控制与决策
Control and Decision

基于动态卷积和超图交互的多实例人体解析方法

黄荣, 袁家奇, 刘浩, 蒋学芹, 周树波

引用本文:
黄荣, 袁家奇, 刘浩, 等. 基于动态卷积和超图交互的多实例人体解析方法[J]. 控制与决策, 2026, 41(1): 276-288.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0303

基于动态卷积和超图交互的多实例人体解析方法

黄 荣[1,2†]，袁家奇[1]，刘 浩[1,2]，蒋学芹[1,2]，周树波[1,2]

(1. 东华大学 信息科学与技术学院，上海 201620；
2. 东华大学 数字化纺织服装技术教育部工程研究中心，上海 201620)

摘 要: 多实例人体解析旨在分割自然场景图像中的多个人体实例及其部件. 现有方法通常依赖静态卷积核并行地分割部件和实例, 导致部件与实例特征缺乏关联难以适应人体姿态和服装外观的多样性. 针对该问题, 提出一种基于动态卷积与超图交互的多实例人体解析方法. 首先, 将分割目标划分为部件、半身、实例 3 种层次, 并对应地配置可学习的动态卷积核; 同时, 设计多尺度掩码注意力机制来引导各层次动态卷积核聚合图像特征, 以适应人体姿态和服装外观的多样性. 然后, 提出超图交互模块, 将部件动态卷积核作为节点, 实例和半身动态卷积核作为超边, 以刻画人体结构先验. 最后, 通过超图上的消息传递来实现部件与实例间的特征交互. 实验结果表明, 所提出方法在 MHP-v2.0、CIHP 和 Densepose 数据集上可超越多种基线方法, 在AP^{p}_{50}、$\mathrm{AP}^{p}_{\mathrm{vol}}$和$\mathrm{PCP}_{50}$三个指标上分别平均地提升了 14.6%、5.8% 和 10.7%. 进一步地, 消融和可视化实验结果验证了动态卷积核和超图交互模块的有效性.

关键词: 多实例人体解析；动态卷积核；超图交互；人体结构化先验；掩码注意力

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0303

引用格式: 黄荣, 袁家奇, 刘浩, 等. 基于动态卷积和超图交互的多实例人体解析方法 [J]. 控制与决策, 2026, 41(1): 276-288.

A multi-instance human parsing method based on dynamic convolution and hypergraph interaction

HUANG Rong[1,2†]，*YUAN Jia-qi*[1]，*LIU Hao*[1,2]，*JIANG Xue-qin*[1,2]，*ZHOU Shu-bo*[1,2]

(1. College of Information Science and Technology， Donghua University， Shanghai 201620， China；
2. Engineering Research Center of Digitized Textile & Apparel Technology of Ministry of Education， Donghua University，Shanghai 201620，China)

Abstract: Multi-instance human parsing aims to segment multiple human instances and their corresponding parts in natural scene images. Existing methods typically rely on static convolution kernels to segment parts and instances in parallel, resulting in a lack of correlation between part and instance features, and thus limiting adaptability to the diversity of human poses and clothing appearances. To address this issue, this paper proposes a multi-instance human parsing method based on dynamic convolution and hypergraph interaction. Segmentation targets are hierarchically divided into three levels: parts, half-body, and instances, with corresponding learnable dynamic convolution kernels configured for each target. Meanwhile, a multi-scale mask attention mechanism is designed to guide the dynamic convolution kernels in aggregating image features across different levels, thereby adapting to the diversity of human poses and clothing appearances. A hypergraph interaction module is proposed, where part dynamic convolution kernels serve as nodes, and instance and half-body dynamic convolution kernels are treated as hyperedges, to model structural priors of the human body. Feature interaction between parts and instances is achieved through message passing on the hypergraph. Experimental results demonstrate that the proposed method outperforms various baseline methods on the MHP-v2.0, CIHP, and Densepose datasets, achieving average improvements of 14.6%, 5.8%, and 10.7% in AP^{p}_{50}, $\mathrm{AP}^{p}_{\mathrm{vol}}$, and PCP_{50} metrics, respectively. Furthermore, ablation and visualization experiments validate the effectiveness of the dynamic convolution kernels and the hypergraph interaction module.

Keywords: multi-instance human parsing； dynamic convolution kernel； hypergraph interaction； human body structural prior；masked attention

收稿日期: 2025-03-25；**录用日期:** 2025-08-20.
基金项目: 国家自然科学基金项目 (62001099)；中央高校基本科研业务费专项自由探索项目 (2232023D-30).
责任编委: 王琦.
†通信作者. E-mail: rong.huang@dhu.edu.cn.

0 引　言

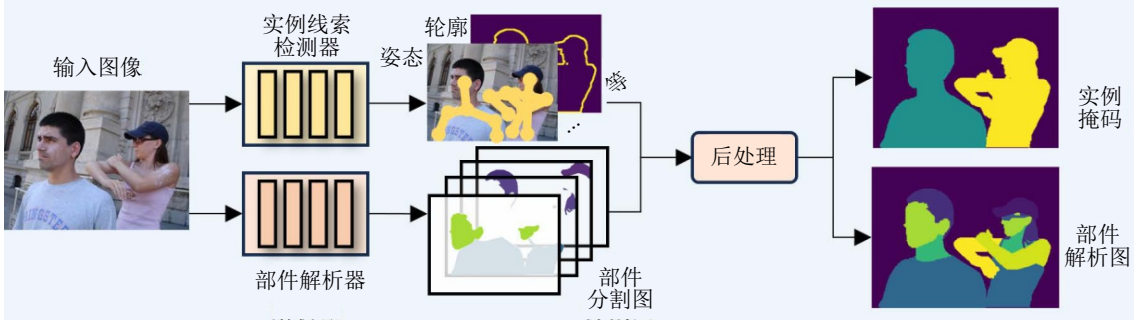
人体解析[1]面向自然场景图像中的人体, 旨在细粒度地分割出人体部件, 如脸部、手臂、躯干等. 近年来, 如远程量体[2]、人体重建[3]、动作预测[4]、虚拟试衣[5]等 AI 应用的发展和普及对于人体解析的性能提出了高要求. 现实的复杂场景中往往存在多个不同的人体实例, 它们的外观各异、姿态多样且相互遮挡. 在区分人体实例的同时细粒度地分割出所有人体部件, 即多实例人体解析, 是近一段时期以来计算机视觉领域的研究热点.

现有的多实例人体解析方法分为自顶向下和自底向上两大类. 自顶向下方法[6-11]将多实例人体解析简化为多个单人人体解析的联合. 如图 1(a) 所示: 这类方法通常先利用 R-CNN 系列[12]或 YOLO 系列[13]等通用的目标检测器粗略地框定出各人体实例的位置, 得到一系列候选区域; 再利用基于 FCN (fully convolution network) 的单人人体解析器[9]依次对候选区域进行分割, 得到各人体实例的解析结果. 然而, 由于多人体实例场景的复杂性, 目标检测器所框定出的矩形候选区域内不可避免地包含非目标实例的部件. 如图 1(a) 所示, 女子的手臂出现在了男子的候选区域内. 这些虚警的人体部件对单人人体解析器形成了干扰, 负面地影响了解析性能. 相对地, 自底向上方法[14-21]将多实例人体解析分为部件分割和部件组合两个阶段, 旨在先分割出所有人体部件, 再将它们组合为人体实例. 如图 1(b) 所示, 在部件分割阶段, 这类方法并行地配置部件解析器和实例线索检测器. 前者输出不区分实例的部件解析结果, 后者提取以人体为中心的 (human-centric) 实例线索, 如位置[14]、廓形[15]和姿态[17]等. 在部件组合阶段, 这类方法根据实例线索, 通过多步聚类[14]、遍历搜索[15]和迭代试配[17]等复杂后处理操作将分割出的人体部件组合为人体实例的掩码, 实现对各人体实例的解析. 然而, 廓形、位置和姿态等通过稀疏的点或线的集合来刻画人体实例, 属于粗粒度的实例线索, 精确区分人体实例的能力有限. 这不可避免地导致部件与实例的错配, 限制了多实例人体解析的性能. 最近, 文献 [20-21] 建立了自底向上方法的新范式, 其通过两条并行的卷积分支分别分割出人体部件和人体实例的细粒度掩码, 再通过简单的逻辑乘法明确部件与实例间的配对关系. 这一新范式排除部件组合阶段, 避免了对实例线索和复杂后处理操作的依赖, 具有优势.

然而, 这些方法[20-21]仍然存在以下两方面问题: 1) 两条并行的卷积分支间缺乏关联, 不利于部件与实例特征的交互. 实质上, 两条卷积分支均以人体为解析对象, 而人体蕴含着如图 2 所示的结构化先验, 即: 人体实例可分解为半身、部件; 部件可组合为半身、人体实例. 缺乏关联的卷积分支忽略了上述的人体结构化先验. 2) 提取部件和实例特征的卷积核呈现出静态性. 这意味着完成训练后, 卷积核的参数保持不变, 难以适应人体姿态和服装外观的多样性[22], 限制了解析模型的泛化能力.

