) [14]机制通过自适应地集成包含高级语义信息和低级细节特征的多级表示，显著提高了边界检测的准确性。在此基础上，RPCNet [15]进一步提出了一种迭代金字塔上下文模块，可以在多个尺度上迭代优化语义和边界特征，实现跨层双向信息增强，从而提高语义分割和边界检测的性能。与此同时，CASENet [16]基于 ResNet-101 的提出多标签语义边界识别框架，将浅细节特征与高级语义信息相结合，实现了类别感知边缘检测。沿着这一技术路线，DcoupleNet [17]创新性地提出解耦监督框架，通过并行双分支网络分别学习物体内部主体特征和外部边界信息，在保持边界锐利度的同时提升内部区域一致性。为了兼顾效率与性能，LiteEdge [18]通过设计高效的深度可分离卷积和特征金字塔结构，在保持高精度的同时大幅降低计算复杂度。与语义和边界特征的简单连接或对分割结果进行后处理不同，本文使用边界信息作为关键线索来增强上下文建模能力，并通过注意力机制进一步加强语义表示。

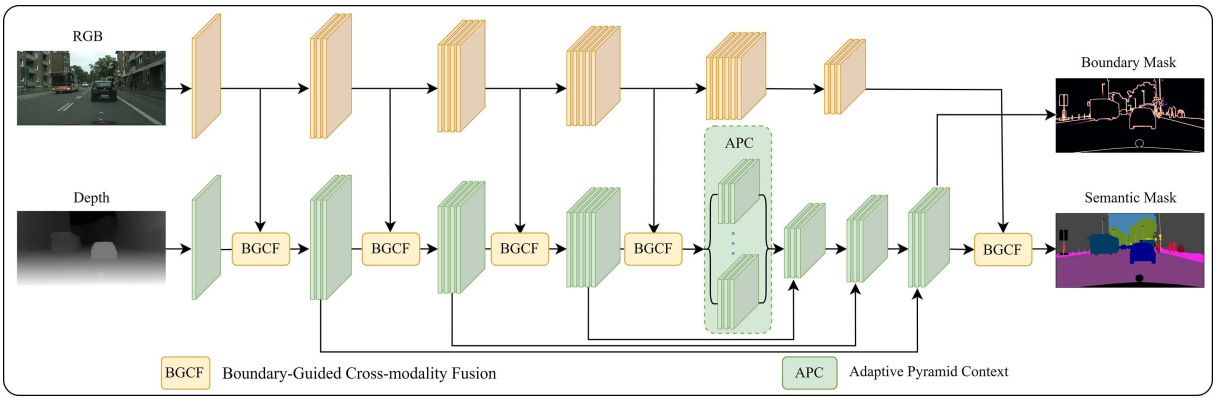

Figure 1. Overall block diagram
图 1. 整体框图

3. 算法原理和模型架构

在本节中，介绍了用于联合语义分割与边界检测的跨模态网络。该网络由三个关键部分组成，用于提取 RGB 特征和深度特征的双流编码器；能够实现选择性特征融合的边界引导跨模态融合模块 BGCF 以及充分挖掘全局信息的自适应金字塔上下文模块 APC。

3.1. 整体架构概述

为高效捕获语义与边界特征并实现信息互补，本研究设计了双流编码器架构，如图 1 所示，具体流程如下，首先，采用由两个 MobileNetV2 [19]块组成的简单共享主干模块将原始图像 $I_{\mathrm{RGB}} \in \mathbb{R}^{3\times H\times W}$ 和深度图 $I_{\mathrm{Depth}} \in \mathbb{R}^{1\times H\times W}$ 嵌入高维特征空间(其中 H，W 分别为图像的高度与宽度)并输出低级特征表示。在本文中选择了 AFFormer-T [11]作为网络的骨干网络。上分支处理 RGB 图像旨在生成丰富的语义特征。另一分支中用来处理深度信息以捕获几何结构。在两个分支的多个尺度上嵌入边界引导的跨模态融合机制 BGCF，将深度几何特征与 RGB 语义特征选择性融合，突出目标边界特征。这种融合机制的核心在于对语义与边界依赖关系的显式建模，其表现为一种双向约束机制，一方面，边界特征为语义特征提供空间引导，使其在物体轮廓附近增强梯度响应并抑制特征向区域外的扩散；另一方面，语义特征为边界检测提供类别先验，有助于过滤由纹理、阴影等产生的伪边缘，从而提升边界结果的结构一致性与语义可解释性。

