计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(1):90−99 [doi: 10.15888/j.cnki.csa.009740] [CSTR: 32024.14.csa.009740]

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

融合注意力与多尺度特征的城市街景实例分割①

王　军[1,2], 吕　佳[1], 程　勇[1,2]

[1](南京信息工程大学 软件学院, 南京 210044)
[2](南京信息工程大学 科技产业处, 南京 210044)
通信作者: 吕　佳, E-mail: 202212490277@nuist.edu.cn

摘　要: 城市街道场景实例分割算法可以显著提升城市环境感知和智能交通系统的准确性与效率, 针对城市街景行人和车辆之间相互遮挡和背景干扰严重等问题, 提出一种基于频率注意力机制和多尺度特征融合的实例分割模型 FMInst. 首先, 构建一种高低频注意力机制进行交互编码从而增加高分辨率细节信息. 其次, 在 Swin Transformer 主干网络的 Patch Merging 层引入软池化操作, 减少特征信息损失, 有效提高小尺度目标分割结果. 最后, 结合 MLP 层构建多尺度的深度卷积, 有效增强目标局部信息提取, 提升实例分割精度. 在 Cityscapes 公共数据集进行对比实验, 结果表明 FMInst 的 mAP 提高 1.2%, 达 35.6%, 同时 AP50 提高 2.2%, 达 61.4%, 极大地改善实例分割的掩码质量和分割效果.

关键词: 城市街景; 实例分割; 频率注意力机制; 多尺度特征融合; 小目标

引用格式: 王军,吕佳,程勇.融合注意力与多尺度特征的城市街景实例分割.计算机系统应用,2025,34(1):90–99. http://www.c-s-a.org.cn/1003-3254/9740.html

Instances Segmentation of Urban Streetscape Incorporating Attention and Multi-scale Feature

WANG Jun[1,2], LYU Jia[1], CHENG Yong[1,2]

[1](School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](Science and Technology Industries Division, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: Algorithms for the instance segmentation of urban street scenes can significantly improve the accuracy and efficiency of urban environment perception and intelligent transportation system. To address mutual occlusions between pedestrians and vehicles and significant background interference in urban street scenes, this study proposes an instance segmentation model, FMInst, based on a frequency attention mechanism and multi-scale feature fusion. Firstly, a high and low-frequency attention mechanism is constructed for interactive coding to increase high-resolution detail information. Secondly, a soft pooling operation is introduced into the Patch Merging layer of the Swin Transformer backbone network to reduce the loss of feature information and effectively improve the segmentation of small-scale targets. Finally, an MLP layer is combined to construct multi-scale deep convolution, which effectively enhances the extraction of local information and improves the segmentation accuracy. Comparison experiments conducted on the public dataset Cityscapes show that FMInst reaches an mAP of 35.6%, with an improvement of 1.2%, and an AP50 of 61.4%, with an improvement of 2.2%. The mask quality and the segmentation effect of the instance segmentation are greatly improved.

Key words: urban streetscape; instance segmentation; frequency attention mechanism; multi-scale feature fusion; small target

① 基金项目: 国家自然科学基金 (41975183)
收稿时间: 2024-06-24; 修改时间: 2024-07-18; 采用时间: 2024-07-25; csa 在线出版时间: 2024-11-28
CNKI 网络首发时间: 2024-11-29

实例分割[1]已成为计算机视觉[2]研究中具有挑战性的任务之一, 旨在预测对象类标签和像素的对象实例掩码, 并定位各种图像中存在的不同类的对象实例, 被广泛应用于自动驾驶、医学分割和遥感等多个领域. 随着卷积神经网络 (CNN)[3]的出现, 衍生出许多实例分割模型, 模型分割效果也在不断提升. 受 Fast R-CNN[4]、Faster R-CNN[5]的启发, 全卷积网络 (FCN)[6]被提出并用于预测分割掩模.

目前实例分割的研究与应用越来越广泛, 但仍面临着很多挑战. 首先, 现有方法优化了掩模分割结果, 但在行人和车辆之间相互遮挡的问题中, 部分方法没有充分利用到对象边界信息, 导致掩模分割质量不够平滑. 此外, 模型在背景干扰严重时, 可能会错误地将背景区域包含在对象内部或将对象的一部分分类为背景, 从而导致分割结果不准确.

为了解决这些问题, 本文采用编码器-解码器结构, 提出了一种基于频率注意力机制和多尺度特征融合的实例分割网络 FMInst (frequency attention and multi-scale feature instance segmentation). 本文的主要贡献如下.

1) 设计一个新颖的高低频率注意力机制, 利用不同频率的特征信息, 获取特征图的更多细节, 以提升分割掩码的质量.

2) 设计一个软池化下采样分支, 在主干网络的补丁令牌下采样中引入软池化操作, 减少特征信息的损失, 提高小尺度对象的分割效果.

3) 设计一个深度卷积前馈神经网络模块, 使用不同尺度的深度卷积结合 MLP 层, 提取多尺度局部信息, 提升模型整体的分割效果.

4) 本文方法在 Cityscapes 数据集上进行实验并可视化, 实验结果表明, FMInst 性能远优于现有的先进方法, 能有效改善实例分割的效果, 并提高分割掩码质量.

1 相关工作

目前的实例分割技术主要分为两类: 两阶段实例分割和单阶段实例分割.