针对上述问题, 本文提出一种基于动态卷积和超图交互的多实例人体解析方法. 将上衣、头发、上半身、下半身、人体实例等作为分割目标, 划分至部件、半身和实例 3 种层次, 并对应地配置可学习的动态卷积核, 如图 2 右侧所示. 设计一种多尺度掩码注

(a)　自顶向下方法



(b)　自底向上方法

图1　现有的多实例人体解析方法框架

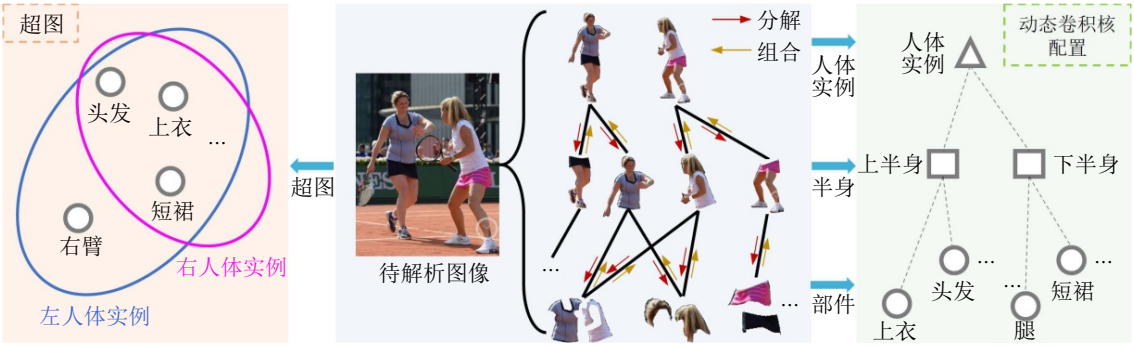


图2 人体结构化先验、动态卷积核配置以及超图

意力机制, 引导动态卷积核从图像特征中聚合分割目标的特征, 从而适应不同图像中人体姿态和服装外观的变化. 如图 2 左侧所示, 按照人体结构化先验, 以部件的动态卷积核为节点 (灰色小圆), 以半身和人体实例的动态卷积核为超边 (彩色大圈), 构建超图. 相比于一般的图 (graph), 超图中的超边可同时连接多个节点 (彩色大圈可同时包含多个灰色小圆). 这与多个人体实例与多种部件间的多对多组合分解关系相契合, 有利于刻画人体结构化先验. 最后, 在超图上通过消息的传递实现部件与实例特征的交互.

本文的主要内容如下.

1) 针对卷积核的静态性问题, 设计多尺度掩码注意力机制来引导各种层次的动态卷积核聚合图像特征. 在推理阶段, 动态卷积核根据图像特征调整参数, 有利于适应人体姿态和服装外观的多样性, 可提升解析模型的泛化能力.

2) 针对部件与实例特征的欠关联性问题, 提出超图交互模块, 以人体结构化先验为引导构建超图. 通过超图上的消息传递, 动态更新部件、半身和人体实例 3 种层次的卷积核, 实现部件与实例特征的交互.

3) 在公开的多实例人体解析数据集 MHP-v2.0[14]、CIHP (crowd instance-level human parsing)[15] 和 Densepose[23] 上完成定量和定性对比实验. 实验结果表明所提出方法的性能优于多种基线方法. 进一步地, 消融和可视化实验结果验证了动态卷积和超图交互的有效性.

1 相关工作

多实例人体解析要求在区分人体实例的同时细粒度地分割出所有人体部件. 相关工作包括单人人体解析、实例分割以及多实例人体解析 (又分为自顶向下和自底向上两大类).

1.1 单人人体解析

现有的单人人体解析方法聚焦于如何融合人体结构化先验建立解析模型. Dong 等[24] 利用低层次过分割算法获取了一组可解析的人体结构片段, 并通过搜索“与或”结构得到最佳的图匹配, 实现了人体解析; 在后续工作中, Dong 等[25] 将结构片段与关节组模板混合, 并利用与或图捕捉人体部件的空间共现和遮挡信息, 联合地实现了人体解析和姿态估计; Liang 等[26] 将人体解析转化为模板回归任务, 并利用人体结构化先验建立模板字典, 获取了人体部件的形状先验, 在此基础上, 通过学习模板的线性组合和主动形变方式生成人体解析结果; ProCnet[27] 按照人体结构化先验渐近地分割人体、半身和部件, 并提出了感知部件的区域卷积结构来交互层次间的信息; Wang 等[28] 以人体结构化先验为指导显式建模了实例分解和部件组合过程中的多源信息融合方式, 有效地实现了跨层次特征交互; 语义神经树 (SNT)[29] 利用树形框架来编码人体结构化先验, 并设计了注意力路由模块和语义聚合模块来融合各部件区域的上下文信息. 虽然这些单人人体解析方法具有启发意义, 但是, 难以将它们直接扩展至多实例人体解析.

1.2 实例分割

实例分割[30] 要求对同一类别的不同实例进行区分, 预测场景中所有实例的掩码. 一些方法[31-35] 先借助通用的目标检测器框定实例的候选区域, 再依次聚焦于各候选区域执行语义分割. Mask R-CNN[31] 在 Faster R-CNN[32] 的基础上添加了预测实例掩码的新分支, 并采用了平均二元交叉熵损失, 实现了分类预测和掩码分割的解耦; Liu 等[33] 提出了自底向上的路径扩增和自适应特征池化, 缩短了特征传递的路径, 畅通了与提议子网络的连接, 提高了特征的利用效率; MS R-CNN[34] 基于实例特征和预测掩码回归交并比评分, 校准掩码质量与评分间的不一致性, 有效提高了实例分割的性能; BM R-CNN[35] 引入了边界保持掩码头, 用于对齐实例的掩码与边界, 提升了实例间的可区分度.

另一些方法[22, 36-38] 先提取图像中像素级部件或片段的特征, 再将它们聚类或分组为实例. Zhang

等[22]提出了一种联合语义分割、实例分割和全景分割的多任务学习框架, 并学习一组语义核和实例核, 统一地实现了目标类别和潜在实例的分割; Wang 等[36]提出了基于位置分割 (SOLO) 方法, 由网格单元直接负责预测区域内的实例掩码和类别, 消除了对目标检测模块的依赖; Wang 等[37]在后续工作中引入了动态卷积机制, 为每个网格单元生成专属的卷积核, 提高了对实例形状和位置的适应性; Mask2former[38]设计了掩码注意力机制和多尺度特征融合, 指导分割模型聚焦目标区域的细粒度信息, 提高了对小目标物体的分割精度. 这些方法以实例为单位进行分割, 未考虑人体的细粒度部件和结构化先验, 难以直接用于解析多实例人体.

1.3　自顶向下多实例人体解析

自顶向下多实例人体解析方法遵循“检测 + 解析”的范式, 先检测人体实例区域再依次解析单个人体实例. BraidNet[6]提出了一种语义特征与细节特征多次交互的交织模块, 并引入区域对比损失增强了人体部件特征的可区分性; Ruan 等[7]提出了实例感知分支和类别感知分支在区分实例的同时解析人体部件. 在此基础上, 通过上下文和高分辨率嵌入逐级细化特征, 有效捕捉复杂边缘和局部纹理. 然而, 这两种方法[6-7]中的检测和解析阶段相分离, 无法实现端到端 (end-to-end) 训练, 在一定程度上限制了解析性能. 相比之下, 文献 [8-10] 将负责目标检测与单人人体解析的网络串级在一起联合优化, 提高了跨任务一致性. Parsing R-CNN[8]在统一实例检测和解析的基础上, 利用感兴趣区域对齐模块提取了高分辨率的特征, 并设计了专用的部分分割模块来生成实例的部件级掩码; RP R-CNN[9]在 Parsing R-CNN[8]的基础上, 提出了全局语义增强特征金字塔和解析重评分模块, 实现了多尺度特征的全局整合和低质量解析结果的自适应过滤; AIParsing[10]提出了一种无锚的多实例人体解析方法, 排除了锚检测头对于超参数的敏感性, 并通过感知局部区域的边界线索区分相邻或重叠的人体部件, 达到了较好的解析精度.

1.4　自底向上多实例人体解析

自底向上多实例人体解析方法遵循“解析 + 组合”的范式, 先分割出所有人体部件, 再将它们按照实例线索组合起来, 得到人体实例掩码. 嵌套对抗网络 (NAN)[14]提出了深度嵌套对抗学习框架, 提取具有区域一致性的部件特征, 再以人体结构先验为指导聚合部件特征实现了实例级的人体解析; 部件组合网络 (PGN)[15]以人体实例的边缘为线索, 扫描人体部件解析图和实例感知边缘图, 再使用广度优先搜索实现了部件组合; 单阶段多人解析 (SMP)[16]利用实例质心位置上的点特征生成解析掩码, 并预测一系列从实例质心到部件质心的偏移量实现了实例与部件的匹配; Zhou 等[17]基于特征金字塔建立密集至稀疏的投影场, 显式建模人体语义与关键点间的关联, 并以可微分的方式求解关节关联的二分匹配问题, 完成了部件组合. 上述方法在部件组合阶段依赖位置[14]、廓形[15]和姿态[17]等粗粒度的实例线索, 不可避免地导致部件与实例的错配. 最近, 个体分割人体解析 (HPSP)[20]以及 Uniparser[21]建立了两条并行的卷积分支分别分割出人体部件和人体实例的细粒度掩码, 再通过简单的逻辑乘法明确部件与实例间的配对关系, 排除了对目标检测器和粗粒度实例线索的依赖. 本文进一步地以人体结构化先验为指导来构建动态卷积核和超图交互模块, 促使解析模型适应复杂场景中人体姿态和服装外观的多样性, 能够提高解析性能.

2　本文方法

本文提出一种基于动态卷积和超图交互的多实例人体解析方法. 首先, 描述所提出方法的总体框架, 然后分别介绍多尺度掩码注意力机制、动态卷积核、超图交互模块和损失函数.