需要强调的是，BGCF 模块并非简单的特征拼接。通过通道级别的联合建模，来参数化“语义–边界”的依赖强度。该模块利用多头缩放点积注意力计算跨模态通道间的亲和矩阵，使网络能够自适应地学习，在边界线索的引导下，增强或抑制相应的语义通道，生成动态的分割分支权重与边界分支权重，从而实现对双任务信息流的自适应调制与融合。

为了充分利用融合的多级特征，引入自适应金字塔(APC)模块，并以跳跃连接方式整合多尺度特征，实现更精细的跨层信息交互与表征增强。在解码器阶段，语义解码器利用 BGCF 输出的聚合特征，并通过像素级预测器生成精细的语义分割结果。边界解码器利用残差结构融合局部和全局上下文信息来重建高精度的边界掩模。整个网络使用交叉熵(CE)和二元交叉熵(BCE)的联合损失函数进行端到端的训练，实现了语义一致性和边界准确性的协同优化。

3.2. 边界引导的跨模态融合模块(Boundary-Guided Cross-Modality Fusion Module)

为了有效地融合高维语义和边界特征，本文提出了边界引导的跨模态融合模块(BGCF)。如图 2 所示其核心目标是实现 RGB 特征与深度特征的自适应融合以及在多尺度上稳健地保留边缘信息。该模块接收两个输入，来自编码器阶段的 RGB 语义特征 $F_s \in \mathbb{R}^{C\times H\times W}$ 以及 Depth 几何特征 $F_d \in \mathbb{R}^{C\times H\times W}$，BGCF 模块对这些输入进行融合处理，并输出一个融合特征图表示为 F_d。

首先对两个特征图分别进行全局平均池化(GAP)、1 × 1 卷积、RELU 和 Sigmoid 来生成通道注意向量：

$$F_{\mathrm{att\_s}} = \sigma\left(\mathrm{RELU}\left(\mathrm{Conv}\left(\mathrm{GAP}\left(F_s\right)\right)\right)\right) \tag{1}$$

$$F_{\mathrm{att\_d}} = \sigma\left(\mathrm{RELU}\left(\mathrm{Conv}\left(\mathrm{GAP}\left(F_d\right)\right)\right)\right) \tag{2}$$

其中，GAP 表示全局平均池化操作，Conv 表示卷积层，RELU 表示修正线性单元激活函数，σ 表示 sigmoid 函数。

边界特征的关注向量：

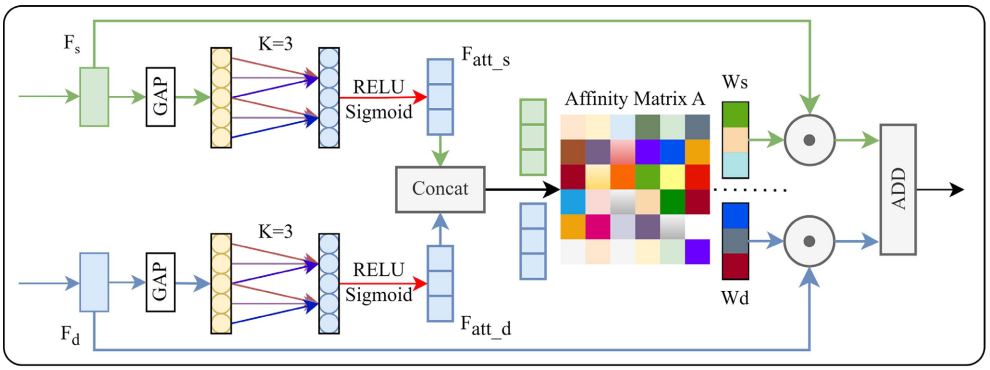

Figure 2. Network framework of BGCF module
图 2. BGCF 模块的网络框架

$$F^{\text{att}} = \left(F_{\text{att\_s}} \| F_{\text{att\_d}} \right) \tag{3}$$