1.1 两阶段实例分割算法

自上而下的方法首先检测每个实例的区域, 然后将实例掩模划分到每个区域中. 2017 年, He 等人[7]提出了 Mask R-CNN 算法, 该算法在 Faster R-CNN 算法基础上, 把 RoI Pooling 替换为 RoI Align 方法, 取消取整操作, 保留浮点数, 并基于分类和回归添加一个掩码分支来预测每一个像素类别, 从而实现较好的分割效果. 但是 Mask R-CNN 过度依赖边界框的准确性, 导致一些小尺度的物体检测效果较差. 2018 年, Liu 等人[8]提出了 PANet, 通过额外引入一条自下而上的路径增强方法, 并采用自适应特征池操作, 使得每一层的特征信息直接传输到其他建议子网, 并为每个建议创建捕获不同视图的互补分支, 进一步改善了掩模预测结果. Sun 等人[9]于 2019 年提出了一种高分辨率分割网络 HRNet, 该网络不同于以往下采样-上采样的传统分割结构, 其在创建低分辨率特征图的前提下保留当前的高分辨率特征图, 并融合不同尺度的特征图, 取代以分类网络为基础的网络架构, 成为新的标准结构.

然而, 由于分割模型在下采样操作中丢失图像细粒度特征, 导致分割掩模非常粗糙. 2021 年, Zhang 等人[10]提出了 RefineMask, 该方法在特征金字塔中最高分辨率的特征图上构建新的语义头, 以生成细粒度的语义特征, 用于补充实例分割过程中丢失的细节, 提高分割掩模的质量. 同年, Tian 等人[11]重新设计实例分割中学习掩模的损失, 提出了 BoxInst, 仅使用训练的边界框注释, 来实现高质量的掩模级实例分割.

在 2023 年, Fang 等人[12]提出了光谱空间特征金字塔网络, 首次将实例分割和高光谱图像进行结合, 通过注意力机制和双向特征金字塔结构, 在特征提取阶段融合多尺度光谱信息和多尺度空间信息, 从而提高网络模型性能. Wang 等人[13]提出了 CutLER, 用于训练无监督对象检测和分割模型, 利用自监督模型的特性发现对象, 并将其放大以训练最先进的本地化模型. 2024 年, Ouyang 等人[14]提出了 MixingMask, 该方法通过利用基于轮廓的分割方法来特别关注边界特征, 减少了对边界特征的忽视, 并在对象检测和实例分割任务之间形成了隐式关联.

自下而上的方法首先预测每个像素的类别标签, 然后使用聚类或其他度量学习方法将它们分组形成实例分割结果. 2017 年, Liu 等人[15]提出了顺序分组 SGN, 模型分为 3 个神经网络, 第 1 个网络通过预测水平和垂直对象断点来对每个图像行和列的像素进行分组, 并创建水平和垂直线段, 第 2 个网络将这些线段分组为连接的组件, 最后一个网络将组件合并为连贯的对象实例. 由于多个单独的子任务会降低应用的潜力, 2019 年, Gao 等人[16]提出了 SSAP, 该方法使用亲和力

金字塔, 以分层方式计算两个像素属于同一实例的概率, 并提出一种级联图划分模块, 按照从粗到细的顺序生成实例掩模. 2021 年, Hu 等人[17]提出第 1 个端到端框架, 称为 ISTR, 该模型利用循环细化策略同时进行检测和分割, 通过预测低维掩码嵌入矩阵, 并将其与集合损失的真实掩码嵌入矩阵进行匹配, 提高了检测和分割任务的性能. 为了提高分段掩模的质量, Ke 等人[18]提出了 Mask Transfiner, 该算法通过检测、分解图像区域, 构建分层四叉树并预测最终标签.

综上, 两阶段处理方法具有更好的分割效果和灵活性, 适合高精度和高质量分割的应用, 如医学影像分析、自动驾驶中的细节识别等. 但也存在包括时间成本较大、模型结构复杂、处理小目标效果较差和训练数据需求大等不足.

1.2 单阶段实例分割算法

与两阶段实例分割方法不同, 单阶段实例分割算法通过简化结构和减少计算开销, 达到高效的分割效果. 2019 年, Bolya 等人[19]提出了 YOLACT 算法, 它将实例分割分解为两个并行子任务: 生成一组原型掩模并预测每个实例的掩模系数. 然后通过线性组合生成实例掩码. 此外, 为了提高分支的运行速度, 还提出了快速非极大值抑制算法, 实现实时实例分割. 2020 年, Xie 等人[20]提出了 PolarMask, 该方法将实例分割问题表述为通过实例中心分类和极坐标中的密集距离回归来预测实例的轮廓. 并提出两种有效的方法来处理高质量中心样本和优化密集距离回归, 显著提高性能并简化训练过程. 同年, Chen 等人[21]提出了 BlendMask, 通过丰富实例级信息并执行更细粒度的位置敏感掩模预测来提升掩模质量. Wang 等人[22]在 SOLO 方法上进一步研究, 提出了 SOLOv2, 新框架中对象掩模生成被解耦的掩模核预测和掩模特征学习, 分别负责生成卷积核和待卷积的特征图, 其次采用矩阵非极大值抑制技术来减少开销.