2.1　总体框架

本文的总体框架如图 3 所示, 分割目标类别包括上衣、头发、上半身、下半身、人体实例等. 为分割目标对应地配置一组随机初始化的可学习动态卷积核 $\boldsymbol{E}_0$, 并将它们划分至部件、半身和实例 3 种层次, 如图 3 中的圆形、方形、三角形小图标所示. 在特征聚合阶段, 多尺度掩码注意力机制引导各动态卷积核从相应的分割目标中聚合图像特征; 在特征交互阶段, 按照人体结构化先验, 以部件的动态卷积核为节点, 半身和实例的动态卷积核为超边, 构建超图, 并根据超图上消息的传递来实现部件与实例特征的交互. 将上述两阶段级联, 根据反向传播的梯度更新动态卷积核. 上述流程以迭代方式执行, 迭代总次数记为 T. 迭代完成后, 得到一组既聚合了分割目标的图像特征又捕获了人体结构化先验的动态卷积核 $\boldsymbol{E}_T$, 用于多实例人体解析. 在推理阶段, 动态卷积核根据图像特征和超图消息传递调整参数, 有利于适应人体姿态和服装外观的多样性.

2.2　多尺度掩码注意力机制

在特征聚合阶段, 配置一组与分割目标对应的

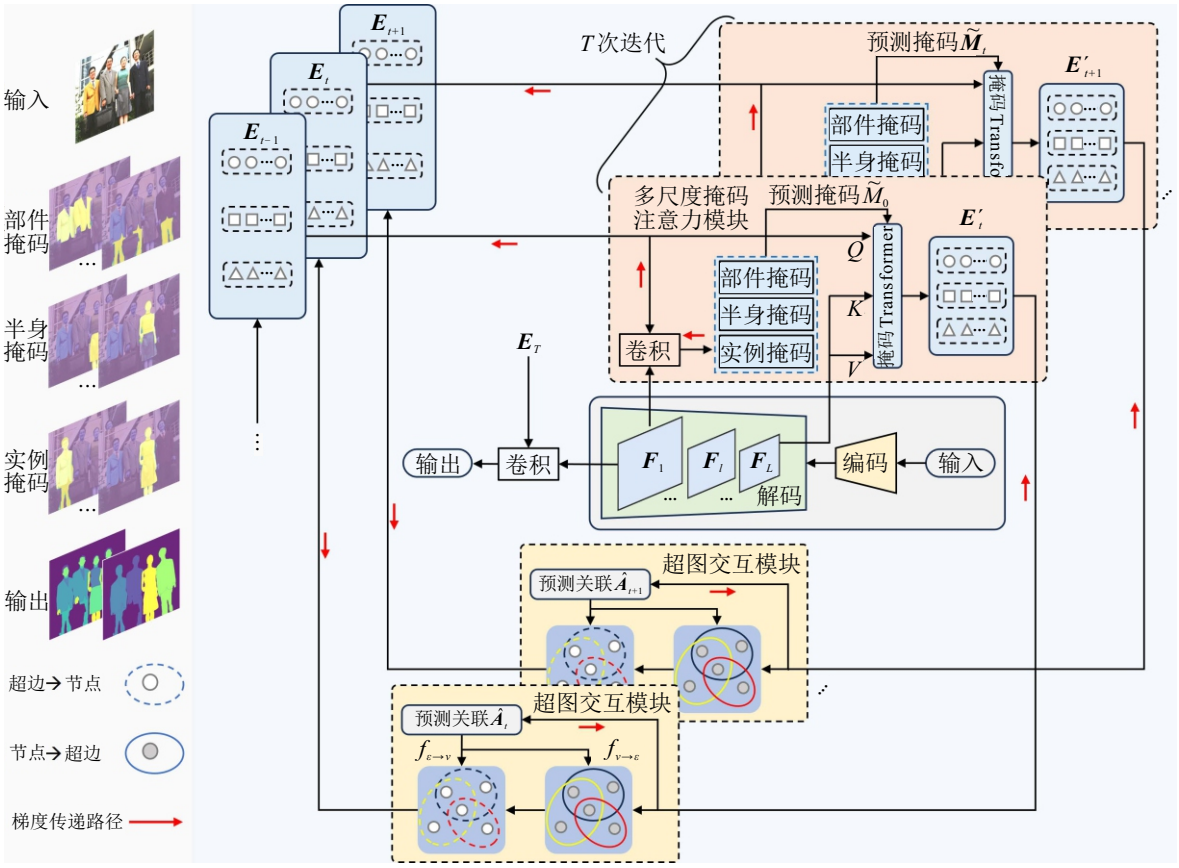


图3 基于动态卷积和超图交互的多实例人体解析方法框架

动态卷积核, 定向地从同类的分割目标中聚合图像特征. 设计多尺度掩码注意力机制, 通过预测掩码来排除非分割目标的干扰, 增强动态卷积核对分割目标的识别能力. 在T次迭代过程中, 该机制引导动态卷积核依次聚合不同尺度的图像特征, 增强动态卷积核对分割目标尺度的适应能力.

将第t次迭代更新的动态卷积核记为$\boldsymbol{E}_t \in \mathbb{R}^{N\times C}$. 其中: $t=0,1,\dots,T$; N和C分别为动态卷积核的个数和维数. 如图 3 所示: 设解码器提取L个不同尺度的特征图, 记第l层的特征图为$\boldsymbol{F}_l$, 其大小为$H_l \times W_l$ $(l=0,1,\dots,L)$; 其通道数均为C. 上述各层特征图中, l越小, 特征图的尺度越大, 所包含的位置信息越丰富; 反之, l越大, 特征图的尺度越小, 所包含的语义信息越丰富.

不失一般性, 以第t次迭代为例展开叙述. 首先, 利用$\boldsymbol{E}_{t-1}$(由第$t-1$次迭代更新后获得) 对最高分辨率的特征图$\boldsymbol{F}_L$执行1×1的卷积操作、二值化、二分匹配 (详见后文第 2.4 节) 和降采样 (从$H_L \times W_L$到$H_l \times W_l$), 得到分割目标的预测掩码$\hat{\boldsymbol{M}}_{t-1} \in [0,1]^{N\times H_l W_l}$. 设置阈值为 0.5, $\hat{M}_{t-1}(x,y)>0.5$表示$\boldsymbol{F}_L$中第y个位置的特征被预测为第x个分割目标类别. 然后, 如图 3 所示, 将$\boldsymbol{E}_{t-1}$作为查询 (query)、特征图$\boldsymbol{F}_l$分别作为键 (key) 和值 (value)、$\hat{M}_{t-1}(x,y)$作为掩码送入 Transformer[39] 进行掩码交叉注意力操作, 得到特征聚合阶段的动态卷积核$\boldsymbol{E}'_t$. 上述计算过程表示为

$$\boldsymbol{E}'_t = \mathrm{softmax}(\tilde{\boldsymbol{M}}_{t-1} + \boldsymbol{Q}_t \boldsymbol{K}_t^{\mathrm{T}})\boldsymbol{V}_t + \boldsymbol{E}_{t-1}. \tag{1}$$

其中: $\boldsymbol{Q}_t = f_Q(\boldsymbol{E}_{t-1}) \in \mathbb{R}^{N\times C}$, $\boldsymbol{K}_t = f_K(\boldsymbol{F}_l) \in \mathbb{R}^{H_l W_l \times C}$, $\boldsymbol{V}_t = f_V(\boldsymbol{F}_l) \in \mathbb{R}^{H_l W_l \times C}$, $f_Q(\cdot)$、$f_K(\cdot)$、$f_V(\cdot)$为可学习的线性变换, 即全连接层; 交叉注意力掩码$\tilde{\boldsymbol{M}}_{t-1}$由分割目标的预测掩码$\hat{\boldsymbol{M}}_{t-1}$经如下变换获得:

$$\tilde{M}_{t-1}(x,y) = \begin{cases} 0, & \hat{M}_{t-1}(x,y) > 0.5; \\ -\infty, & \text{otherwise}. \end{cases} \tag{2}$$

综上, 多尺度掩码注意力机制根据迭代次数索引不同尺度的特征图$\boldsymbol{F}_l$, 并按照预测掩码定向地从同类的分割目标中聚合图像特征, 排除非分割目标的干扰, 增强动态卷积核对分割目标的识别能力.

2.3 超图交互模块

在特征交互阶段, 借鉴超图神经网络[40] 构建超图交互模块, 在各种层次的动态卷积核间传递消息, 以实现部件与实例特征的交互.

将超图定义为节点和超边的集合$G=(\mathcal{V},\mathcal{E})$. 其中: $\mathcal{V}$为节点集, $\mathcal{E}$为超边集. 相比于一般的图, 超图中的一条超边可同时连接任意多个节点. 这一特

性与人体结构化先验中实例分解和部件组合的多对多关系相契合. 构建基于超图的特征交互模块有利于融合人体结构化先验.

具体地, N个动态卷积核$\boldsymbol{E}'_t \in \mathbb{R}^{N\times C}$与分割目标的$N$个类别一一对应. 按照人体结构化先验将$\boldsymbol{E}'_t$划分至部件、半身和实例 3 种层次, 所得数量分别记为N_P、N_H和N_I, 并有$N_P+N_H+N_I=N$成立. 将$\boldsymbol{E}'_t$中部件的动态卷积核作为超图的节点集, 记为$\boldsymbol{E}'_{t,\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|\times C}$, 其中$|\mathcal{V}|=N_P$; 将$\boldsymbol{E}'_t$中半身和实例的动态卷积核作为超图的超边集, 记为$\boldsymbol{E}'_{t,\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|\times C}$, 这里$|\mathcal{E}|=N_H+N_I$, 则有$\boldsymbol{E}'_t=\boldsymbol{E}'_{t,\mathcal{V}}\bigcup\boldsymbol{E}'_{t,\mathcal{E}}$.

