小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2024-0605
2026 年 2 月 第 2 期
Vol.47 No.2 2026

深度融合实例-类别特征的多人解析算法

叶岚清[1],刘 骊[1,2],付晓东[1,2],刘利军[1,2],彭 玮[1,2]

[1](昆明理工大学 信息工程与自动化学院,昆明 650500)
[2](昆明理工大学 信息工程与自动化学院 云南省计算机技术应用重点实验室,昆明 650500)
E-mail:ieall@kust.edu.cn

摘 要: 针对多人解析中存在阶段间强耦合性、多尺度表征不足、解析精准性受限的问题,提出一种深度融合实例-类别特征的多人解析算法.首先,在特征空间中定义包含人体部位、关节点和服装配饰的编码集,通过引入可变形注意力进行人体多尺度特征学习;然后,将人体多尺度特征作为共享语义信息,统一表示实例特征和类别特征;最后结合交叉注意力深度融合实例与类别特征,以生成精准的多人解析结果.在 MHP V2.0 数据集上的实验结果表明,所提算法在评估指标 AP^{P}_{vol} 和 PCP_{50} 上分别达到了 51.2% 和 55.6%,较现有方法提升了 1.9% 和 2.7%,能够有效识别并准确解析不同尺度下的人体部位和服装配饰,提高多人解析精度.此外,在 CIHP 数据集上的测试,进一步验证所提算法具有一定的鲁棒性.

关 键 词: 多人解析;可变形注意力;多尺度特征;交叉注意力;实例-类别特征融合
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)02-0435-08

Multi-human Parsing Algorithm with Deep Fusion Instance-category Feature

YE Lanqing[1], LIU Li[1,2], FU Xiaodong[1,2], LIU Lijun[1,2], PENG Wei[1,2]

[1](Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)
[2](Computer Technology Application Key Lab of Yunnan Province, Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: To address the challenges of strong stages coupling, insufficient multi-scale feature representation, and limited parsing precision in multi-human parsing, a deeply fused instance-category features parsing algorithm is proposed. Firstly, the algorithm defines coding sets for human body parts, keypoints, and clothing accessories in the feature space and applies deformable attention to effectively learn human multi-scale features. Secondly, the multi-scale features are used as shared semantic information to achieve unified representation of instance and category features. Finally, cross-attention is introduced to deeply fuse instance and category features, generating accurate multi-human parsing results. Experimental results on the MHP V2.0 dataset show that AP^{P}_{vol} and PCP_{50} are 51.2% and 55.6% respectively, which are improved by 1.9% and 2.7% compared with the state-of-the-art methods. These results indicate the effectiveness of the method in recognizing and accurately parsing human parts and clothing accessories at various scales, significantly improving parsing precision. Furthermore, the results on the CIHP dataset further validate the robustness of the proposed algorithm.
Keywords: multi-human parsing; deformable attention; multi-scale feature; cross attention; instance-category feature fusion

0 引 言

多人解析[1]旨在识别多人图像中的每一个人体实例,并对每个人体实例进行细粒度的部位解析,可广泛应用于虚拟试穿、行人重识别、视频监控、人机交互和动作分析等领域.目前,多人解析方法根据不同解析流程可分为 2 类:1)自顶向下方法.首先检测图像中的所有人体实例,再对每一个人体实例单独进行人体解析;2)自底向上方法.首先检测图像中的所有人体部位,然后通过聚类等后处理方式将这些部位分配给不同的人体实例.

区别于单人解析,多人解析能够有效处理复杂场景下的多个人体实例,更符合实际应用需求,具有重要的社会价值.随着深度学习和全卷积网络[2]的快速发展,许多高效的人体解析模型相继被提出.如 Zhang 等[3]提出了一种无锚点实例级人体解析网络,通过人体检测器和边缘引导解析头实现像素级的人体解析.Zhou 等[4]提出了新的自底向上方法,即多粒度人体表示学习,该方法通过充分利用人体在不同细粒度上的结构信息,以端到端的方式同时完成多人解析和姿态估计.尽管上述方法取得了一定的进展,但通常需要依赖预检测或后分组处理,将人体解析分成了两个阶段.这种阶段间的强耦合性使得第 1 阶段的检测结果直接影响第 2 阶段的解析,而第 2 阶段的分组往往又高度依赖第 1 阶段的输出,从而导致实例和类别特征分离,降低了解析结果的精确性.

