计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):254−262 [doi: 10.15888/j.cnki.csa.009475]

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

融合位置注意力机制与轻量化 STDC 网络的非结构化场景语义分割①

陈　晔[1], 杨长春[2], 杨　森[2], 王宇鹏[2], 王　彭[2]

[1](常州纺织服装职业技术学院, 常州 213164)
[2](常州大学 计算机与人工智能学院, 常州 213164)
通信作者: 陈　晔, E-mail: cy@cztgi.edu.cn

摘　要: 近年来, 非结构化道路分割已成为计算机视觉领域的重要研究方向之一. 现有的大多数方法适合结构化道路的分割并无法满足非结构化道路分割的准确性与实时性需求. 为了解决上述问题, 本文对 STDC 网络进行改进, 引入残差连接来更好地融合多尺度语义信息, 还提出一种嵌入位置注意力模块的空洞空间卷积池化金字塔 (PA-ASPP) 来增强网络对道路等特定区域的位置感知能力. 本文在 RUGD 与 RELLIS-3D 两个数据集上进行实验, 所提出方法的 *MIoU* 在两个数据集的测试集上分别达到了 50.78% 和 49.96%.

关键词: 非结构化环境; 语义分割; PA-ASPP; STDC

引用格式: 陈晔,杨长春,杨森,王宇鹏,王彭.融合位置注意力机制与轻量化 STDC 网络的非结构化场景语义分割.计算机系统应用,2024,33(4):254–262. http://www.c-s-a.org.cn/1003-3254/9475.html

Unstructured Scene Semantic Segmentation Combining Location Attention Mechanism and Lightweight STDC Network

CHEN Ye[1], YANG Chang-Chun[2], YANG Sen[2], WANG Yu-Peng[2], WANG Peng[2]

[1](Changzhou Vocational Institute of Textile and Garment, Changzhou 213164, China)
[2](School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213164, China)

Abstract: In recent years, unstructured road segmentation has become one of the important research directions in the field of computer vision. Most existing methods are suitable for structured road segmentation and cannot meet the accuracy and real-time requirements of unstructured road segmentation. To address the above issues, this study improves the short-term dense concatenate (STDC) network by introducing residual connections to better integrate multi-scale semantic information. Additionally, it proposes a position attention-aware spatial pyramid pooling (PA-ASPP) module to enhance the network's position awareness ability for specific regions such as roads. Experiments are conducted on two datasets, RUGD and RELLIS-3D, and the proposed method achieves a mean intersection over union (*MIoU*) of 50.78% and 49.96% on the test sets of the two datasets, respectively.

Key words: unstructured environment; semantic segmentation; position attention-aware spatial pyramid pooling (PA-ASPP); short-term dense concatenate (STDC)

无人车在非结构化的野外环境中有广泛的应用, 如勘探、救援和侦查等领域. 然而, 在野外环境中, 类别和地形结构之间缺乏明确的边界, 因此非结构化场景中的语义分割具有挑战性. 此外, 将算法部署在真实

① 基金项目: 江苏省现代教育技术研究课题 (2021-R-88294); 江苏研究生科研创新项目 (KYCX23_3169)
收稿时间: 2023-10-10; 修改时间: 2023-11-09, 2023-12-04; 采用时间: 2023-12-15; csa 在线出版时间: 2024-03-01
CNKI 网络首发时间: 2024-03-07

环境中时, 会出现低实时性和低准确性等问题, 这会对无人车的运动决策产生严重干扰, 使其无法完成既定工作. 视觉感知是无人车感知周围环境的重要手段, 其中语义分割是计算机视觉中的一个关键应用, 能够全面理解场景. 近年来, 语义分割在视觉场景理解方面取得了重大进展[1–4]. 然而, 这些网络在非结构化的场景中并不能很好的运作, 原因如下.

1) 非结构化场景中的类别和地形结构的边界重叠现象严重, 需要高精度的边界分割. 近些年, 已有一些改进方法来应对这个问题. 例如, 张凯航等人改进 SegNet[2]实现了对非结构化道路可行驶区域的语义分割[5]. 龚志力等人提出了基于改进 DeepLabV3+[4]的非结构化道路分割方法[6]. 艾青林等人结合注意力机制和 ICNet[7]提出了一种非结构化场景语义分割方法[8]. 然而, 这些方法仍然存在边界重叠分割效果不佳和实时性不足的问题.

2) 无人车在非结构化场景下需要高效的分割算法, 以保证实时性和准确性, 并避免错误决策. 然而, 目前的轻量化网络如 DDRNet[9]、STDC[10]在实时性方面表现较好, 但精度较低且边界重叠处理效果不佳.