在超图中, 节点与超边间的连接关系由关联矩阵$\boldsymbol{A}\in\{0,1\}^{|\mathcal{E}|\times|\mathcal{V}|}$来描述. 若第$e$条超边与第$v$个节点相连接, 则$A(e,v)=1$; 否则, $A(e,v)=0$. 真实的关联矩阵可从像素级的掩码标签中获取: 若第e个实例的掩码与第v个部件的掩码交集不为空, 则相应的超边与节点间相连接. 预测关联矩阵的步骤如下: 首先, 将$\boldsymbol{E}'_{t,\mathcal{V}}$和$\boldsymbol{E}'_{t,\mathcal{E}}$经由全连接层变换得到$\boldsymbol{Z}_{t,\mathcal{V}}$和$\boldsymbol{Z}_{t,\mathcal{E}}$; 然后, 经相似度度量和归一化, 预测出第$t$次迭代的关联矩阵$\hat{\boldsymbol{A}}_t\in[0,1]^{|\mathcal{E}|\times|\mathcal{V}|}$, 表示为

$$\hat{\boldsymbol{A}}_t=\varepsilon(\boldsymbol{Z}_{t,\mathcal{E}}\cdot(\boldsymbol{Z}_{t,\mathcal{V}})^{\mathrm{T}}), \tag{3}$$

其中$\varepsilon(\cdot)$为 Sigmoid 函数. 式 (1) 的含义为将部件动态卷积核与其真实关联的实例/半身动态卷积核可学习地投影至同一空间中的相近区域. 在此过程中, 真实的关联矩阵$\boldsymbol{A}$作为监督信号 (详见后文第 2.4 节).

基于预测的关联矩阵$\hat{\boldsymbol{A}}_t$, 依次执行节点到超边, 超边到节点的消息传递, 更新节点和超边所对应的动态卷积核. 节点到超边的消息传递表示为

$$\boldsymbol{E}''_{t,\mathcal{E}}=f_{\mathcal{V}\to\mathcal{E}}(\boldsymbol{E}'_{t,\mathcal{E}},\boldsymbol{E}'_{t,\mathcal{V}};\hat{\boldsymbol{A}}_t). \tag{4}$$

其中: $\boldsymbol{E}''_{t,\mathcal{E}}$为超边的更新结果; 具体地, $f_{\mathcal{V}\to\mathcal{E}}$由 Transformer 中的掩码交叉注意力操作实现, 表示为

$$f_{\mathcal{V}\to\mathcal{E}}=\mathrm{softmax}(\tilde{\boldsymbol{A}}_t+\boldsymbol{Q}_{t,\mathcal{E}}\boldsymbol{K}^{\mathrm{T}}_{t,\mathcal{V}})\boldsymbol{V}_{t,\mathcal{V}}+\boldsymbol{E}'_{t,\mathcal{E}}. \tag{5}$$

这里: $\boldsymbol{Q}_{t,\mathcal{E}}\in\mathbb{R}^{|\mathcal{E}|\times C}$为$\boldsymbol{E}'_{t,\mathcal{E}}$经由全连接层变换获得; $\boldsymbol{K}_{t,\mathcal{V}},\boldsymbol{V}_{t,\mathcal{V}}\in\mathbb{R}^{|\mathcal{V}|\times C}$为$\boldsymbol{E}'_{t,\mathcal{V}}$经由全连接层变换获得; $\tilde{\boldsymbol{A}}_t\in\{0,-\infty\}^{|\mathcal{E}|\times|\mathcal{V}|}$为$\hat{\boldsymbol{A}}_t$经由式 (2) 变换后获得. 式 (5) 的含义如下: 以存在连接的节点与超边间的注意力为权重消息, 通过对节点的加权来传递消息, 更新超边. 类推地, 超边到节点的消息传递表示为

$$\boldsymbol{E}''_{t,\mathcal{V}}=f_{\mathcal{E}\to\mathcal{V}}(\boldsymbol{E}'_{t,\mathcal{V}},\boldsymbol{E}'_{t,\mathcal{E}};\hat{\boldsymbol{A}}^{\mathrm{T}}_t). \tag{6}$$

其中: $\boldsymbol{E}''_{t,\mathcal{V}}$为节点的更新结果; 在形式上, $f_{\mathcal{E}\to\mathcal{V}}$与$f_{\mathcal{V}\to\mathcal{E}}$相类似, 仅需交换符号$\mathcal{V}$和$\mathcal{E}$, 交换$\boldsymbol{E}'_{t,\mathcal{V}}$和$\boldsymbol{E}'_{t,\mathcal{E}}$即可, 相应地, $f_{\mathcal{E}\to\mathcal{V}}$的含义如下: 以存在连接的节点与超边间的注意力为权重消息, 通过对超边的加权来传递消息, 更新节点.

将$\boldsymbol{E}''_{t,\mathcal{E}}$与$\boldsymbol{E}''_{t,\mathcal{V}}$拼接, 得到第$t$次迭代的动态卷积核$\boldsymbol{E}_t=[\boldsymbol{E}''_{t,\mathcal{E}};\boldsymbol{E}''_{t,\mathcal{V}}]$. 至此, 将部件、半身、实例映射为超图中的节点和超边, 通过节点与超边间的消息传递, 实现了 3 种层次动态卷积核的交互.

2.4 损失函数

对初始动态卷积核$\boldsymbol{E}_0$迭代T次后得到$\boldsymbol{E}_T$. 再利用$\boldsymbol{E}_T$对特征图$\boldsymbol{F}_L$执行1×1的卷积操作, 得到预测结果, 记为$\hat{\boldsymbol{Y}}_i=(\hat{\boldsymbol{M}}_i,\hat{\boldsymbol{C}}_i,\hat{\boldsymbol{A}}_i)\ (i=1,2,\ldots,N)$, 其中$\hat{\boldsymbol{M}}_i$、$\hat{\boldsymbol{C}}_i$和$\hat{\boldsymbol{A}}_i$分别为由第$i$个动态卷积核预测的掩码、类别和关联. 相应地, 作为监督信号的标签, 记为$\boldsymbol{Y}_j=(\boldsymbol{M}_j,\boldsymbol{C}_j,\boldsymbol{A}_j)\ (j=1,2,\ldots,N_{\mathrm{gt}})$. 这里: $\boldsymbol{M}_j$、$\boldsymbol{C}_j$和$\boldsymbol{A}_j$分别为第j个分割目标的真实的掩码、预测和关联; N_{gt}为真实的分割目标的数量. 设置$N>N_{\mathrm{gt}}$, 以确保动态卷积核的数量大于分割目标的数量.

采用交叉熵计算预测与标签间的损失, 具体分为掩码预测损失$\mathcal{L}_M$、分类预测损失$\mathcal{L}_C$和关联预测损失$\mathcal{L}_A$. 不失一般性, $\mathcal{L}_A$表示为

$$\mathcal{L}_A=-(\boldsymbol{A}_j\log(\hat{\boldsymbol{A}}_i)+(1-\boldsymbol{A}_j)\log(1-\hat{\boldsymbol{A}}_i)). \tag{7}$$

$\mathcal{L}_M$和$\mathcal{L}_C$的计算形式与式 (7) 一致, 只需将式 (7) 中的$\hat{\boldsymbol{A}}_i$和$\boldsymbol{A}_j$分别替换为$\hat{\boldsymbol{M}}_i$和$\boldsymbol{M}_j$得到掩码预测损失, 分别替换为$\hat{\boldsymbol{C}}_i$和$\boldsymbol{C}_j$得到分类预测损失. 将上述 3 项损失总和得到

$$\mathcal{L}(\hat{\boldsymbol{Y}}_i,\boldsymbol{Y}_j)=\lambda_M\mathcal{L}_M+\lambda_C\mathcal{L}_C+\lambda_A\mathcal{L}_A, \tag{8}$$

其中λ_M、λ_C以及λ_A分别为平衡掩码预测损失、分类预测损失以及关联预测损失重要程度的超参数. 由于不同图像中实例、半身和部件的数量不确定, 本文通过求解二分匹配问题[41]确立预测$\hat{\boldsymbol{Y}}_i$与标签$\boldsymbol{Y}_j$间的对应关系. 具体而言, 利用匈牙利算法[42]搜索一个最优排列σ^*, 表示为

$$\sigma^*=\arg\min_{\sigma}\sum_{j=1}^{N_{\mathrm{gt}}}\mathcal{L}(\hat{\boldsymbol{Y}}_{\sigma(j)},\boldsymbol{Y}_j). \tag{9}$$

然后, 根据σ^*整理预测与标签间的对应关系, 得到用于梯度反向求导的损失$\mathcal{L}(\hat{\boldsymbol{Y}}_{\sigma^*(j)},\boldsymbol{Y}_j)$, 执行模型参数的更新. 对于第 2.2 节中预测的掩码$\hat{\boldsymbol{M}}_{t-1}$, 亦采用上述二分匹配方法来实现监督.

3 实验分析

3.1 实验设置

本文在 3 个标准的多实例人体解析数据集 MHP-v2.0[14]、CIHP[15] 和 Densepose[23] 上评估所提出方法的性能, 并与现有方法[7-11, 14-21, 29, 31, 34-35] 进行比较. 其中: MHP-v2.0 包含 25 403 幅图像, 细粒度标注了 58 个部件类别; CIHP 包含 38 280 幅图像, 细粒度

标注了 19 个部件类别; Densepose 源自于 COCO 数据集中的 27 659 幅类别为“人”的图像, 且细粒度标注了 12 个部件类别. 上述 3 个数据集中, 每幅图像含有至少 2 个、平均 3 个人体实例, 且每幅图像涵盖丰富的自然场景、光照条件和人物姿态, 涉及部件的交叠和遮挡等情况, 贴近实际应用. 本文按照上述各数据集给出的官方数据划分进行模型的训练和测试. 需要指出的是, 上述各数据集实际上未提供官方测试集. 与现有工作[7-11, 14-21] 保持一致, 本文在验证集上进行性能测评. 为避免歧义, 下文统一将验证集称为测试集.