为了解决上述问题,Dai 等[5]提出一种新颖的多人解析

收稿日期:2024-12-23 收修改稿日期:2025-02-12 基金项目:国家自然科学基金项目(62262036,62362043)资助;云南省兴滇英才支持计划项目(KKXY202203008)资助. 作者简介:**叶岚清**,男,1999 年生,硕士研究生,CCF 学生会员,研究方向为计算机图形学与计算机视觉、图像处理;**刘 骊**(通信作者),女,1979 年生,博士,教授,博士生导师,CCF 高级会员,研究方向为计算机图形学与计算机视觉、图像处理;**付晓东**,男,1975 年生,博士,教授,博士生导师,CCF 高级会员,研究方向为服务计算、决策理论与方法;**刘利军**,男,1978 年生,博士,副教授,CCF 会员,研究方向为图像处理、云计算;**彭 玮**,女,1980 年生,博士,教授,博士生导师,CCF 会员,研究方向为机器学习、数据挖掘.

网络 ReSParser,将多人解析任务建模为层次化集合预测问题,通过估计人体部位的多个关键标记点,逐步构建从粗到细的代表性集合,以表示人体实例特征并编码人体与部位的关系. 随后,利用这些代表性集合生成与部位相关的卷积核,从而实现高质量的多人解析结果. Chu 等[6]提出了 SMP 框架,借鉴单阶段多人姿态估计方法,将多人解析简化为定位人体实例与使用点集及偏移量对部位进行分类两个任务. 虽然上述方法在推理速度上具有一定的优势,但其却较难准确捕捉人体小尺度部位(如手部、脚部等)和服装配饰(如领带、手表等),且多尺度表征能力欠缺,导致解析结果不够精确.

综上所述,多人解析仍存在以下难点:1)现有方法对人体多尺度表征不充分,难以准确捕捉人体小尺度部位和服装配饰,导致解析结果的完整性和准确性受限;2)由于现有单阶段方法未能充分利用人体多尺度特征作为共享语义信息,缺乏对实例特征和类别特征的统一建模,降低了多人解析结果的精度;3)大多数现有方法采用双阶段策略,阶段间的强耦合性不仅显著降低了模型的推理效率,还增加了阶段间误差累积的风险,限制了模型的鲁棒性.

针对以上难点,本文提出了一种深度融合实例-类别特征的多人解析算法. 通过引入可变形注意力机制,对定义的 3 种人体编码集进行特征提取,增强对人体多尺度特征的表征能力;充分利用人体多尺度特征作为共享语义信息,实现对人体实例和类别特征的统一表示,提升多人解析的完整性和精准性;采用单阶段策略,结合交叉注意力机制和注意力权重矩阵,对人体实例-类别特征进行深度融合,提高多人解析的鲁棒性.

1 相关工作

大多数方法将多人解析任务分为两个阶段,分别采用“先检测后解析”或“先解析后分组”的策略,逐步得到解析结果. Sun 等[7]利用特征金字塔网络(Feature Pyramid Network, FPN)和卷积神经网络[8](Convolutional Neural Networks, CNN)进行人体检测,随后基于检测结果辅助完成姿态估计. He 等[9]提出了“RoIAlign”区域表示方法,通过区域卷积神经网络直接从原始图像中提取人体表示,显著增强了人体解析的能力. Ren 等[10]进一步提出了一种精确的两阶段多人解析方法,第一阶段使用基于锚点的检测器提取目标区域,第二阶段利用解析模板完成多人解析. 之后,Yang 等[11]提出了解析 R-CNN 网络,通过引入全局语义特征金字塔和解析重评分网络,提升了解析性能. Luo 等[12]提出一种基于多层次的深度特征交换网络,既兼顾不同分辨率下高维特征学习,又可以满足不同分辨率下的特征交换学习. Wei 等[13]则提出了一种多任务学习方法,该方法首先检测图像中的所有人体实例,然后结合姿态、边缘和实例掩模信息,来共同促进实例级人体解析. 虽然上述方法在解析结果上表现较好,但其双阶段策略逐步处理的性质导致阶段间存在强耦合性,不仅缺乏端到端的优化,还阻碍了检测阶段与解析阶段之间的信息交互. 此外,由于推理速度与检测到的人体实例数量成正比,降低了多人解析的效率.