与传统的结构化场景语义分割相比, 现有的方法会面临一些特殊的挑战和问题, 如环境复杂性及多样性和变化性. 在非结构化场景中, 可能存在大量的遮挡物、光照变化、背景干扰等因素, 这会增加语义分割的难度. 相比之下, 传统的结构化场景更容易处理, 因为它们通常具有明确定义的区域和边界. 同时, 非结构化场景中的对象形状、大小、姿态等方面的变化更加多样化, 这使得模型需要更好的泛化能力来适应各种情况. 而现有深度学习模型不适应的地方如下.

1) 对计算资源要求高: 针对非结构化场景的语义分割任务, 由于场景的复杂性, 需要更大的模型和更多的计算资源来处理, 这对于一些资源受限的情况可能造成困难.

2) 泛化能力不足: 目前的深度学习模型在面对极端多样性和复杂性的非结构化场景时, 可能存在泛化能力不足的问题, 导致在实际应用中的稳定性和鲁棒性有所欠缺.

针对以上问题, 本文对现有方法进行了改进, 提出了一种基于位置注意力机制的轻量级非结构化场景语义分割方法, 称为 PA-SNet. 该方法改进了 STDC 网络, 引入残差结构以传递不同尺度的信息, 从而更好地获取场景语义上下文信息. 同时, 在 ASPP 结构中嵌入位置注意力机制来增强网络对道路等特定区域的位置感知能力. 具体工作如下.

1) 针对非结构化场景语义分割难以兼顾分割精度和实时性问题, 本文提出一种改进的 STDC 网络. 该网络在保证较少参数量的前提下能够尽可能多地融合多尺度特征, 减少边界信息的丢失, 实现了准确率和实时性之间的较好平衡.

2) 针对场景语义上下文信息的获取问题, 本文提出一种改进的 PA-ASPP 模块. 该模块通过嵌入位置注意力机制来捕获特征图中任意两个位置之间的依赖关系, 增强模型对道路等特定区域的感知能力. 在分割道路的场景中, 可以使用位置注意力机制来突出显示道路的位置, 从而更好地区分道路和其他物体.

3) 针对类别分布不均衡问题, 本文提出加权交叉熵损失函数. 该损失函数通过融合 OHEM (online hard example mining) 算法对不同类别赋予不同权重, 以提升网络对小样本类别的关注度, 缓解类别不平衡问题.

1 相关工作

1.1 语义分割

语义分割是计算机视觉中的一项重要任务, 为智能驾驶的环境感知提供了重要基础. 在过去的几年里, 语义分割经历了飞速发展, 并在各个领域得到了广泛应用, 包括医学图像分割、机器人技术和智能驾驶汽车等. 首先, FCN[1]是第 1 个采用全卷积网络进行语义分割的方法. 随后, 基于 FCN 的方法在图像语义分割方面取得了显著进展. Chen 等人[11]和 Yu 等人[12]对 FCN 进行了改进, 去除了两个下采样层以获得密集预测, 并利用扩张卷积来扩大感受野. 此外, 还有其他方法如 U-Net[3]、DeepLabV3+[4]、MSCI[13]、SPGNet[14]、RefineNet[15]和 DFN[16], 它们采用了编码器-解码器结构来预测分割的掩码, 同时融合了低层信息和高层信息.

除了准确性, 实时性也是语义分割任务中不可忽视的因素. 为了满足实时语义分割的需求, 出现了一些轻量级网络, 如 BiSeNet[17]、DDRNet[9]、ConvNeXt[18]、和 FasterNet[19]等. 这些网络在结构化环境 (如 Cityscapes[20]、ADE20K[21]等) 中取得了良好的效果. 然而, 它们在非结构化环境中的有效性尚未得到验证.

1.2 非结构化环境中的语言分割

在结构化的城市道路场景中, 语义分割已经取得

了显著进步. 然而, 在实际的道路驾驶和机器人野外工作等非结构化场景中, 面临着许多挑战. 相比于结构化道路, 非结构化道路存在着分割精度低、实时性差和边界区分不明显等问题. 因此, 适用于结构化环境的分割方法并不一定适用于非结构化环境. 为了解决这个问题, Baheti 等人进行了一系列的尝试, 在文献[22]中对 DeepLabV3+进行了改进, 采用残差网络作为骨干网络, 以克服下采样过程中可能丢失小细节的问题. 此外, 文献[23]提出了一个记忆模块, 以更好地处理非结构化环境中的光照变化问题, 并有效捕捉不清晰的物体. 此外, 张凯航等人[5]提出了基于 SegNet 的非结构化道路可行驶区域语义分割方法, 而龚志力等人[6]提出了基于改进 DeepLabV3+的非结构化道路分割方法. 针对网络效率问题, Baheti 等人还对 DeepLabV3+框架进行了修改, 采用 Xception 网络[24]作为特征提取的骨干网络, 并提出了一个轻量化的网络[25], 艾青林等人[8]进行了改进, 提出了 AF-ICNet, 引入了注意力机制, 从而提升了对小目标类别的分割精度和分割速度.