其中||表示信道级联操作。为了度量语义和边界特征的每个通道之间的亲和力，分割了融合注意力向量 $F^{\text{att}} \in \mathbb{R}^{2C}$，并通过缩放点积注意力计算跨流特征的亲和矩阵：

$$A = \text{Softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d}} \right) \tag{4}$$

其中 d 为缩放因子，第 i 个主动融合权重 w 的计算公式为：

$$w_i = A_i V_i \tag{5}$$

所有结果拼接后得到最终融合权重向量：

$$w = \text{Concat}\left(w_1, w_2, \cdots, w_h \right) \tag{6}$$

其中 h 为注意力头数。然后，将 w 划分为语义分支权重 w_s 与边界分支权重 w_{bi}，并分别作用于原始特征：

$$F_s' = F_s \odot w_s \ , \ F_d' = F_d \odot w_d \tag{7}$$

其中⊙表示逐通道乘法。最后，通过残差连接并拼接融合特征：

$$F_s = \left(F_s' + F_s \right) + \left(F_d' + F_d \right) \tag{8}$$

3.3. 特征金字塔上下文模块(Adaptive Pyramid Context Module)

在编码和解码过程中，类似于上采样和池化等操作容易导致重要语义信息的丢失。因此引入上下文模块有助于强化图像特征表达，由于输入特征在不同尺度上具有不同的分辨率，特征金字塔能够依据对应尺度提取目标相关信息，从而提升整体特征图的质量。为此，本研究引入了一个自适应金字塔上下文模块，如图 3 所示。该模块融合多层级特征，其分支数量可随条件变化而调整。具备较大感受野的上下文信息能够促进对相近局部区域内不同类别间共享特征的理解，进而优化分割效果。为减少计算复杂度，该模块采用了结构简洁的最近邻上采样方法。

4. 实验

本节在城市道路交通场景数据集 Cityscapes [20]上进行了实验，以综合评估本方法在复杂交通环境中的感知能力。第 4.1 节介绍了实施细节和评估指标。在第 4.2 节中，在 Cityscapes 数据集上将本研究的方法与基线方法进行了比较。在第 4.3 节中，通过对 BGCF 与 APC 模块进行消融研究，旨在验证其设计对

网络性能的必要性与贡献。结果证明提出的方法可以：1) 联合学习语义分割和边界检测任务；2) 提升语义分割和边界检测的精度。

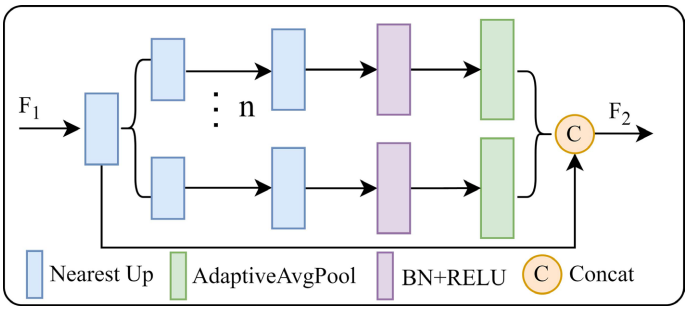

Figure 3. Network framework of APC module
图 3. APC 模块的网络框架

4.1. 实验设置

1) 数据集：Cityscapes 数据集包含 5000 张图像，分辨率为 1024 × 2048 像素，19 个类别有细粒度注释。使用 2975 张图片进行训练，500 张用于验证，1525 张用于测试。