2021 年, Cheng 等人[23]提出了统一框架 MaskFormer, 该方法预测一组二进制掩模, 每个掩模与单个全局类标签预测相关联, 并且可以将任何现有的每个像素分类模型无缝转换为掩模分类. 但 MaskFormer 在处理细粒度的实例分割任务和复杂的场景时, 存在特征提取能力不足的问题, 于是 Cheng 等人[24]对 MaskFormer 进行改进, 提出了 Mask2Former, 该算法基于简单的元架构和新的 Transformer 解码器, 在 Transformer 解码器中使用屏蔽注意力, 将注意力限制在预测段为中心的局部特征上, 接着使用多尺度高分辨率特征, 帮助模型分割小对象或区域, 然后切换自注意力和交叉注意力的顺序, 使查询特征可学习, 最后, 通过在 K 个随机采样点上计算掩码损失, 在不影响性能的情况下节省了 3 倍的训练内存. 虽然 Mask2Former 的分割精度很高, 但 Mask2Former 需要通过大量的解码器层来解码对象查询, 耗费大量的时间. 于是在 2023 年, He 等人[25]提出了 FastInst, 该模型遵循 Mask2Former 的元架构, 采用实例激活引导查询、双路径更新策略和地面真值掩模引导学习, 在使用更轻的像素解码器、更少的 Transformer 解码器层的情况下获得更好的性能.

2024 年, Gu 等人[26]提出了 DiffusionInst, 该框架将实例表示为向量, 模型经过训练可以反转噪声真实掩模, 不会产生区域选取框的任何归纳偏差. 它采用随机生成的向量作为输入和输出掩码, 在推理过程中进行多步去噪.

综上所述, 单阶段实例分割算法比两阶段实例分割方法更简洁、高效, 并且分割速度更快, 适用于高效实时处理的应用, 如实时视频处理、移动设备上的图像处理等. 但由于单阶段方法缺少多轮优化环节, 存在精度较低的缺点, 因此本文提出一种精确的单阶段实例分割框架.

2 FMInst 算法设计

2.1 总体架构

本文采用 Mask2Former 作为基线, 该模型由 3 部分组成: 主干特征提取器、像素解码器和 Transformer 解码器, 如图 1 所示. Swin Transformer[27]主干特征提取器用于从输入图像中提取低分辨率特征. 像素解码器[28]用于将编码器输出的特征图转换成像素级别的预测结果. Transformer 解码器[29]用于对预测结果进行对象查询操作, 生成最终的二进制掩模预测结果.

Mask2Former 相比于传统的基于编码器-解码器的图像实例分割模型, 使用了掩码注意力和多头自注意力机制, 自适应地捕捉图像中的不同特征, 但由于只捕捉到了低分辨率特征而忽略了高分辨率特征, 造成细节信息丢失的问题, 同时 Swin Transformer 编码器在下采样过程中丢失大量特征信息, 导致无法精确分割小目标的细节信息, 造成分割掩码质量较差、分割效果不好等问题. 本文针对这些问题, 对 Mask2Former 训练

方法进行改进, 如图 2 所示.

首先, 在保留原 Transformer 解码器的基础上, 将多头自注意力机制替换为高低频率注意力 (high and low frequency attention, HLFAM) 机制, 通过利用不同频率的特征进行编码互补, 即高频捕捉局部细节, 低频聚焦全局结构, 以减少计算成本, 提高模型效率. 其次, 在与主编码器的特征分辨率匹配的同时, 通过构建基于软池化的补丁合并 (softpool-based patch merging, SPM) 模块缩短补丁令牌的长度来形成分层特征编码结构, 以获得多尺度特征, 减少小尺度特征的遗漏. 接着, 通过引入深度卷积前馈神经网络 (depthwise convolution feedforward network, DCFN) 模块, 在两个线性层之间分别插入不同尺度的深度卷积, 以增强多尺度局部信息的提取, 提升分割的效果.