1.3 注意力机制

注意力机制是一种常见的机器学习技术, 用于解决输入数据中的部分信息对于输出结果的贡献更大的情况. 它在自然语言处理、计算机视觉、语音识别等领域都有广泛的应用. 其中, SE 网络[26]通过对注意力机制中的通道关系进行建模, 增强了网络的表示能力. 此外, Chen 等人[27] 提出了利用几个注意力掩码来融合来自不同分支的特征图或预测图的方法. 另外, Wang 等人[28]提出了非本地模块, 该模块通过计算特征图中每个空间点之间的相关矩阵来生成巨大的注意力图, 然后通过注意力引导来聚合密集的上下文信息. 除了上述提到的注意力机制, 还存在许多其他形式的注意力机制, 如多头注意力机制和位置注意力机制等.

本文采用的是位置注意力机制, 其在处理小目标的位置识别任务上表现出色的原因是它能够聚焦于图像中的特定位置, 并对该位置的细节信息进行加权处理. 位置注意力机制具有一些优势, 使其在本文所研究的环境下有其独特的适用性.

(1) 非结构化场景通常具有丰富的上下文信息, 例如背景元素、周围环境等. 位置注意力机制可以在关注小目标的同时, 考虑到周围上下文的影响, 从而提高对小目标的理解和识别能力.

(2) 非结构化场景中可能存在尺度变化较大的物体, 位置注意力机制可以在不同尺度上自适应地聚焦于感兴趣的目标区域. 通过对不同尺度的位置进行注意力加权, 模型能够具备对多尺度目标的感知能力.

(3) 非结构化场景中的目标通常具有复杂的形状和纹理, 位置注意力机制可以帮助模型关注目标的关键部分, 在复杂背景中更准确地定位目标.

(4) 位置注意力机制可以增强模型对目标位置的鲁棒性. 在非结构化场景中, 目标可能存在遮挡、不完整性或形变等问题. 通过位置注意力机制, 模型可以更加集中地处理目标的重要区域, 从而提高对目标的识别准确性和鲁棒性.

2 方法

2.1 网络整体结构

本文基于 STDC 网络提出了一种改进的嵌入注意力机制的轻量级编解码器结构, STDC 是一种轻量化的网络, 是一种用于实时语义分割的新型高效网络结构. 该网络结构旨在解决传统方法中存在的时间消耗和任务特定设计不足的问题, 通过去除冗余结构和引入细节聚合模块, 从而提高分割准确性和推理速度的平衡. 编码阶段通过逐步减小特征图分辨率获得更加准确的高级场景语义特征; 解码器阶段解码低分辨率的场景特征来获得与输入场景大小一致的场景语义类别信息, 如图 1 所示. 本文选取对语义分割效果较好、分割速度较快的 STDC 网络进行针对性改进. 编码器采用轻量化的网络 STDCNet 作为特征提取网络, 同时加入残差连接以更有效的结合来自高层和浅层的语义信息, 共有 5 个阶段. 前两个阶段负责采取底层特征, 仅使用一个卷积、归一化层和 ReLU 激活函数. 第 3、4、5 个阶段分别含有若干个 STDC 模块进行下采样, STDC 模块的具体组成将在第 2.2 节详细介绍. 5 个阶段之后经过注意力细化模块特征融合, 送入改进过的 PA-ASPP 模块. 解码器由多个上采样层即反卷积组成, 接受来自编码器提取出的特征实现不同尺度特征的融合, 将上下文语义信息结合起来.

2.2 编码器

本文引入了轻量化的 STDC 网络并进行改进作为特征提取的骨干网络. 如图 1 所示, 编码器部分有 5 个阶段构成, 每个阶段进行步长为 2 的下采样. 为了减少计算量, 前两个阶段只使用一个卷积块、BN 层和 ReLU

激活函数, 用于提取表层特征已足够. 第 3、4、5 个计算中每个阶段包含若干个 STDC 模块, 其中第 1 个 STDC 模块包含下采样操作, 其余的则保持特征图尺寸不变. 图 2 展示了 STDC 模块的结构, M 表示输入特征通道数, N 表示输出特征通道数, STDC 模块包括 4 个块, 除最后一个块外, 第 i 个块的输出通道数为 $x_{\text{output}} = F(x_1, x_2, \cdots, x_n)$, 最后一个块的输出特征通道数与倒数第 2 个保持一致.