2) 评估指标：使用了两个定量指标来评估方法的性能。1) 采用交并比(IoU)来评估语义分割结果。2) 为了证明本研究可以提取高质量语义边界，在 Cityscapes-val 数据集上引入边界 IoU (BIoU) [21]，以进一步评估语义边界性能。

3) 实施细节：我们基于 MMsegmentation 框架构建网络模型。在 NVIDIA RTX 3090 GPU 上进行所有实验。选择 AFFormer-T [11]作为 RGB 分支，并在 ImageNet-1k 上进行预训练，而深度分支则从头开始学习。对 Cityscapes 数据集使用 AdamW 优化器来更新模型参数。数据增强方法包括随机调整大小、随机缩放、随机水平翻转和颜色抖动。

4.2. 实验结果

表 1 显示了 Cityscapes-val 数据集上语义分割结果与基线方法的比较。与基线相比，我们的方法具有更高的 mIoU 性能，达到了 79.50%，验证了我们的双流设计在准确性方面得到明显提升。此外，图 4 所示的定性结果表明我们的方法显著减少了道路和人行道与路边绿地之间的类别混淆，减少了复杂背景下车辆的误分类，有效地抑制了行人等小目标的碎片化和丢失。

为了证明我们的方法实现了更精确的边界定位，使用表 2 中的 BIoU 度量来评估语义边界准确性。这表明我们的方法在很大程度上优于基线方法，BIoU 达到了 46.39%，验证了联合学习语义分割和边界检测可以提高边界区域的分割性能。语义边界预测的可视化结果如图 5 所示。与 Mobile-Seed 相比，我们的方法在行人剪影、车辆前后边缘和车道线等关键结构处产生更完整的分割边界，显著减少了类别混淆和过度的边界平滑。

尽管所提方法在 Cityscapes 上取得了更优的 mIoU 与 BIoU，但在部分复杂场景中仍会出现边界模糊和分割错误，如图 6 所示。对于细长结构与小目标，由于分辨率限制与多次下采样带来的细节损失，边界响应可能出现断裂或不闭合，使得语义分支缺乏足够强的边界约束，产生细节缺失或目标被背景吞并等问题。该现象表明，仅通过通道级依赖建模仍不足以完全恢复像素级几何细节。语义预测更倾向于依赖全局上下文而牺牲边界精度，边界预测则可能被噪声梯度主导而出现过平滑。此外，当深度图存在空洞、量化噪声或对齐误差时，网络可能将不可靠的几何边缘误认为真实边界，导致边界分支出现伪边缘

增强，并进一步通过交互路径反向影响语义分割，造成分割和检测的不准确。

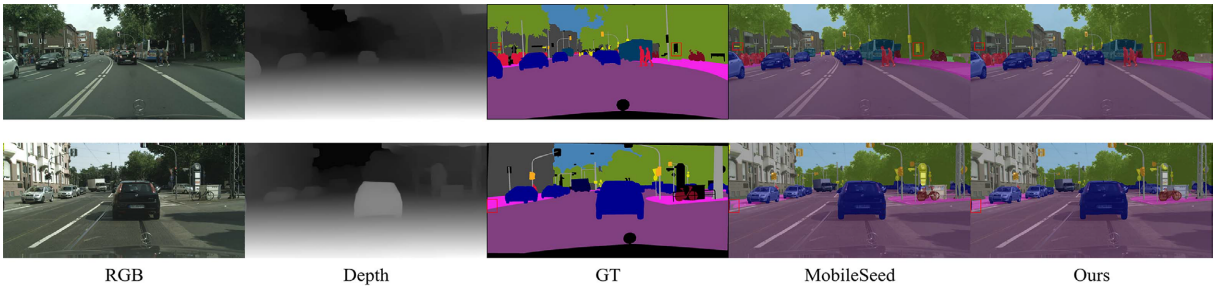

Figure 4. Qualitative semantic segmentation results on the Cityscapes dataset
图 4. Cityscapes 数据集上定性语义分割结果