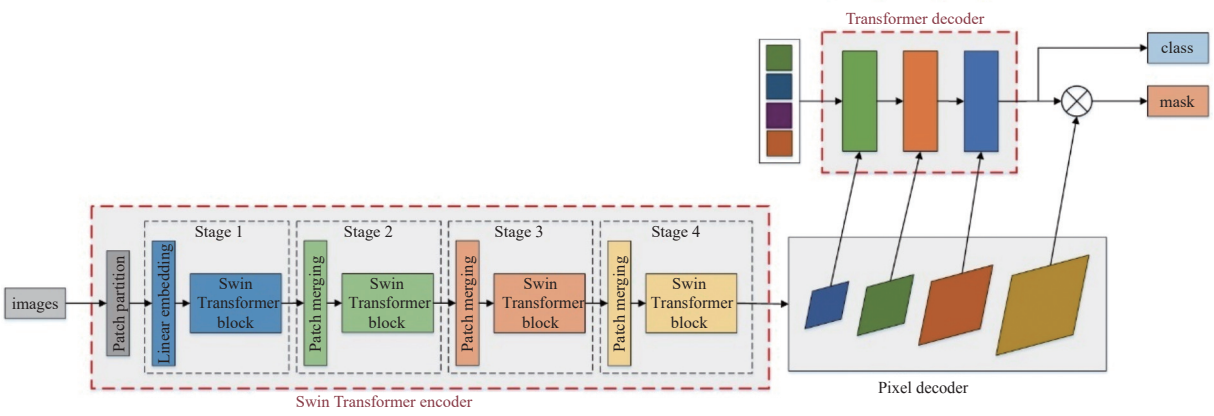

图 1 Mask2Former 网络结构

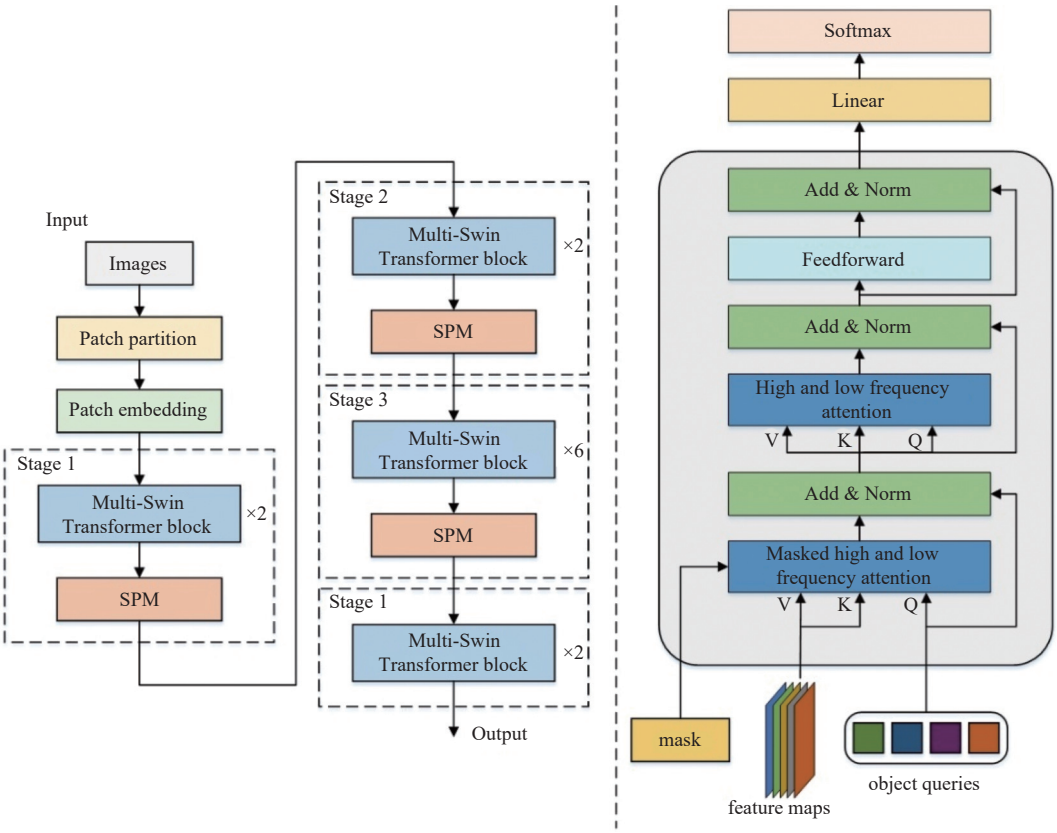

图 2 FMInst 网络架构

2.2 高低频率注意力机制 HLFAM

城市街景图像中蕴含丰富的频谱信息, 其中高频区域包含实例的局部细节 (如线条、形状等), 低频区域聚焦全局结构 (如纹理、色彩等). 然而模型使用全局注意力, 忽略了不同频率特征信息的差异. 同时, 传统的 multi-head self attention[30]在处理高分辨率图像时, 计算量非常大. 针对上述问题, 本文提出一种全新的高效率注意力机制 HLFAM, 如图 3 所示.

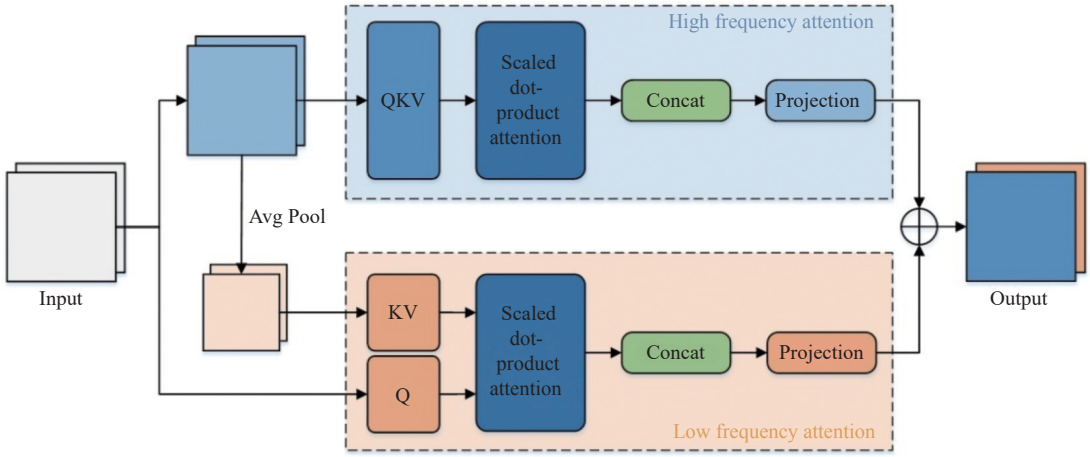

图 3 HLFAM 模块

HLFAM 通过构建两条分支, 分别实现高低不同频率的特征处理. 具体来说, 一条路径利用高分辨率的特征对数据进行深度映射. 确保输入数据具有丰富的特征描述. 接着, 模型将注意力集中在高频部分. 为了捕捉细粒度的高频信息, 采用 LW-SA (local window self-attention) 技术[31]. 通过使用 2×2 的窗口尺寸, 聚焦于每个局部窗口内的上下文关系. 此外, 另一条路径负责处理低频特征. 首先, 将平均池化应用于每个局部窗口中, 以降低局部窗口内特征的维度, 从而提取出更为通用的低频信息. 接着, 通过把剩余的 head 分配给低频注意力, 对低频特征及其对应的查询位置之间的潜在关系进行建模.