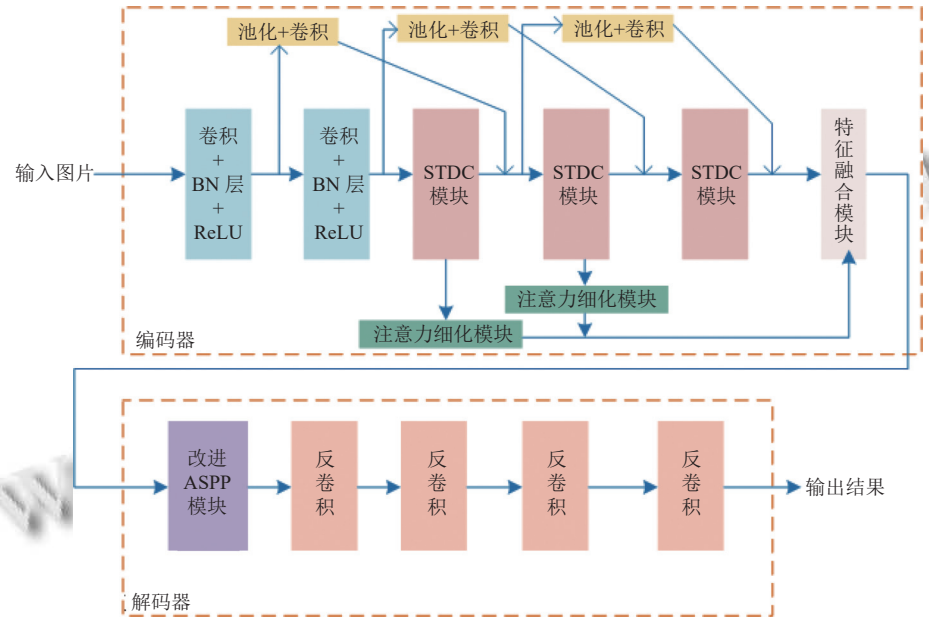

图 1 整体网络结构图

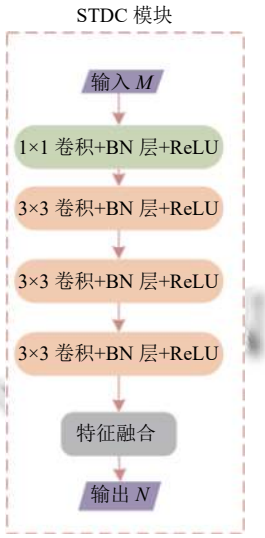

图 2 STDC 模块

STDC 模块的最终输出为各个块输出特征的融合, 即拼接在一起:

$$x_{\text{output}} = F(x_1, x_2, \cdots, x_n) \tag{1}$$

其中, F 表示融合函数, $x_1, x_2, \cdots, x_n$ 表示 n 个块的输出, x_{output} 表示 STDC 模块的输出.

同时, STDC 模块的参数量计算公式为:

$$\begin{aligned} Param =& M\times 1\times 1\times \frac{N}{2^1} + \sum_{i=2}^{n-1}\frac{N}{2^{i-1}}\times 3\times 3\times \frac{N}{2^i} \\ &+ \frac{N}{2^{n-1}}\times 3\times 3\times \frac{N}{2^{n-1}} \\ =& \frac{NM}{2} + \frac{9N^2}{2^3}\times \sum_{i=0}^{n-3}\frac{1}{2^{2i}} + \frac{9N^2}{2^{2n-2}} \\ =& \frac{NM}{2} + \frac{3N^2}{2}\times\left(1+\frac{1}{2^{2n-3}}\right) \end{aligned} \tag{2}$$

2.3 PA-ASPP 模块

在非结构化环境语义分割中出现的一些问题, 如可行驶的道路区域不明显等, 与感受野获取的上下文以及模型对道路的位置感知能力有一定相关性. 为了能够在网络深层继续提取图像详细, 在编码器和解码器之间加入了改进的 PA-ASPP 模块来进一步增强模型对道路的位置感知能力以及对细节边界的分割能力. PA-ASPP 模块是在 ASPP 的基础上每个支路嵌入位置注意力机制 (PA), 使网络关注到道路的位置感知信息. PA-ASPP 模块的结构如图 3 所示, 虽然 ASPP 能有效获取不同采样率获取的图像特征, 可以有效提升网络

整体的分割精度, 但是对于非结构化环境中的边界、位置等信息的分割能力有限. 因此通过在不同尺度特征信息后增添 PA 注意力机制, 弥补了 ASPP 的不足, 可以使用位置注意力机制来突出显示道路的位置, 从而更好地区分道路和其他物体.