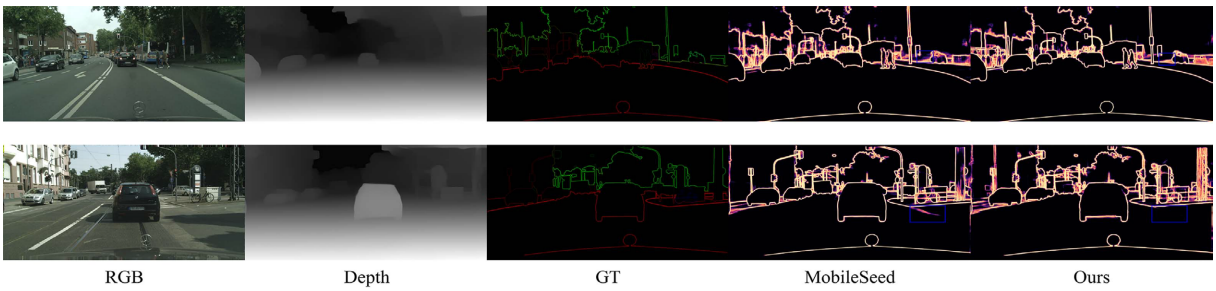

Figure 5. Qualitative boundary detection results on the Cityscapes dataset
图 5. Cityscapes 数据集上定性边界检测结果

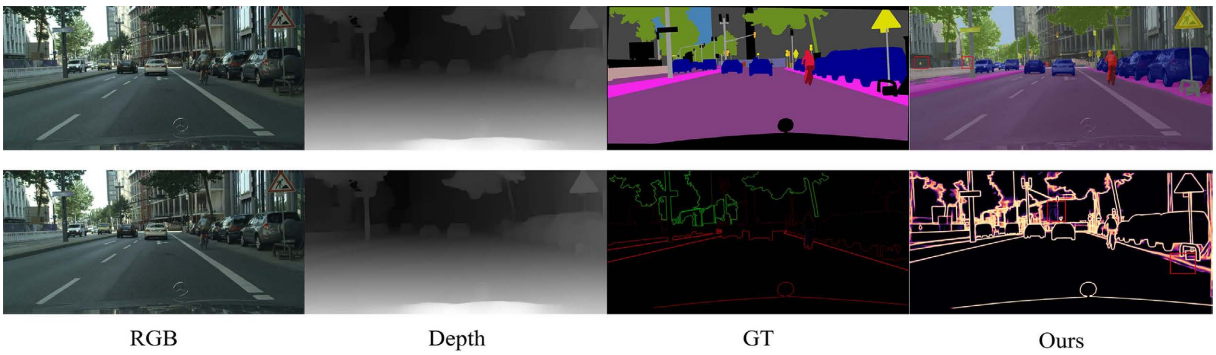

Figure 6. Recognition results for small targets and complex scenes on the Cityscapes dataset
图 6. Cityscapes 数据集上针对小目标和复杂场景的识别结果

Table 1. The performance of the proposed semantic segmentation method was systematically compared with that of the baseline model on the full-resolution (1024 × 2048) Cityscapes dataset
表 1. 在全分辨率(1024 × 2048) Cityscapes 数据集上系统对比了所提语义分割方法与基线模型的性能

方法	骨干网络	输入	mIoU (%)
MobileSeed	AFF-T	RGB	78.40
MobileSeed-d	AFF-T	RGB-D	77.45
AFFormer-T	AFF-T	RGB	78.70
SGACNet	R34-NBt1D	RGB-D	78.70
Ours	AFF-T	RGB-D	**79.50**

Table 2. The performance of the proposed boundary detection method was systematically compared with the baseline model on the full-resolution (1024 × 2048) Cityscapes dataset
表 2. 在全分辨率(1024 × 2048) Cityscapes 数据集上系统对比了所提边界检测方法与基线模型的性能