可变形注意力机制能够使模型更关注人体的局部区域,显著增强表征能力,提高精确性与鲁棒性. 如 Truong 等[14]提出基于可变形注意力蒸馏学习的自监督视频对象分割方法,既学习到精确的对象表征,又适应空间与时间的动态变化. Mao 等[15]通过结合人体边缘与掩模检测,并引入可变形注意力,自适应地解决因姿态变化导致的人体部位变形问题,从而增强了对人体部位的表征能力. 上述方法充分利用了可变形注意力机制,提升了模型的对局部区域的表征能力. 然而,这一机制并未较好地应用于多人解析任务中,导致解析精度受限. Cheng 等[16]结合实例标记区分技术与高分辨率特征表示,提出了一个更加鲁棒的多人解析模型. Chu 等[17]设计了一种推理效率较高的人体解析方法,并通过联合优化过程获得解析结果. 尽管以上方法在人体表征方面取得了一定进展,但未充分考虑小尺度部位特征,或仅在单一尺度下表征,忽略了人体部位、关节点以及服装配饰对解析结果的影响,从而导致解析结果的完整性和准确性受限.

针对上述问题,区别于 Chu 等[17]提出的快速高效特征表示学习模型,以及 Mao 等[15]提出的基于掩模引导的可变形自适应单人解析网络,本文提出通过使用可变形注意力引导特征提取来学习人体多尺度特征,并将其作为共享语义信息,对实例-类别特征进行统一表示,最后利用交叉注意力机制对实例-类别特征进行深度融合,从而生成更加精准且鲁棒的多人解析结果.

2 深度融合实例-类别特征的多人解析算法

本文提出的深度融合实例-类别特征的多人解析算法对应的流程图如图 1 所示,以多人图像作为输入,首先通过 ResNet-101 和特征金字塔网络[8](FPN)提取人体基本特征;其次,定义人体部位、关节点、服装配饰 3 种编码集,引入可变形注意力机制对 3 种编码集进行特征建模并拼接得到人体多尺度特征;然后将人体多尺度特征作为共享语义信息,用于后续实例-类别特征的统一表示,其分别提取得到人体实例特征和类别特征;最后利用交叉注意力和注意力权重矩阵深度融合人体实例特征与类别特征,生成精准的多人解析结果.

2.1 人体多尺度特征编码

首先,在特征空间中对人体进行多尺度表征,将人体结构定义为 3 种尺度下的人体编码集,并结合带有位置信息的嵌入矩阵,以增强不同尺度特征的对齐能力. 最后,通过引入可变形注意力对特征进行提取与拼接,生成丰富的人体多尺度特征.

如图 1 所示,将 ResNet-101 作为主干网络,从输入的多人图像中提取人体分层特征,表示为 $C=\{C_2,C_3,C_4,C_5\}$,其中 C_2、C_3、C_4 和 C_5 分别对应 ResNet 不同分层的输出特征图;再采用特征金字塔网络对分层特征进行逐层融合,通过双线性插值的上采样操作逐步恢复空间分辨率,设置上采样的比例为 2,当前层的特征图 F_l 定义为:

$$F_l = UpSample(F_{l+1}, scale=2) \tag{1}$$

其中,F_{l+1} 为上一层特征图,F_l 为当前层特征图. 然后通过 1×1 卷积对低层特征进行通道降维,并与上采样后的高层特征进行加权融合得到人体基本特征,其融合过程表示为:

$$F_{base} = \sum_{i=2}^{5}\omega_i(Conv_{1\times1}(F_{l-1}) + UpSample(F_l)) \tag{2}$$

其中,ω_i 是每层特征的加权系数,$Conv_{1\times1}$ 表示 1×1 降维卷积.