最后, 将经过处理的高频特征和低频特征进行拼接, 并将合并后的特征图传输给下一层. 由于高频注意力和低频注意力只采用了简单的非重叠串行方式, 简化了窗口移位和多尺度窗口划分等操作. 因此在利用高低频率信息的同时, 大大提高了模型的效率, 节省了大量的计算时间成本.

2.3 软池化分支结构 SPM

本文在原有的 Swin Transformer 模块 patch merging 层中引入软池化[32]操作, 通过设计特定的令牌, 对原始输入的图像进行下采样, 同时保留图像中局部纹理、边缘细节以及颜色分布等特征信息, 减少特征信息的损失, 提高小尺度对象的分割效果.

如图 4 所示, SPM 模块具有两个分支, 其中一个分支引入扩张卷积的瓶颈块, 通过扩大卷积的感受野来收集小尺度物体的特征和结构信息, 在瓶颈块中, 第 1 个 1×1 卷积层用来增加维数, 中间的 3×3 膨胀卷积层在保持数据结构不变的情况下扩大感受野, 用于获得更多的结构信息, 最后一个 1×1 卷积层用来减小特征尺度. 对于给定 n 阶段的输入特征 s, 则第 1 个分支的输出特征为 $F_1 \in \mathbb{R}^{(h/2)\times(w/2)\times 2c_1}$. F_1 如式 (1) 所示:

$$F_1 = \varphi(BN(DConv(\varphi(s)))) \tag{1}$$

其中, $\varphi(\cdot)$表示具有批量归一化和 GELU 的 1×1 卷积.

另一个分支引入软池化操作以获得更好的表征能力. 软池化操作以指数加权的方式激活池内核中的像素, 保存更详细的特征信息. 对于特定核邻域 R 中的每个像素, 软池化操作的计算方法如式 (2) 所示:

$$\tilde{s} = \sum_{i\in R} \frac{e^{s_i}\times s_i}{\sum_{j\in R} e^{s_j}} \tag{2}$$

然后, 将软池化后的特征输入卷积层 (增加维数) 以获得输出 $F_2 = \mathbb{R}^{(h/2)\times(w/2)\times 2c_1}$. F_2 如式 (3) 所示:

$$F_2 = \varphi(SoftPool(s)) \tag{3}$$

其中, $\varphi(\cdot)$表示具有批量归一化和 GELU 的 1×1 卷积.

综上, SPM 模块的功能是获取小尺度特征, 并保存更多特征细节. 扩展卷积分支和软池化操作分支的加

权和作为 SPM 的输出L. 过程如式 (4) 所示:

$$L = F_1 \oplus F_2 \tag{4}$$

其中, $\oplus$代表元素级加法.

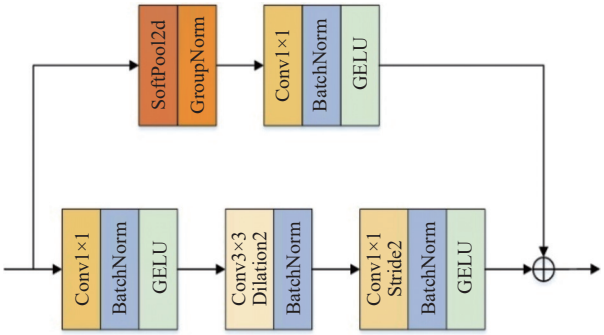

图 4 SPM 分支结构

2.4 深度卷积前馈神经网络 DCFN

由于 Transformer 解码器中使用位置编码来确定每个补丁的位置时, 而位置编码是从开始训练时就已经固定, 因此导致模型测试时, 给出一个不同分辨率的图片, 位置编码就会做插值处理, 造成测试样例出现误差, 达不到训练时的分割效果.

针对上述问题, 提出了 DCFN 模块, 该模块主要是由深度卷积 (DC)[33]和多层感知机层 (MLP)[34]组合而成. 深度卷积通过堆叠多个卷积层, 其浅层卷积层捕捉低级特征 (如边缘、纹理等), 深层卷积层提取更复杂和抽象的高级特征 (如物体的形状、类别等), 逐层提取图像的不同层次特征. 通过增加卷积层的深度, 模型可以学习到更加丰富和复杂的特征表示. 深度卷积中卷积层的权重是共享的, 这种权重共享机制显著减少了模型的参数量, 使得深度卷积网络在不增加过多计算资源的情况下, 能够实现更好的分割效果.

如图 5 所示, 通过在两个线性层之间插入两条多尺度深度卷积路径, 以减少零填充导致的位置信息丢失. 在 LayerNorm 之后, 我们以 r 的比例扩展通道, 将其分成两个分支, 3×3 和 5×5 深度卷积 (DWConv) 用于增加原 Swin Transformer 编码器的多尺度局部信息提取, 接着对不同尺度得到的特征信息进行拼接, 然后进行 *GELU* 非线性激活函数和全连接层操作得到最后的输出特征图. 这个过程如式 (5)–式 (7) 所示:

$$F_{\text{out1}} = Conv_{3\times3}(MLP(F_{\text{in}})) \tag{5}$$

$$F_{\text{out2}} = Conv_{5\times5}(MLP(F_{\text{in}})) \tag{6}$$

$$F_{\text{out}} = MLP(GELU(F_{\text{out1}} + F_{\text{out2}})) + F_{\text{in}} \tag{7}$$

其中, F_{in}是来自编码器的特征图, F_{out}是 DCFN 模块的输出特征图, *GELU* 表示非线性激活函数, *MLP* 是全连接层, $Conv_{3\times3}$表示 3×3 深度卷积, $Conv_{5\times5}$表示 5×5 深度卷积.