方法	骨干网络	输入	BIoU (%)
MobileSeed	AFF-T	RGB	43.30
MobileSeed-d	AFF-T	RGB-D	42.15
AFFormer-T	AFF-T	RGB	41.30
Ours	AFF-T	RGB-D	**46.39**

4.3. 消融实验

4.3.1. BGCF 模块

为评估 BGCF 中通道注意力机制对网络的影响，我们使用 ESqueezeAndExcitation 与 ChannelAtt 两种机制进行对比，并保持其余超参数完全一致。实验结果表明，ESqueezeAndExcitation 具有更优的性能。这主要归因于其采用一维卷积捕捉局部通道交互，避免了 ChannelAtt 中全局全连接操作可能引入的噪声，从而更好地保持了通道信息的完整性。同时，ESqueezeAndExcitation 参数更少，降低了过拟合风险，进一步增强了模型的泛化能力。实验结果如表 3 所示，其中实验 2 与实验 3 分别对应两种注意力机制，验证了 ESqueezeAndExcitation 在提升模型性能方面的有效性。

Table 3. In ablation studies on the Cityscapes dataset, all metrics are expressed as (%)
表 3. 在 Cityscapes 数据集上的消融研究，所有指标均以(%)表示

方法	APC	ESqueezeAndExcitation	ChannelAtt	mIoU (%)	BIoU (%)
#1	×	√	×	78.58	45.98
#2	√	×	√	78.77	46.27
#3	√	√	×	**79.50**	**46.39**

4.3.2. APC 模块

为明确自适应金字塔上下文(APC)模块对于模型的实际效能，在 Cityscapes 数据集上进行了消融实验，将 APC 模块替换为单层卷积结构以连接模型中的上下支路，同时保持所有其他超参数一致。该设置旨在剥离 APC 多尺度上下文融合能力的影响，实验结果如表 3 所示，实验 1 和实验 3 分别使用 APC 模块和单层卷积结构，验证了 APC 模块的有效性。

5. 结论

本文提出了一种跨模态联合感知网络，该框架由两个独立编码器，边界引导的跨模态融合模块(BGCF)以及自适应金字塔上下文模块(APC)组成，编码器分别提取 RGB 和 Depth 特征，边界引导的跨模态融合模块为两种特征分配动态融合权值，以获取丰富的融合特征，自适应金字塔模块，以跳跃连接方式整合多尺度特征，实现更精细的跨层信息交互与表征增强。在 Cityscapes 数据集上实现了该方法，并与基线方法进行了比较，实验结果表明，该方法的性能优于基线方法 mIoU、BIoU 分别达到了 79.50%和 46.39%。基于上述研究，未来的工作将进一步探索对噪声和缺失数据更为鲁棒的深度信息表示与融合方法，以提升模型在真实环境下的稳定性与泛化能力。针对细长物体、小目标等困难样本的感知难题，设计专用的数据增强策略或损失函数，以加强对困难区域及边界像素的特征学习，提升模型对复杂场景的辨别能力。

同时进一步研究网络分支、知识蒸馏等技术，保持准确性的前提下，显著减少参数数量和计算复杂性，以适应 NVIDIA Jetson 等边缘计算平台的资源限制。

参考文献

[1] Liao, Y., Kang, S., Li, J., Liu, Y., Liu, Y., Dong, Z., *et al*. (2024) Mobile-Seed: Joint Semantic Segmentation and Boundary Detection for Mobile Robots. *IEEE Robotics and Automation Letters*, **9**, 3902-3909. https://doi.org/10.1109/lra.2024.3373235

[2] Xie, E., Wang, W., Yu, Z., *et al*. (2021) SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Annual Conference on Neural Information Processing Systems* 2021, 6-14 December 2021, 12077-12090.