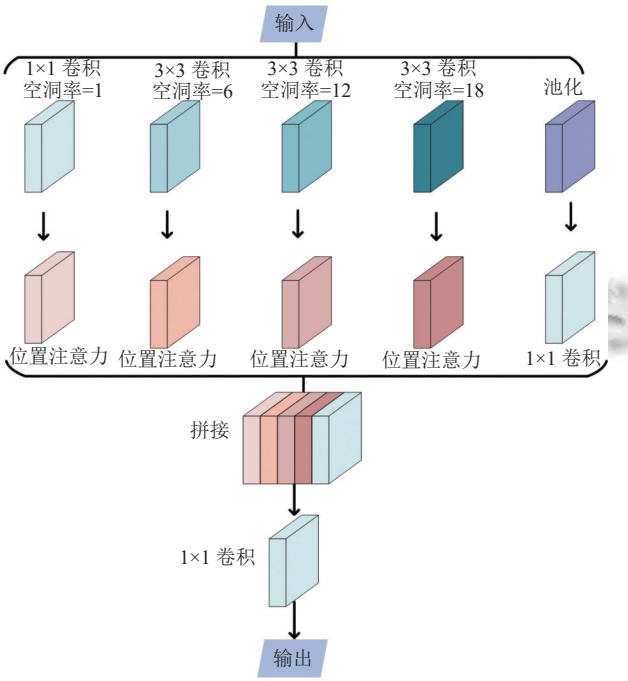

图 3 PA-ASPP 模块

其中, 输出 Y 可表示为:

$$Y = Conv_{1\times1}\left\{Concat\left\{\begin{array}{c} PA\left[Conv_{1\times1}(X)\right] \\ \sum_{i=6,12,18}^{n} PA\left[Conv_{3\times3}(X)\right] \\ Conv_{1\times1}\left[Pool(X)\right] \end{array}\right\}\right\} \tag{3}$$

2.4 解码器

输入图像经过 PA-ASPP 模块之后, 进入到解码器. 本文借鉴 U-Net[15]提出的编解码器思想, 设计了与编码器对称的解码器, 由多个反卷积构成, 每个反卷积块里面包含一个 1×1 卷积和反卷积, 反卷积的过程可以表示为:

$$y = TranConv_{1\times1}(Conv(x)) \tag{4}$$

2.5 损失函数

为了训练本文提出的模型, 使用添加 OHEM 算法的加权交叉熵损失函数 OHEMCELoss. 本文训练用的数据集中存在类别分布不均衡的问题, 例如 grass、tree、sky 等类别出现频率较高, mulch、rock、fence 等类别出现频次较低, 普通的交叉熵损失函数无法缓解这个问题. 于是本文在交叉熵损失函数的基础上添加了初始学习率设置为 0.01, 动量和权重的衰减率分别设定为 0.9 和 0.001. 优化器采用随机梯度下降 (stochastic gradient descent, SGD) 算法更新参数. OHEM 算法以获得更快速、更好的收敛, 缓解类不平衡问题. 在训练过程中关注 hard examples, 对其施加更高的权重. 不仅解决了正负样本类别不均衡问题, 同时提高了算法准确率. 本文设置第i类的权重为:

$$\omega_i = \frac{1}{\sqrt{P_i}} \times 10000 \tag{5}$$

其中, P_i为第i类对应的像素个数, 10 000 是经过多次尝试而确定的一个固定数值. 损失函数可表示为:

$$L_{\text{OHEMCE}} = -\sum_{i=1}^{N} \omega_i y_i \log(p_i) \tag{6}$$

其中, y_i表示第i类的真实标签 (为 0 或 1), p_i表示模型的预测概率, 即样本属于第i类的概率.

3 实验

3.1 数据集

RUGD 数据集[29]是一个为非结构化环境中的语义分割而定制的数据集. 它侧重于对非结构化户外环境的语义理解, 它是由移动机器人平台上的摄像机捕获的. 此数据集包含 4 759 张训练图片和 1 964 张测试图片, 有 24 个类别, 包括车辆、建筑、天空、草地等, 还包括 8 种独特的地形类型.

RELLIS 数据集[30]是另一个为非结构化环境中的语义分割而定制的数据集. 这些数据是在德克萨斯 A&M 大学的 Rellis 校区收集的, 对于不平衡和环境地形相关的现有算法提出了挑战. 它包含 3 302 张训练图片和 1 672 张测试图片, 有 19 个类别, 包括车辆、人、天空、建筑等.

3.2 训练环境

实验在一台配置有 GTX 3070 显卡和 16 GB 内存的计算机上运行. 操作系统为 Ubuntu 22.04, 使用 Python 3.7 作为语言环境, PyTorch 1.11 作为编译环境, CUDA 版本为 11.6. 为了使网络能够正常训练数据集, 我们选择了 STDC 的变体网络 (即 STDC1) 进行改进. 在训练过程中, 两个数据集的批量大小均设置为 8, 并进行了 150 个 epochs 的训练. 为了避免过拟合, 在训练

过程中还利用了数据增强技术, 包括水平翻转、缩放和旋转操作.

3.3 评价指标

网络分割精度评估指标为平均交并比 (mean intersection over union, *MIoU*) 与平均像素精确度 (mean pixel accuracy, *mAcc*), 轻量性评估指标为每秒浮点运算次数 (floating point operations per second, *FLOPs*) 与参数量 (parameters, *Param*).

IoU: 交并比在语义分割中表示真实值和预测值两个集合的交集与并集之比, 根据混淆矩阵可表示为:

$$IoU = \frac{TP}{FP+FN+TP} \tag{7}$$

平均交并比为对所有类别的 *IoU* 取平均值.