本文采用AP^p(average precision based on part)[43] 和 PCP (percentage of correctly parsed semantic parts)[43] 分别从实例和部件层面测评解析模型的性能. 指标AP^p衡量从属于同一人体实例的不同部件的预测掩码与真实掩码间的交并比 (IoU) 来确定该人体实例是否为真正例 (true positive). 与现有的工作[14-15, 17-18, 20] 一致, 本文采用了AP^p_{50}和$\mathrm{AP}^p_{\mathrm{vol}}$, 前者将 IoU 大于阈值 0.5 的人体实例定义为真正例; 后者在$0.1 \sim 0.9$的范围内 (以 0.1 为递增量) 取阈值, 求取多个AP^p的平均值. 指标PCP统计正确解析的部件数量与部件总数量的比率. 本文采用PCP_{50}, 即将 IoU 大于阈值 0.5 的部件定义为真正例. 上述指标的数值越大, 解析性能越好.

所提出方法以 ResNet-50[44] 作为编解码器的骨干. 采用随机水平翻转和尺度抖动进行数据增强. 训练时, 批大小设置为 8. 采用随机梯度下降 (SGD) 优化器对模型进行训练, 训练总回合数为 72. SGD 优化器的权重衰减系数设置为 0.000 1, 动量因子设置为 0.9. 初始学习率设置为 0.002. 模型训练过程中, 采用分段衰减策略[23, 43] 在第 50 个和第 65 个训练回合以 1/10 为因子对学习率进行递次衰减. 动态卷积核的数量设置如下: 实例和半身的动态卷积核数量分别设置为 10 和 20; 部件的动态卷积核数量与数据集中的部件类别数量一致, 即对于 MHP-v2.0 设置为 58, 对于 CIHP 设置为 19, 对于 Densepose 设置为 12. 解码器中图像特征的金字塔层数$L = 3$. 训练和测试各进行 3 次迭代更新动态卷积核, 即$T = 3$. 动态卷积核参数的初始化采用服从正态分布的随机初始化策略. 超参数λ_M、λ_C和λ_A分别设置为 5、2、1 [38, 45]. 所提出方法的源代码下载地址为 https://github.com/Yjhan2/Ins2part.

3.2 定量结果分析与比较

表 1 为所提出方法与 15 种现有的多实例人体解析方法[7-11, 14-21, 29, 31] 在 MHP-v2.0 测试集上的定量结果. 其中: 前 7 种属于自顶向下方法, 后 8 种属于自底向上方法. 由表 1 可知, 所提出方法在 3 个指标上均优于现有方法, 分别平均高出了 16.2%、6.7% 和 11.8%. 这是由于自顶向下方法[7-11, 29, 31] 依赖于目标检测器, 存在非目标干扰和偏差传播问题; 自底向上方法如 PGN[15]、MHParser[18]、NAN[14] 和 DSPF[17] 依赖粗粒度的实例线索组合部件, 存在错配问题. 相比之下, 现有方法中 HPSP[20]、SMP[16] 和 Uniparser[21] 分别引入细粒度的实例掩码和动态卷积核缓解了错配问题, 取得了与所提出方法相接近的解析性能. 本文在这些方法的基础上提出了多尺度掩码注意力机制和超图交互模块. 这里: 多尺度掩码注意力机制通过聚合不同尺度的图像特征, 增强了动态卷积核对分割目标尺度的适应能力; 动态超图交互模块结合实例与部件间的组合分解关系, 引入了人体结构化先验, 增强了泛化能力; 二者结合使得所提出方法在 3 个指标上分别高出了次优方法 Uniparser[21] 4.5%、2.5% 和 5.1%.

表1 MHP-v2.0 测试集上的定量结果与性能比较

	方法	骨干网络	AP^p_{50}	$\mathrm{AP}^p_{\mathrm{vol}}$	PCP_{50}
自顶向下方法	Mask R-CNN[31]	ResNet-50	14.9	33.9	25.1
	Parsing R-CNN[8]	ResNeXt-101	30.2	41.8	44.2
	M-CE2P[7]	ResNet-101	34.5	42.7	43.7
	SNT[29]	ResNet-101	34.4	42.5	43.5
	RP R-CNN[9]	ResNet-50	45.3	46.8	43.8
	AIParsing[10]	ResNet-50	41.1	45.9	45.3
	CID[11]	HRNet-W48	47.1	48.2	<u>51.5</u>
自底向上方法	PGN[15]	ResNet-101	17.6	35.5	26.9
	MHParse[18]	ResNet-101	18.0	36.1	27.0
	NAN[14]	—	25.1	41.8	32.3
	ReSParser[19]	ResNet-101	34.3	42.7	43.7
	DSPF[17]	ResNet-101	39.0	44.3	42.3
	HPSP[20]	ResNet-101	41.3	45.8	47.7
	HPSP[20]	Swin ViT	44.5	45.7	50.8
	SMP[16]	ResNet-101	47.1	48.2	<u>51.5</u>
	Uniparser[21]	ResNet-50	<u>49.2</u>	<u>48.7</u>	51.1
	本文方法	ResNet-50	**51.4**	**49.9**	**53.7**

注: 粗体和下划线分别表示性能最优和次优

表 2 为所提出方法与 6 种自顶向下方法[8-11, 34-35] 和 5 种自底向上方法[15-16, 19-21] 在 CIHP 测试集上的定量比较结果. 由于 CIHP 数据集的类别数量较少, 各方法在表 2 中的表现优于表 1. 通过统计可知, 所提出方法在 3 个指标上均优于现有方法, 分别平均高出了 7.5%、5.1% 和 3.7%. 现有方法中, AIParsing[10]、SMP[16]、CID[11] 和 Uniparser[21] 与所提出方法在解析

性能上较为接近, 其中 AIParsing[10] 通过无锚的目标检测器和联合边缘强化的单人解析器缓解了非目标干扰和偏差传播问题. 相比之下, 所提出方法无需依赖于目标检测器, 通过实例、半身和部件的细粒度分割实现多实例人体解析. 因此, 所提出方法在 3 个指标上超过了 AIParsing[10] 1.2%、0.8% 和 1.7%.

表2　CIHP 测试集上的定量结果与性能比较

方法		骨干网络	AP^p_{50}	AP^p_{vol}	PCP_{50}
自顶向下方法	Parsing R-CNN[8]	ResNeXt-101	69.1	55.9	66.2
	MS R-CNN[34]	ResNet-50	68.9	56.8	59.9
	BM R-CNN[35]	ResNet-50	64.6	54.3	61.8
	RP R-CNN[9]	ResNet-50	71.6	58.3	62.2
	AIParsing[10]	ResNet-50	73.1	59.2	66.3
	CID[11]	HRNet-W48	73.5	59.4	66.9
自底向上方法	PGN[15]	ResNet-101	39.0	34.0	61.0
	HPSP[20]	ResNet-101	61.2	52.0	61.4
	HPSP[20]	Swin ViT	63.0	52.7	62.0
	ReSParser[19]	ResNet-101	69.2	56.4	65.0
	SMP[16]	ResNet-101	71.7	57.3	64.5
	Uniparser[21]	ResNet-50	73.7	59.4	67.2
	本文方法	ResNet-50	**74.0**	**59.7**	**67.4**

注: 粗体和下划线分别表示性能最优和次优

表 3 为所提出方法与 5 种现有方法[7-9, 15, 17] 在 Densepose 测试集上的定量比较结果. 通过统计可知, 所提出方法取得了最优性能, 在 3 个指标上分别平均高出了现有方法 20.0%、5.5% 和 16.7%. 特别地, 所提出方法比次优方法 DSPF[17] 提升了 11.4%、0.7%、11.1%. 这得益于所提出方法将粗粒度的人体关键点替换为了细粒度的实例掩码, 并通过超图实现了实例与部件特征的交互.

表3　Densepose 测试集上的定量结果与性能比较

方法		骨干网络	AP^p_{50}	AP^p_{vol}	PCP_{50}
自顶向下方法	Parsing R-CNN[8]	ResNeXt-101	43.5	53.1	51.8
	M-CE2P[7]	ResNet-50	43.7	52.9	51.2
	RP R-CNN[9]	ResNet-50	48.5	54.4	51.1
自底向上方法	PGN[15]	ResNet-101	23.4	35.9	32.5
	M-CE2P[7]	ResNet-101	37.6	48.3	43.9
	DSPF[17]	ResNet-50	49.7	54.7	52.8
	本文方法	ResNet-50	**61.1**	**55.4**	**63.9**

注: 粗体和下划线分别表示性能最优和次优

值得一提的是, 与所提出方法采用 ResNet-50 作为骨干网络相比, 不少现有方法[7-8, 15-20, 29] 采用了参数规模更大、特征提取能力更强的 ResNet101[44]、ResNeXt-101[46] 或 Swin ViT[47]. 然而, 在解析性能方面, 所提出方法仍然具有优势, 这进一步验证了经历多次迭代更新后的多层次动态卷积核具备较强的人体特征捕捉能力.

3.3　定性结果分析与比较

图 4 为 Parsing R-CNN[8]、RP R-CNN[9]、HPSP[20]、SMP[16] 以及所提出方法的定性结果, 其中前两个属于自顶向下方法, 依赖于目标检测器 (如图 4 第 1 行和第 2 行结果中的绿色矩形框). 图 4 中: 由掩码的颜色区分人体部件的类别, 由轮廓的颜色区分人体实例.