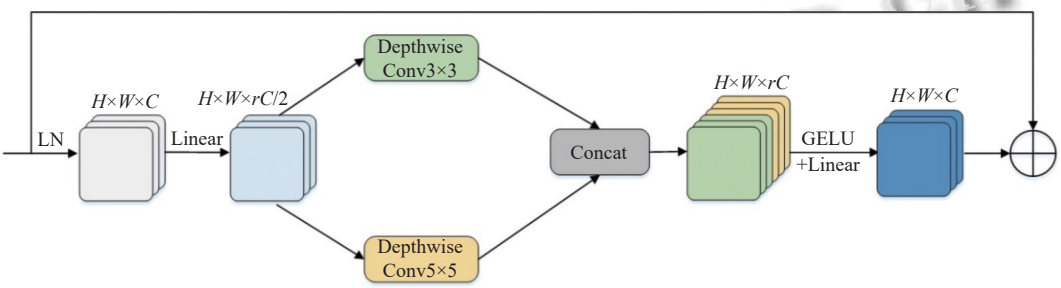

图 5 DCFN 模块

3 实验结果与分析

3.1 实验数据集和评估指标

本文的主要实验是在开源的 Cityscapes 数据集[35]上进行的. Cityscapes 数据集分别包含 2 975 张训练图像、500 张验证图像和 1 525 张测试图像. 图像由车载摄像头采集, 包含 50 个不同城市不同天气条件和季节的街景数据. Cityscapes 数据集通常为 2048×1024 像素, 每张图像都附带类别标签和实例分割标签, 使得模型能够区分不同的对象实例.

本文实验所用的评价指标为 mAP 和 AP50. 在实例分割任务中, mAP 是一种常用的评价指标, 用于衡量模型在检测和分割对象时的性能. AP50 是指以 0.5 作为 IoU 阈值的平均精度.

3.2 实验环境以及参数设置

本文的实验环境为 Ubuntu 18.04 系统, 使用的语言开发工具为 Python 3.8、深度学习框架为 PyTorch

1.9.0, Cuda 版本为 11.1, GPU 版本为 NVIDIA GeForce RTX 3090, 运行显存为 24 GB.

本文方法是依赖于 Detectron2 进行搭建的. 采用 AdamW 优化器和递减学习速率调度器在单个 GPU 上进行训练. 对于主干网络, 设置初始学习率为 2.5E–5, 批量大小为 4, 动量为 0.9, 权重衰减为 0.05, 正则化系数为 5E–4, 迭代 400k 次. 对于数据增强, 使用 LSJ 增强, 随机尺度为 0.1–0.2 范围采样, 将图像裁剪为固定大小为 1024×1024.

3.3 消融实验结果与分析

为了评估所提出的网络结构和 3 个重要模块的性能, 本文应用 Mask2Former 作为基线网络, 并使用 ImageNet 中的预训练权重来初始化其参数, 在 Cityscapes 验证集上进行消融实验, 实验对比结果如表 1 所示.

表 1 消融实验结果对比 (%)

Model name	Modules			Evaluation index	
	SPM	HLFAM	DFFN	mAP ↑	AP50 ↑
Baseline	—	—	—	34.38	59.20
Baseline + HLFAM	—	√	—	35.00	60.26
Baseline + SPM	√	—	—	34.83	60.00
Baseline + DCFN	—	—	√	35.46	60.90
Baseline + SPM + HLFAM + DCFN	√	√	√	**35.58**	**61.41**

通过表 1 可知, 单独添加 HLFAM 模块, mAP 值提高了 0.62%. 与基线模型相比, 该模块不仅关注低分辨率特征信息, 还额外增加了高频率注意力, 以捕获更多的高分辨率细节信息, 充分利用不同频率的特征信息能够显著提升模型的分割效果. 通过单独添加 SPM 模块, mAP 值提高了 0.45%. 与基线模型相比, 该模块引入软池化操作, 保留了城市街景图像中的更多细节信息, 避免了在 patch merging 层过程中特征信息的丢失, 提高了小尺度对象的分割效果. 通过单独引入 DCFN 模块, mAP 提高 1.08%. 与基线模型相比, 该模块在每个前馈神经网络中添加了一层带有零填充的深度卷积, 能够提取更多不同尺度的局部特征, 同时减少了模型的参数量和计算量. 从表 1 最后一行可以看出, 同时引入 3 个模块后, mAP 值提高了 1.2%, 可视化结果也证明了这些模块在提升模型性能方面的有效性.

本文在城市街景数据集上, 还对比了不同实例在基线模型和 FMInst 的分割效果, 对比结果如表 2 所示. 从表 2 中可以看出本文方法在 6 个类别的分割效果上有所提升.