PixAcc: 像素精确度表示图像中正确分类的像素所占的百分比, 根据混淆矩阵可以表示为:

$$PixAcc = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

平均像素精确度为对所有类别的 *PixAcc* 取平均值.

FLOPs: 指浮点运算数, 也指计算量, 可以用来衡量模型的复杂度, 可以表示为:

$$FLOPs = \sum(2\times C_i \times K^2 - 1)\times H \times W \times C_o \tag{9}$$

其中, C_i、C_o表示输入和输出通道, K 为卷积核大小, H、W 是输出特征图的大小.

Param: 是指模型中需要学习的可调整参数的总数, 通常用来衡量模型的规模大小. 可表示如下:

$$Param = \sum C_i \times C_o \times K^2 \tag{10}$$

3.4 消融实验

验证各个模块对本文模型的影响, 本文进行 4 次训练. 其中, 第 1 次为原始的 STDC 作为编码器的模型, 第 2 次为增加了 PA-ASPP 模块, 第 3 次在第 2 次的基础上对 STDC 改进添加残差连接, 第 4 次在第 3 次的基础上使用 OHEMCELoss, 即本文提出的方法. 在两个数据集上的结果如表 1 所示.

表 1 消融实验结果

PA-ASPP	残差	OHEM-CELoss	RUGD		RELLIS		
			MIoU (%)	*mAcc* (%)	*MIoU* (%)	*mAcc* (%)	*FLOPs* (G)
×	×	×	32.33	37.82	37.45	40.58	4.46
√	×	×	49.04	55.14	42.03	45.26	6.45
√	√	×	51.25	57.44	47.45	51.50	6.82
√	√	√	50.78	56.28	49.96	55.88	6.82

分析结果可知, 在保证实时性的情况下, 各改进模块对于网络分割精度均有明显提升. 对于 RUGD 数据集, 增加 PA-ASPP 特征融合模块, *MIoU* 提高了 16.71%, 在 PA-ASPP 基础上增加多尺度残差模块, *MIoU* 进一步提升了 2.21%, 达到 51.25%, *mAcc* 达到 57.44%, 但是数据集本身类别不平衡问题不是很严重, 所以采用添加 OHEM 的交叉熵损失之后效果略微下降. 对于 RELLIS 数据集, 增加 PA-ASPP 特征融合模块, *MIoU* 提高了 4.58%, 在 PA-ASPP 基础上增加多尺度残差模块, *MIoU* 再次提升了 5.42%, 达到 47.45%, *mAcc* 达到 51.50%, 在采用了添加 OHEM 的交叉熵损失函数之后缓解了类别不平衡问题, *MIoU* 提高了 2.51%. 以上结果表明, PA-ASPP 模块和多尺度残差结构有效提升了分割精度, OHEMLoss 提升了 RELLIS 数据集的表现. 虽然本文添加了各个模块之后 *FLOPs* 和 *Param* 略增高, 但仍然满足实时性要求.

3.5 对比实验

将本文提出的方法与其他先进方法进行对比, 经过 150 轮训练后, 模型在 RUGD 和 RELLIS 验证集上的结果与现有模型结果对比分别如表 2 所示.

表 2 对比实验结果

Method	RUGD *MIoU* (%)	RELLIS *MIoU* (%)	*FLOPs* (G)	*Param* (M)
Memory-based[23]	37.71	45.61	243.21	42.13
DDRNet[9]	46.86	48.37	4.59	5.69
ConvNeXt[18]	49.17	38.64	80.44	87.99
TrSeg[31]	33.91	N/A	419.00	74.00
PSPNet[32]	31.78	N/A	463.00	72.00
UperNet[33]	31.95	N/A	N/A	N/A
Ours	50.78	49.96	6.82	10.86

3.5.1 RUGD 数据集对比

由表 2 可以观察到, 本文的算法在 RUGD 数据集上有较好的表现, 在验证集上 *MIoU* 达到了 50.78%, 高于其他方法. 值得注意的是, 在分割精度明显高于其他网络的同时, *FLOPs* 和 *Param* 远远低于其他网络. 证明了本文提出的网络能在保证实时性的情况下实现更高精度的语义分割.

图 4 为其他方法与本文方法在 RUGD 测试集上的可视化结果. 可见, 对于整体分割准确率, 本文提出的方法略优于其他方法, 虽然对于远景的小目标提取能力略差, 但是对于边缘的分割更贴近标签图像. 同时, 对于易混淆的类别区分能力也相对更强.

3.5.2 RELLIS 数据集对比

从表 2 中可以看出, 本文算法在 RELLIS 数据集上同样有良好的结果, 在测试集上的 *MIoU* 达到了 49.96%, 高于其他网络. 同时, *FLOPs* 和 *Param* 仅为 6.82G 和 10.86M. 有力证明了本文的改进对提升非结构化环境分割精度的有效性.