图 4 第 1 列和第 2 列为“部件遮挡”场景下的定性结果. 观察图 4 第 1 行和第 2 行可知, 自顶向下方法难以应对大面积遮挡, 存在大量漏检, 如长发女性和黑人男性的人体部件. 这是由于目标检测器难以区分存在大面积遮挡的两个人体实例, 且这种检测偏差递次传播至了人体解析阶段, 造成漏检. HPSP[20]、SMP[16] 以及所提出方法无需目标检测器, 避免了非目标干扰和偏差传播问题, 鲁棒地应对了大面积遮挡. 然而, HPSP[20] 中的静态卷积核难以适应人体姿态和服装外观的多样性, 仍然存在如第 3 行第 1 列的实例漏检 (长发女性) 和第 3 行第 2 列的部件漏检 (黑人男性的左手). SMP[16] 依赖于实例质心偏移的粗粒度线索组合人体部件, 存在错配问题. 如第 4 行第 1 列的手部和第 4 行第 2 列的手臂被错误地分配给不存在的人体实例 (由红色轮廓表示). 相比之下, 所提出方法与掩码标签高度贴合, 不但准确地区分开了大面积遮挡的人体实例, 还高质量地分割出了所有人体部件.

图 4 第 3 列和第 4 列为“实例密集”场景下的定性结果. 通过观察图 4 可知, 现有方法对于小尺度分割目标, 在实例/部件的边界处出现了不同程度的错漏. 如 Parsing R-CNN[8]、RP R-CNN[9] 和 HPSP[20] 均未能完整地分割出第 3 列图像中女性的脚部. 第 3 列第 3 行、第 4 行呈坐姿态人体实例的边界处出现了额外的蓝色轮廓痕迹, 表现出部件与实例的错配. 这种现象在第 4 列的定性结果中更为明显, 不少轮廓侵入了人体部件内部区域, 存在错检、漏检和错配等. 相比之下, 所提出方法在“实例密集”场景下的表现优于现有方法. 这是由于所提出方法按照人体结构化先验分层次地配置了多个动态卷积核, 并通过多次的特征聚合和交互, 使得这些动态卷积核能够专属地响应不同的实例、半身和部件, 鲁棒地应对“实例密集”场景.

3.4　消融实验

本节针对迭代次数、动态卷积核的配置数量和特征交互方式开展消融实验, 定量比较不同消融条

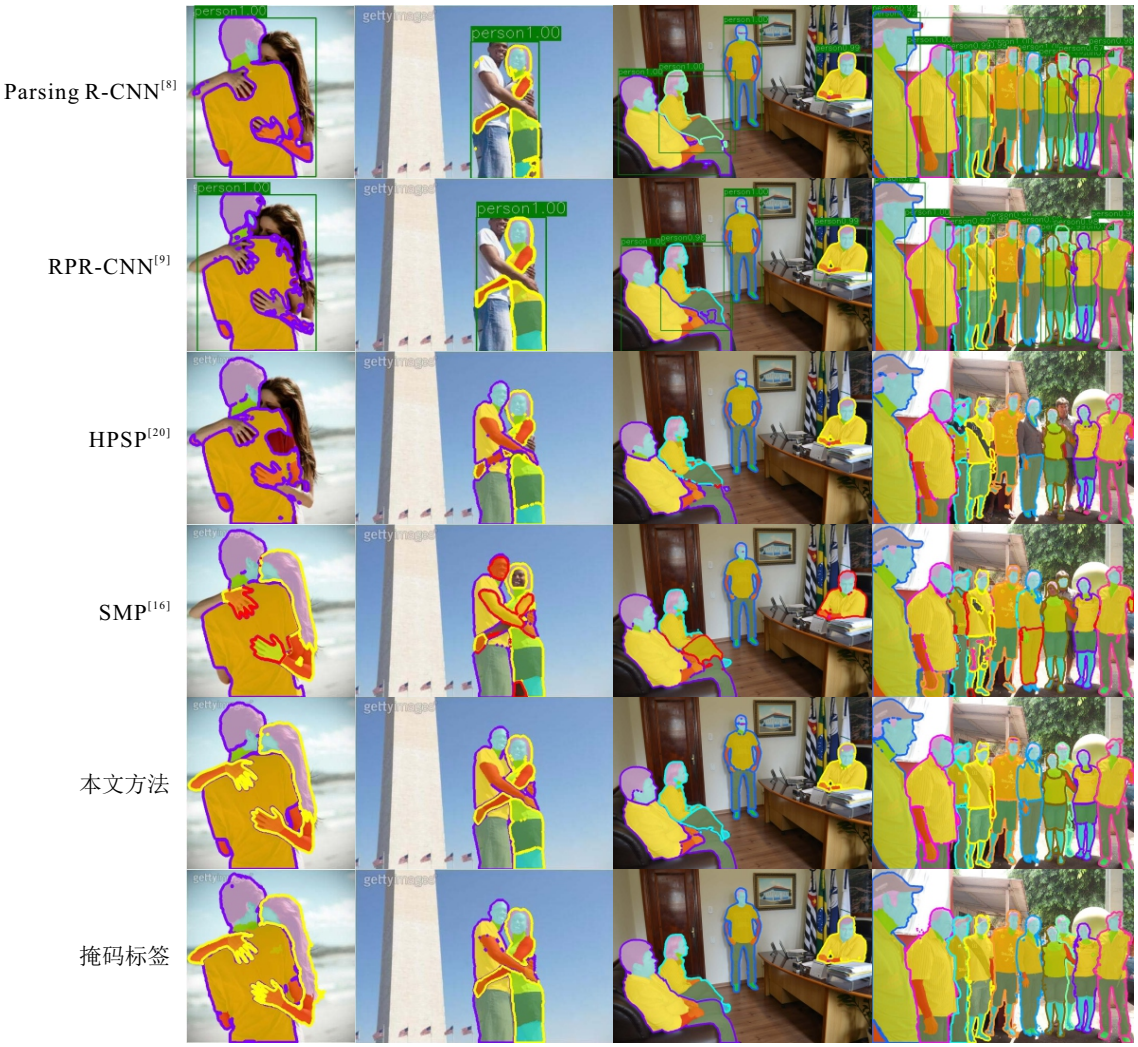


图4 MHP-v2.0 测试集上的定性结果与比较 (由轮廓的颜色区分人体实例)

件对解析性能的影响. 若无特殊说明, 则消融实验均在 MHP-v2 数据集上进行.

本节将动态卷积核迭代次数T以 1 为递进, 依次设置为$0 \sim 5$, 其中$T=0$表示使用初始动态卷积核进行解析. 定量结果如图 5 所示. 当$T=0$时, 在 3 个指标上仅能达到 14.6%、29.2% 和 21.8%, 表明初始动态卷积核仅能粗略地完成分割目标定位. 随着迭代的进行, 解析性能明显提升. 当$T=1$时, 相比于$T=0$时, 在 3 个指标上分别提高了 34.8%、19.8% 和 30.5%. 比较$T=2 \sim T=5$的定量结果可知, 历经 3 次迭代后, 对于动态卷积核的更新效应趋于收敛和饱和. 由图 5 可知, 在 CIHP 数据集上呈现出了一致的性能变化趋势. 因此, 将T默认地设置为 3.

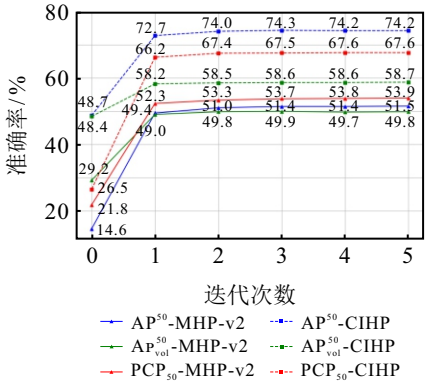


图5 关于动态卷积核迭代次数的消融实验结果

本节将特征交互方式设置为无交互、朴素交互和超图交互 3 种. 其中: 无交互是指直接令$E_t=E'_t$; 朴素交互是指采用 Transformer[39] 中的标准注意力操作, 即所有动态卷积核间不加区分地进行两两交互; 关于超图交互的介绍见第 2.3 节. 相比于超图交互, 朴素交互未考虑实例与部件的关联, 不受人体结构化先验的指导.

表 4 为各组件消融实验的定量结果. 在$N_I=30$的条件下, 超图交互在 3 个指标上达到了最高值, 分别超过无交互 2.4%、4.2% 和 3.7%, 超过朴素交互 3.0%、1.2% 和 1.5%. 这表明超图交互以人体结构化先验为指导, 有利于动态卷积核捕捉人体特征. 在超图交互的基础上, 若将N_I从 30 缩减至 10, 则解析性能略有下降. 比较表 4 的第 4 行与第 7 行实验结果表明, 配置半身动态卷积核使得解析性能提升了

表4　各组件消融实验的定量结果

交互方式	实例核数N_I	半身核数N_H	AP^p_{50}	AP^p_{vol}	PCP_{50}
无交互	30	—	49.6	47.8	51.1
朴素交互	30	—	49.3	49.2	52.2
超图交互	30	—	50.8	49.8	53.0
	10	—	51.0	49.7	52.8
	5	10	48.8	46.5	50.2
	10	20	51.4	49.9	53.7
	15	30	51.5	49.8	53.8
	20	40	51.5	49.9	53.9

1.2%、0.2% 和 1.3%. 这表明在实例与部件间引入半身能够层次化地刻画人体结构, 有助于增强模型对人体结构的理解, 提高对人体姿态和服装外观的适应能力. 当$N_I > 10$时, 解析性能趋于饱和, 增加动态卷积核的数量不再带来增益.