[3] Liu, Y., Cheng, M., Fan, D., Zhang, L., Bian, J. and Tao, D. (2021) Semantic Edge Detection with Diverse Deep Supervision. *International Journal of Computer Vision*, **130**, 179-198. https://doi.org/10.1007/s11263-021-01539-8

[4] Xiao, X., Zhao, Y., Zhang, F., Luo, B., Yu, L., Chen, B., *et al*. (2023) BASeg: Boundary Aware Semantic Segmentation for Autonomous Driving. *Neural Networks*, **157**, 460-470. https://doi.org/10.1016/j.neunet.2022.10.034

[5] Zhang, Y., Xiong, C., Liu, J., Ye, X. and Sun, G. (2023) Spatial Information-Guided Adaptive Context-Aware Network for Efficient RGB-D Semantic Segmentation. *IEEE Sensors Journal*, **23**, 23512-23521. https://doi.org/10.1109/jsen.2023.3304637

[6] Takikawa, T., Acuna, D., Jampani, V. and Fidler, S. (2019) Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 5229-5238. https://doi.org/10.1109/iccv.2019.00533

[7] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[8] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., *et al*., Eds., *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer International Publishing, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[9] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2018) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 834-848. https://doi.org/10.1109/tpami.2017.2699184

[10] Cheng, B., Schwing, A. and Kirillov, A. (2021) Per-Pixel Classification Is Not All You Need for Semantic Segmentation. Advances in Neural Information Processing Systems, 34, 17864-17875.

[11] Dong, B., Wang, P. and Wang, F. (2023) Head-Free Lightweight Semantic Segmentation with Linear Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 516-524. https://doi.org/10.1609/aaai.v37i1.25126

[12] Du, S., Wang, W., Guo, R., Wang, R. and Tang, S. (2024) AsymFormer: Asymmetrical Cross-Modal Representation Learning for Mobile Platform Real-Time RGB-D Semantic Segmentation. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 17-18 June 2024, 7608-7615. https://doi.org/10.1109/cvprw63382.2024.00756

[13] Hu, X., Yang, K., Fei, L. and Wang, K. (2019) ACNET: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation. 2019 *IEEE International Conference on Image Processing* (*ICIP*), 22-25 September 2019, 1440-1444. https://doi.org/10.1109/icip.2019.8803025

[14] Hu, Y., Chen, Y., Li, X. and Feng, J. (2019) Dynamic Feature Fusion for Semantic Edge Detection. *Proceedings of the* 28*th International Joint Conference on Artificial Intelligence*, Macao, 10-16 August 2019, 782-788. https://doi.org/10.24963/ijcai.2019/110

[15] Zhen, M., Wang, J., Zhou, L., Li, S., Shen, T., Shang, J., *et al*. (2020) Joint Semantic Segmentation and Boundary Detection Using Iterative Pyramid Contexts. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 13666-13675. https://doi.org/10.1109/cvpr42600.2020.01368

[16] Yu, Z., Feng, C., Liu, M. and Ramalingam, S. (2017) CASENet: Deep Category-Aware Semantic Edge Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5964-5973. https://doi.org/10.1109/cvpr.2017.191

[17] Li, X., Li, X., Zhang, L., Cheng, G., Shi, J., Lin, Z., *et al*. (2020) Improving Semantic Segmentation via Decoupled Body and Edge Supervision. 16*th European Conference ECCV* 2020, Glasgow, 23-28 August 2020, 435-452. https://doi.org/10.1007/978-3-030-58520-4_26

[18] Wang, H., Mohamed, H., Wang, Z., Rueckauer, B. and Liu, S. (2021) LiteEdge: Lightweight Semantic Edge Detection Network. 2021 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October

2021, 2657-2666. https://doi.org/10.1109/iccvw54120.2021.00300

[19] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L. (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4510-4520. https://doi.org/10.1109/cvpr.2018.00474

[20] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., *et al*. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 3213-3223. https://doi.org/10.1109/cvpr.2016.350

[21] Cheng, B., Girshick, R., Dollar, P., Berg, A.C. and Kirillov, A. (2021) Boundary IoU: Improving Object-Centric Image Segmentation Evaluation. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 15334-15342. https://doi.org/10.1109/cvpr46437.2021.01508