图 5 为其他方法与本文方法在 RELLIS 测试集上的可视化结果. 可见, 本文的方法在 RELLIS 数据集上同样也取到了不错的结果. 虽然对于远景中的小目标分割能力略弱, 但是整体分割效果优于其他的网络. 尤其在易混淆类别中, 其他网络对于水坑和水坑中的倒影容易错误分割, 而本文的方法有很强的能力将其分割出来.

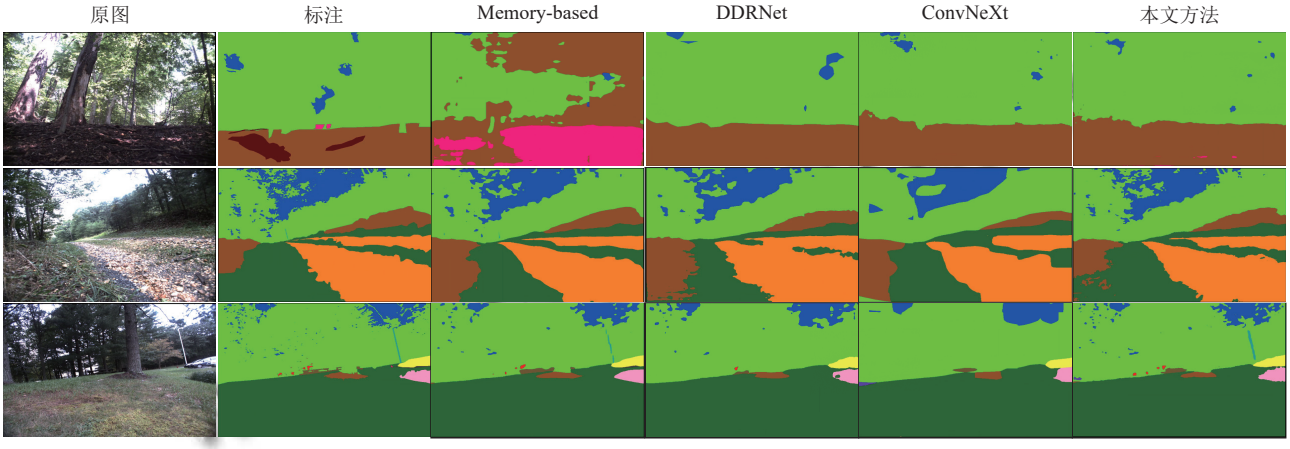

图 4 RUGD 测试集可视化结果

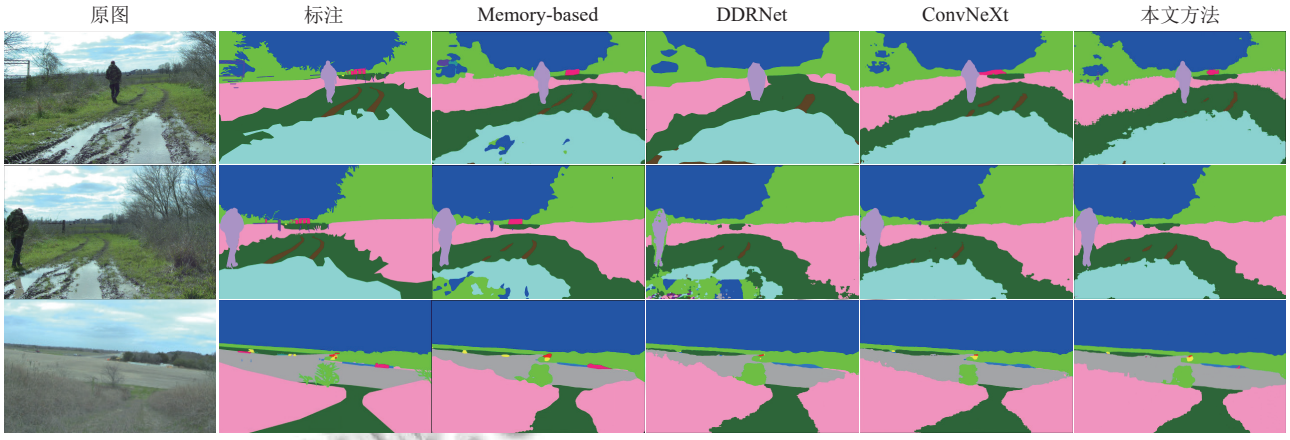

图 5 RELLIS 测试集可视化结果

4 结论与展望

本文针对非结构化环境中语义分割存在准确率低、实时性差等问题, 提出一种基于 STDC 的轻量级非结构化环境语义分割网络. 采用改进过的轻量级 STDC 作为特征提取网络, 在网络中加入残差块来融合不同尺度特征以增强全局感知能力, 同时, 在特征提取之后添加嵌入注意力机制的 PA-ASPP 模块来提升分割的精度, 数据集测试实验表明, 本文提出的网络对 RUGD 数据集的 *MIoU* 和 *mAcc* 分别达到了 50.78% 与 56.28%, 在 RELLIS 数据集的分割精度 *MIoU* 和 *mAcc* 分别达到了 49.96% 和 55.88%, 同时, 网络的 *FLOPs* 和 *Param* 仅为 6.82G 和 10.86M. 完全满足实时性要求并且能够部署在小型系统、边缘设备上. 可视化结果证明, 本文提出的网络在保证实时性的情况下实现了更好的精度. 未来研究工作中将关注远景中的小目标, 进一步优化网络结构的基础上, 实现对非结构化场景中小目标的语义信息的提取和分割.