3.5　可视化实验

3.5.1　各次迭代动态卷积核的激活图

为了分析动态卷积核的迭代更新效应, 图 6 可视化了 3 次迭代过程中各层次动态卷积核对于相应分割目标的特征激活程度. 其中: 图 6(a) ～ 图 6(d) 分别为实例 (左起第 2 人)、半身 (左起第 2 人的上下半身)、部件 (夹克) 的激活图; 紫色表示激活程度为 0, 从绿色到黄色的逐渐过渡表示激活程度逐渐增大.

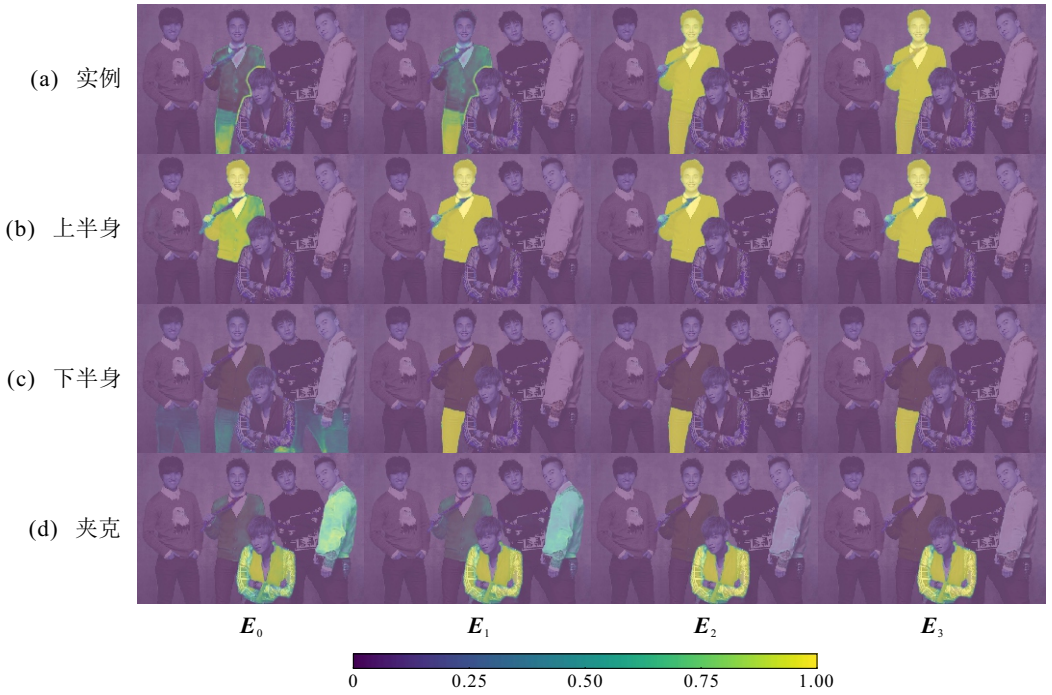


图6　不同迭代次数的动态卷积核激活图

通过观察可知, 图 6 第 1 列的初始动态卷积核仅能完成粗略的分割目标定位, 各激活图中黄色的分布稀碎. 此外, 还存在一定程度的虚警. 如图 6(d) 第 1 列中初始动态卷积核在夹克、上衣与毛衣间产生了混淆. 比较图 6 第 1 列与第 3 列的差异性可知: 随着迭代的进行, 激活图逐渐细化, 分割目标的激活得到增强, 非分割目标的干扰被排除. 如图 6(d) 第 3 列中迭代更新后的动态卷积核较好地激活了“夹克”区域, 抑制了“上衣”和“毛衣”区域. 这表明多尺度掩码注意力机制迭代地从分割目标中聚合了图像特征, 有利于增强动态卷积核对分割目标的识别能力. 进一步地, 比较图 6 第 3 列与第 4 列的相似性可知: 历经 3 次迭代后, 更新效应趋于收敛和饱和, 使得动态卷积核具备了适应人体姿态和服装外观的能力. 上述分析与第 3.4 节的消融实验结果相一致.

3.5.2　超图交互中实例与部件的关联

超图交互模块以人体结构化先验为指导, 通过如式 (5) 所示的掩码交叉注意力操作在实例与部件的动态卷积核间传递消息, 以实现交互. 本节可视化了超图交互中实例与部件的关联, 即人体结构化先验中实例分解和部件组合.

首先, 收集 MHP-v2.0 测试集中的人体实例. 然后, 以实例为单位, 根据真实关联展开得到 58 维的行向量. 行向量中元素 1 表示实例拥有对应的部件; 反之, 则设置为 0. 接着, 对所有实例的行向量进行聚类, 得到的每个聚类簇内的实例拥有相似的部件. 最后, 从 3 个最大的聚类簇中分别选取离簇心距离最近的 20 个实例, 共 60 个. 图 7 为关联注意力分布与真实分布的可视化对比. 如图 7(a) 所示: 将 3 个聚类簇编号为 Ⅰ、Ⅱ、Ⅲ; 将 20 个行向量排列构成一个矩阵; 矩阵中元素 1 用黄色表示, 0 用灰色表示. 可

见, 同一矩阵中的行向量间拥有较为相似的黄色分布. 图 7(b) 和图 7(c) 分别为基于超图交互和朴素交互的关联注意力分布 (统计自选定的 60 个实例), 其中朴素交互如第 3.4 节所述是指不加区分地对所有动态卷积核进行两两交互. 图 7(b) 和图 7(c) 中, 从绿色到黄色的逐渐过渡表示关联程度逐渐增大.

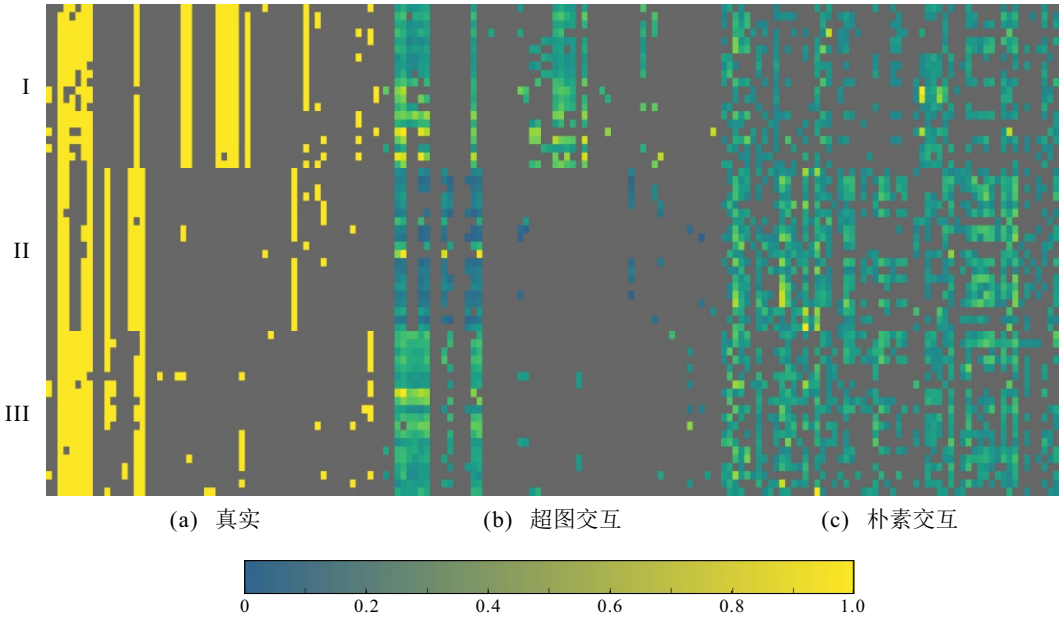


图7 关联注意力分布与真实分布的可视化对比

通过比较可知, 超图交互模块的关联注意力分布与真实分布相接近. 相对地, 朴素交互模块由于缺乏人体结构化先验的指导, 存在不少冗余的特征交互, 导致其关联注意力分布与真实分布相差较大. 这表明超图交互模块能够融合人体结构化先验, 有效地捕捉实例与部件的关联, 有利于不同层次动态卷积核的交互, 提高动态卷积核适应人体姿态和服装外观变化的能力.

3.6 跨数据集交叉测试及讨论

进一步地, 本节测试所提出方法的跨数据集迁移能力, 定量实验结果如表 5 所示, 其中 $A \to B$ 表示模型在 A 数据集上训练后在 B 数据集上测试. 对比表 1、表 2 与表 5 的结果可知, 跨数据集交叉测试准确率在 3 个指标上分别平均下降了 18.6%、16.5% 和 9.7%. 进一步地, 图 8 为跨数据集交叉测试的定性结果. 通过观察可知, 所提出方法对于面部、头发等常见类别维持了较好的解析性能, 对于夹克、上衣等类别的解析性能较差.

表5 跨数据集交叉测试的定量结果

交叉测试方式	AP^{p}_{50}	AP^{p}_{vol}	PCP_{50}
CIHP→MHP	48.5	44.5	43.7
MHP→CIHP	50.6	46.4	66.9

跨数据集交叉测试性能下降的原因在于: 不同数据集的标注标准存在差异, 对于相同部件类别的定义存在分歧, 导致语义漂移问题. 未来工作将构建跨数据集的超图, 通过部件特征交互融合不同数据集的多元特征来实现同类部件特征的跨数据集对齐.