表 2 城市街景类别分割结果对比 (%)

Class name	mAP ↑		AP50 ↑	
	Mask2Former	FMInst	Mask2Former	FMInst
person	30.3	31.6	60.9	62.3
rider	22.9	25.8	58.4	62.6
car	55.5	55.8	81.9	82.3
truck	36.0	33.9	50.5	47.3
bus	56.6	57.9	71.7	74.5
train	38.2	42.0	57.7	63.5
motorcycle	15.9	18.6	40.4	47.8
bicycle	19.6	19.0	52.1	50.9

3.4 对比实验结果与分析

为了验证本文提出的方法的有效性, 本文采用编码器-解码器结构, 将 FMInst 算法与其他实例分割方法进行了比较, 结果分别如表 3 和表 4 所示.

表 3 单阶段算法的实验结果对比 (%)

Method	Model	mAP ↑	AP50 ↑
One-stage	YOLACT	23.5	39.9
	PolarMask	30.5	56.9
	CondInst	33.2	57.2
	SOLOv2	32.6	59.0
	Mask2Former	34.4	59.2
	Boxteacher	26.8	54.2
	Box2mask	22.7	46.6
Ours	FMInst	**35.6**	**61.4**

表 4 两阶段算法的实验结果对比 (%)

Method	Model	mAP ↑	AP50 ↑
Two-stage	Mask R-CNN	26.2	49.9
	PANet	31.8	57.1
	SSAP	32.7	51.8
	E2EC	32.9	59.2
	PolySnake	34.3	61.0
	FSFnet	34.1	52.7
	BGF	33.4	59.8
Ours	FMInst	**35.6**	**61.4**

从表 3 和表 4 的结果所示, FMInst 算法分割的效果都明显优于其他算法. FMInst 算法通过使用高低频率注意力机制, 丰富了高分辨率的特征信息, 提高对小目标分割的准确性; 引入软池化操作, 提高模型的特征表达能力; 将不同尺度的深度卷积和前馈神经网络相结合, 增强上下文信息, 解决多尺度问题; 使用大规模抖动增强的数据增强方法改善对遮挡目标的分割效果. 相较于单阶段的 YOLACT 算法[19]、PolarMask 算法[20]、CondInst 算法[36]、SOLOv2 算法[22]、Mask2Former 算法[24]、Boxteacher 算法[37]和 BoxLevelset 算法[38]在平均精度上分别提高了 12.1、5.1、2.4、3、1.2、8.8 和 12.9 个百分点. 相较于两阶段的 Mask R-CNN 算法[7]、PANet 算法[8]、SSAP 算法[16]、E2EC 算法[39]、Poly-

Snake 算法[40]、FSFnet 算法[41]和 BGF 算法[42]在平均精度上分别提高了 9.4、3.8、2.9、2.7、1.3、1.5 和 2.2 个百分点.

在 Cityscapes val 数据集上分别对基线模型和本文算法进行可视化处理, 结果如图 6 所示.

图 6 中第 (1) 列为输入的原图, 第 (2) 列为 Mask2-Former 算法分割的输出图像, 第 (3) 列为 FMInst 算法分割的输出图像. 从 (a) 行可知, FMInst 算法能大大的改善实例之间的相互遮挡; 从 (b) 行可知, FMInst 算法能改善实例分割的掩码质量; 从 (c) 行可知, FMInst 算法能解决实例分割后掩码边缘模糊的问题; 从 (d)、(e) 行可知, FMInst 算法能解决因为光线导致实例分割模型出现一些实例未分割和错误的实例分割问题; 从 (f) 行可知, FMInst 算法能提高小目标分割的准确度; 从 (g) 行可知, 在面对复杂的城市街道场景时, FMInst 算法的分割效果更优.

综合以上分析, 本文提出的 FMInst 算法在城市街景图像中的分割质量及分割数量方面都要优于 Mask2-Former 算法, 有效改善模型的分割效果, 减少实力分割掩模的误差, 提升了分割的精度和掩码的分割质量.

图 6 可视化对比图

4 结束语

本文在 Mask2Former 算法的基础上进行改进, 提出了一个新的 FMInst 算法. 该算法在 Transformer 解码器中构建 HLFAM 机制, 对高分辨率的特征进行深度映射, 对低分辨率的特征进行建模, 分别实现高低不同频率的特征处理, 提高模型分割效率, 同时在 Swin Transformer 编码器的 patch merging 层中引入 SPM 分支, 以指数加权的软池化方式激活池内核的像素, 保存更多的特征信息, 提高小尺度对象的分割效果, 最后提出 DCFN 模块, 通过堆叠不同尺度的深度卷积进行权重共享的局部特征信息提取, 在减少模型参数量的同时提升分割的精度. 改进后的 FMInst 模型在 mAP 上达到 35.6%, AP50 上达到 61.4%, 相比原始基线模型分别提高 1.2%、2.2%. 在未来的工作中, 将通过对模型

进行轻量化设计、对损失函数进行优化来进行进一步的研究.

参考文献

1 苏丽, 孙雨鑫, 苑守正. 基于深度学习的实例分割研究综述. 智能系统学报, 2022, 17(1): 16–31. [doi: 10.11992/tis.202109043]

2 陈洛轩, 林成创, 郑招良, 等. Transformer 在计算机视觉场景下的研究综述. 计算机科学, 2023, 50(12): 130–147. [doi: 10.11896/jsjkx.221100076]

3 Li ZW, Liu F, Yang WJ, *et al*. A survey of convolutional neural networks: Analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 6999–7019. [doi: 10.1109/TNNLS.2021.3084827]

4 Girshick R. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448.

5 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: ACM, 2015. 91–99.