参考文献

1 Shelhamer E, Long J, Darrell T. Fully convolutional

networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

2 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495. [doi: 10.1109/TPAMI.2016.2644615]

3 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

4 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 833–851.

5 张凯航, 冀杰, 蒋骆, 等. 基于 SegNet 的非结构道路可行驶区域语义分割. 重庆大学学报, 2020, 43(3): 79–87. [doi: 10.11835/j.issn.1000-582X.2020.03.009]

6 龚志力, 谷玉海, 朱腾腾, 等. 融合注意力机制与轻量化 DeepLabV3+的非结构化道路识别. 微电子学与计算机, 2022, 39(2): 26–33.

7 Zhao HS, Qi XJ, Shen XY, *et al*. ICNet for real-time semantic segmentation on high-resolution images. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 418–434.

8 艾青林, 张俊瑞, 吴飞青. 基于小目标类别注意力机制与特征融合的 AF-ICNet 非结构化场景语义分割方法. 光子学报, 2023, 52(1): 0110001.

9 Hong YD, Pan HH, Sun WC, *et al*. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv:2101.06085, 2021.

10 Fan MY, Lai SQ, Huang JS, *et al*. Rethinking bisenet for real-time semantic segmentation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9711–9720.

11 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

12 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR, 2016.

13 Lin D, Ji YF, Lischinski D, *et al*. Multi-scale context intertwining for semantic segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 622–638.

14 Chen BW, Chen LC, Wei YC, *et al*. SPGNet: Semantic prediction guidance for scene parsing. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 5217–5227.

15 Lin GS, Milan A, Shen CH, *et al*. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5168–5177.

16 Yu CQ, Wang JB, Peng C, *et al*. Learning a discriminative feature network for semantic segmentation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1857–1866.

17 Yu CQ, Wang JB, Peng C, *et al*. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 334–349.

18 Liu Z, Mao HZ, Wu CY, *et al*. A ConvNet for the 2020s. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11966–11976.

19 Chen JR, Kao SH, He H, *et al*. Run, don't walk: Chasing higher FLOPS for faster neural networks. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 12021–12031.

20 Cordts M, Omran M, Ramos S, *et al*. The cityscapes dataset for semantic urban scene understanding. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3213–3223.

21 Zhou BL, Zhao H, Puig X, *et al*. Semantic understanding of scenes through the ADE20K dataset. International Journal of Computer Vision, 2019, 127(3): 302–321. [doi: 10.1007/s11263-018-1140-0]

22 Baheti B, Gajre S, Talbar S. Semantic scene understanding in unstructured environment with deep convolutional neural network. Proceedings of the 2019 IEEE Region 10 Conference. Kochi: IEEE, 2019. 790–795.

23 Jin Y, Han D, Ko H. Memory-based semantic segmentation for off-road unstructured natural environments. Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague: IEEE, 2021. 24–31.

24 Chollet F. Xception: Deep learning with depthwise separable

convolutions. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1800–1807.

25 Baheti B, Innani S, Gajre S, *et al*. Semantic scene segmentation in unstructured environment with modified DeepLabV3+. Pattern Recognition Letters, 2020, 138: 223–229. [doi: 10.1016/j.patrec.2020.07.029]

26 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

27 Chen LC, Yang Y, Wang J, *et al*. Attention to scale: Scale-aware semantic image segmentation. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3640–3649.

28 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. arXiv:1711.07971, 2018.

29 Wigness M, Eum S, Rogers JG, *et al*. A RUGD dataset for autonomous navigation and visual perception in unstructured outdoor environments. Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau: IEEE, 2019. 5000–5007.

30 Jiang P, Osteen P, Wigness M, *et al*. RELLIS-3D dataset: Data, benchmarks and analysis. Proceedings of the 2021 IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021. 1110–1116.

31 Jin Y, Han D, Ko H. TrSeg: Transformer for semantic segmentation. Pattern Recognition Letters, 2021, 148: 29–35. [doi: 10.1016/j.patrec.2021.04.024]

32 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6230–6239.

33 Xiao TT, Liu YC, Zhou BL, *et al*. Unified perceptual parsing for scene understanding. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 432–448.

(校对责编: 牛欣悦)