图8 跨数据集交叉测试的定性结果

4 结 论

本文提出了一种基于动态卷积和超图交互的多实例人体解析方法. 将分割目标划分至部件、半身和实例 3 种层次, 并对应地配置可学习的动态卷积核. 针对卷积核的静态性问题, 设计了多尺度注意力机制, 利用掩码排除非分割目标干扰, 引导动态卷积核定向地聚合分割目标的图像特征. 在推理阶段, 动态卷积核能够根据图像特征调整参数, 从而适应人体姿态和服装外观的多样性, 提升了解析模型的泛化能力. 针对部件与实例特征的欠关联性问题, 设计了超图交互模块. 该模块将部件的动态卷积核作为节

点, 半身和实例的动态卷积核作为超边, 并以人体结构化先验为引导构建了超图. 通过超图上的消息传递, 实现了部件与实例特征的交互. 实验结果表明所提出方法在 3 个主流的公开数据集上均取得了最佳的解析性能. 定量结果表明, 所提出方法在 AP^{p}_{50}、AP^{p}_{vol} 和 PCP_{50} 三个指标上平均高于现有方法 14.6%、5.8% 和 10.7%. 定性结果验证了所提出方法能够在区分人体实例的同时分割人体部件. 消融和可视化实验结果进一步验证了动态卷积和超图交互的有效性.

未来工作将研究基于超图的跨数据集特征交互方法, 指导解析模型学习统一的部件语义表示, 增强跨数据集迁移能力.

参考文献 (References)

[1] 甘霖, 刘骊, 刘利军, 等. 结合边缘轮廓和姿态特征的人体精确解析模型[J]. 计算机辅助设计与图形学学报, 2021, 33(9): 1428-1439.
(Gan L, Liu L, Liu L J, et al. Accurate human parsing model by edge contour and pose feature[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(9): 1428-1439.)

[2] Zhang J, Hu Y Y, Zhang M Y, et al. Non-contact body measurement for garment customization based on 2D images: A systematic review[J]. Textile Research Journal, DOI: 10.1177/00405175251318629.

[3] Wang J Y, Yoon J S, Wang T F, et al. Complete 3D human reconstruction from a single incomplete image[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 8748-8758.

[4] Ding G D, Sener F, Yao A. Temporal action segmentation: An analysis of modern techniques[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(2): 1011-1030.

[5] Chen C Y, Chen Y C, Shuai H H, et al. Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network[C]. IEEE/CVF International Conference on Computer Vision. Paris, 2023: 7479-7488.

[6] Liu X C, Zhang M, Liu W, et al. BraidNet: Braiding semantics and details for accurate human parsing[C]. Proceedings of the 27th ACM International Conference on Multimedia. Nice, 2019: 338-346.

[7] Ruan T, Liu T, Huang Z L, et al. Devil in the details: Towards accurate single and multiple human parsing[C]. AAAI Conference on Artificial Intelligence. Honolulu, 2019: 4814-4821.

[8] Yang L, Song Q, Wang Z H, et al. Parsing R-CNN for instance-level human analysis[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 364-373.

[9] Yang L, Song Q, Wang Z H, et al. Renovating parsing R-CNN for accurate multiple human parsing[C]. Computer Vision — ECCV 2020. Glasgow, 2020: 421-437.

[10] Zhang S Y, Cao X C, Qi G J, et al. AIParsing: Anchor-free instance-level human parsing[J]. IEEE Transactions on Image Processing, 2022, 31: 5599-5612.

[11] Wang D K, Zhang S L. Contextual instance decoupling for instance-level human analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9520-9533.

[12] 赵亮, 高升伦, 陈俊英, 等. 基于特征共享双头 Cascade R-CNN 的混凝土细观损伤特征检测[J]. 控制与决策, 2022, 37(7): 1745-1751.
(Zhao L, Gao S L, Chen J Y, et al. Feature detection of concrete mesoscopic damage based on feature sharing double-head Cascade R-CNN[J]. Control and Decision, 2022, 37(7): 1745-1751.)

[13] Terven J, Córdova-Esparza D M, Romero-González J A. A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS[J]. Machine Learning and Knowledge Extraction, 2023, 5(4): 1680-1716.

[14] Zhao J, Li J S, Cheng Y, et al. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing[C]. Proceedings of the 26th ACM International Conference on Multimedia. Seoul, 2018: 792-800.

[15] Gong K, Liang X D, Li Y H, et al. Instance-level human parsing via part grouping network[C]. European Conference on Computer Vision. Munich, 2018: 805-822.

[16] Chu J M, Jin L, Fan X J, et al. Single-stage multi-human parsing via point sets and center-based offsets[C]. Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, 2023: 1863-1873.

[17] Zhou T F, Yang Y, Wang W G. Differentiable multi-granularity human parsing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8296-8310.

[18] Li J S, Zhao J, Lang C Y, et al. Multi-human parsing with a graph-based generative adversarial model[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(1): 1-21.

[19] Dai Y, Chen X J, Wang X H, et al. ReSParser: Fully convolutional multiple human parsing with representative sets[J]. IEEE Transactions on Multimedia, 2023, 26: 1384-1394.

[20] Li Z, Cao L L, Wang H B, et al. End-to-end instance-level human parsing by segmenting persons[J]. IEEE Transactions on Multimedia, 2023, 26: 41-50.

[21] Chu J M, Jin L, Teng Y L, et al. UniParser: Multi-human parsing with unified correlation representation learning[J]. IEEE Transactions on Image Processing, 2024, 33: 5159-5171.

[22] Zhang W W, Pang J M, Chen K, et al. K-Net: Towards unified image segmentation[C]. Advances in Neural Information Processing Systems 34. Montreal, 2021: 10326-10338.

[23] Güler R A, Neverova N, Kokkinos I. DensePose: Dense

human pose estimation in the wild[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7297-7306.

[24] Dong J, Chen Q, Xia W, et al. A deformable mixture parsing model with parselets[C]. IEEE International Conference on Computer Vision. Sydney, 2013: 3408-3415.

[25] Dong J, Chen Q, Shen X H, et al. Towards unified human parsing and pose estimation[C]. IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 843-850.

[26] Liang X D, Liu S, Shen X H, et al. Deep human parsing with active template regression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(12): 2402-2414.

[27] Zhu B K, Chen Y Y, Tang M, et al. Progressive cognitive human parsing[C]. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, 2018: 7607-7614.

[28] Wang W G, Zhang Z J, Qi S Y, et al. Learning compositional neural information fusion for human parsing[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 5703-5713.

[29] Ji R Y, Du D W, Zhang L B, et al. Learning semantic neural tree for human parsing[C]. Proceedings of the 16th European Conference on Computer Vision. Glasgow, 2020: 205-221.

[30] 张旭东, 王玉婷, 范之国, 等. 基于双金字塔特征融合网络的 RGB-D 多类实例分割[J]. 控制与决策, 2020, 35(7): 1561-1568.
(Zhang X D, Wang Y T, Fan Z G, et al. RGB-D multi-class instance segmentation based on double pyramid feature fusion model[J]. Control and Decision, 2020, 35(7): 1561-1568.)

[31] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. IEEE International Conference on Computer Vision. Venice, 2017: 2980-2988.

[32] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[33] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8759-8768.

[34] Huang Z J, Huang L C, Gong Y C, et al. Mask scoring R-CNN[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 6402-6411.

[35] Cheng T H, Wang X G, Huang L C, et al. Boundary-preserving mask R-CNN[C]. Proceedings of the 16th European Conference on Computer Vision. Glasgow, 2020: 660-676.

[36] Wang X L, Kong T, Shen C H, et al. SOLO: Segmenting objects by locations[C]. Proceedings of the 16th European Conference on Computer Vision. Glasgow, 2020: 649-665.

[37] Wang X L, Zhang R F, Kong T, et al. SOLOv2: Dynamic and fast instance segmentation[C]. Advances in Neural Information Processing Systems 33. Vancouver, 2020: 17721-17732.

[38] Cheng B W, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 1280-1289.

[39] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems 30. Long Beach, 2017: 6000-6010.

[40] Li M R, Zhang Y, Li X Y, et al. Hypergraph transformer neural networks[J]. ACM Transactions on Knowledge Discovery from Data, 2023, 17(5): 1-22.

[41] Karp R M, Vazirani U V, Vazirani V V. An optimal algorithm for on-line bipartite matching[C]. Proceedings of the 22nd Annual ACM Symposium on Theory of Computing — STOC '90. Baltimore, 1990: 352-358.

[42] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics: NRL, 2005, 52(1): 7-21.

[43] Li J S, Zhao J, Wei Y C, et al. Multiple-human parsing in the wild[J/OL]. 2017, arXiv: 1705.07206.

[44] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[45] Zhang H, Li F, Xu H Z, et al. MP-former: Mask-piloted transformer for image segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 18074-18083.

[46] Xie S N, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 5987-5995.

[47] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 9992-10002.

作者简介

黄荣 (1985−), 男, 副教授, 博士, 主要研究方向为机器视觉、深度学习与应用、多媒体信息安全, E-mail: rong.huang@dhu.edu.cn;

袁家奇 (1999−), 男, 硕士生, 主要研究方向为自然场景人体解析, E-mail: 2222090@mail.dhu.edu.cn;

刘浩 (1977−), 男, 副教授, 博士, 主要研究方向为深度学习、机器视觉, E-mail: liuhao@dhu.edu.cn;

蒋学芹 (1981−), 男, 教授, 博士, 博士生导师, 主要研究方向为图机器学习、工业视觉、量子编码, E-mail: xqjiang@dhu.edu.cn;

周树波 (1988−), 男, 助理研究员, 博士, 主要研究方向为深度学习、工业视觉、图像分析, E-mail: zhoushubo@dhu.edu.cn.