6 Wang JF, Song L, Li ZM, *et al*. End-to-end object detection with fully convolutional network. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15844–15853.

7 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988.

8 Liu S, Qi L, Qin HF, *et al*. Path aggregation network for instance segmentation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8759–8768.

9 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 5686–5696.

10 Zhang G, Lu X, Tan JR, *et al*. RefineMask: Towards high-quality instance segmentation with fine-grained features. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 6857–6865.

11 Tian Z, Shen CH, Wang XL, *et al*. BoxInst: High-performance instance segmentation with box annotations. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5439–5448.

12 Fang LY, Jiang YF, Yan YL, *et al*. Hyperspectral image instance segmentation using spectral-spatial feature pyramid network. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5502613.

13 Wang X, Girdhar R, Yu SX, *et al*. Cut and learn for unsupervised object detection and instance segmentation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 3124–3134.

14 Ouyang WZ, Xu ZL, Xu J, *et al*. MixingMask: A contour-aware approach for joint object detection and instance segmentation. Pattern Recognition, 2024, 155: 110620. [doi: 10.1016/j.patcog.2024.110620]

15 Liu S, Jia JY, Fidler S, *et al*. SGN: Sequential grouping networks for instance segmentation. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 3516–3524.

16 Gao NY, Shan YH, Wang YP, *et al*. SSAP: Single-shot instance segmentation with affinity pyramid. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 642–651.

17 Hu J, Cao LJ, Lu Y, *et al*. ISTR: End-to-end instance segmentation with Transformers. arXiv:2105.00637, 2021.

18 Ke L, Danelljan M, Li X, *et al*. Mask transfiner for high-quality instance segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4402–4411.

19 Bolya D, Zhou C, Xiao FY, *et al*. YOLACT: Real-time instance segmentation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9156–9165.

20 Xie EZ, Sun PZ, Song XG, *et al*. PolarMask: Single shot instance segmentation with polar representation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12190–12199.

21 Chen H, Sun KY, Tian Z, *et al*. BlendMask: Top-down meets bottom-up for instance segmentation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8570–8578.

22 Wang XL, Zhang RF, Kong T, *et al*. SOLOv2: Dynamic and fast instance segmentation. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: ACM, 2020. 1487.

23 Cheng BW, Schwing AG, Kirillov A. Per-pixel classification is not all you need for semantic segmentation. Proceedings of the 35th International Conference on Neural Information

Processing Systems. 2021. 1367.
24 Cheng BW, Misra I, Schwing AG, *et al*. Masked-attention mask Transformer for universal image segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 1280–1289.
25 He JJ, Li PY, Geng YF, *et al*. FastInst: A simple query-based model for real-time instance segmentation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 23663–23672.
26 Gu ZX, Chen HX, Xu ZE. Diffusioninst: Diffusion model for instance segmentation. Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE, 2024. 2730–2734.
27 Liu Z, Hu H, Lin YT, *et al*. Swin Transformer V2: Scaling up capacity and resolution. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11999–12009.
28 Zou XY, Dou ZY, Yang JW, *et al*. Generalized decoding for pixel, image, and language. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 15116–15127.
29 Chen Z, Zhang J, Tao DC. Recurrent glimpse-based decoder for detection with Transformer. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5250–5259.
30 Liu J, Chen SW, Wang BQ, *et al*. Attention as relation: Learning supervised multi-head self-attention for relation extraction. Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. Yokohama: ACM, 2021. 524.
31 Vaswani A, Ramachandran P, Srinivas A, *et al*. Scaling local self-attention for parameter efficient visual backbones. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12889–12899.
32 Stergiou A, Poppe R, Kalliatakis G. Refining activation downsampling with SoftPool. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 10337–10346.
33 Meena G, Mohbey KK, Indian A, *et al*. Identifying emotions from facial expressions using a deep convolutional neural network-based approach. Multimedia Tools and Applications, 2024, 83(6): 15711–15732.
34 Gao JT, Zhao XY, Li MY, *et al*. SMLP4Rec: An efficient all-MLP architecture for sequential recommendations. ACM Transactions on Information Systems, 2024, 42(3): 86.
35 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.
36 Tian Z, Shen CH, Chen H. Conditional convolutions for instance segmentation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 282–298.
37 Cheng TH, Wang XG, Chen SY, *et al*. BoxTeacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 3145–3154.
38 Li WT, Liu WY, Zhu JK, *et al*. Box2Mask: Box-supervised instance segmentation via level-set evolution. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(7): 5157–5173. [doi: 10.1109/TPAMI.2024.3363054]
39 Zhang T, Wei SQ, Ji SP. E2EC: An end-to-end contour-based method for high-quality high-speed instance segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4433–4442.
40 Feng H, Zhou KY, Zhou WG, *et al*. Recurrent generic contour-based instance segmentation with progressive learning. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(9): 7947–7961. [doi: 10.1109/TCSVT.2024.3383829]
41 Du CF, Liu PX, Song XS, *et al*. A two-pipeline instance segmentation network via boundary enhancement for scene understanding. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 4504913.
42 Gao LC, Wang SJ, Chen SG. A novel boundary-guided global feature fusion module for instance segmentation. Neural Processing Letters, 2024, 56(2): 91. [doi: 10.1007/s11063-024-11564-6]

(校对责编: 张重毅)