根据人体拓扑结构,定义了 3 种不同尺度的编码集. 首先,将人体的 11 个部位划分为上半身和下半身两个子集,以凸显人体部位的层次化特征,便于模型捕捉人体局部与全局关系. 其中每个部位被定义为一个人体部位编码 α,并通过 1×1卷积生成人体部位编码集 $H^i_{body}=\{\alpha_1,\alpha_2,\cdots,\alpha_7\}\cup\{\alpha_8,\alpha_9,\cdots,\alpha_{11}\}$;然后,基于关节点对称性,将 17 个关节点分为两

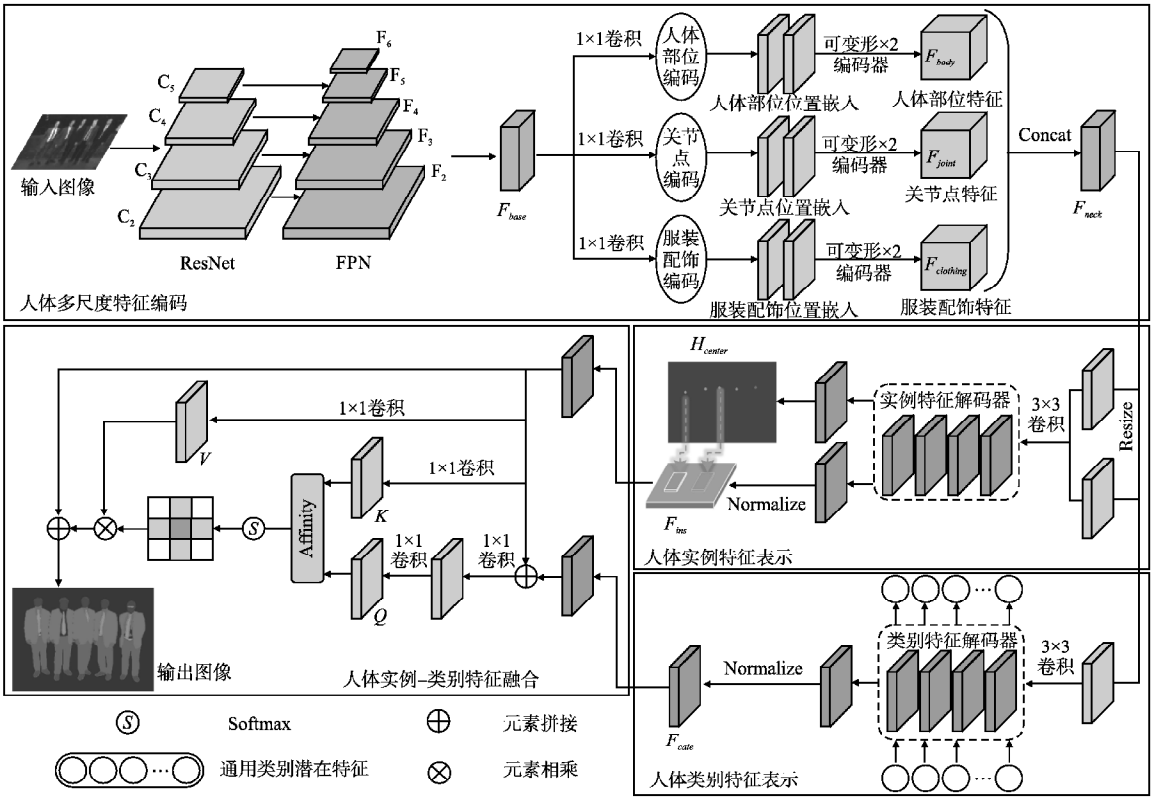


图 1　深度融合实例和类别特征的多人解析算法流程图

Fig. 1　Flowchart of multi-human parsing algorithm for deep fusion of instance and category features

个子集,以减少关节点间的冗余干扰. 每个关节点的二维坐标定义为一个关节点编码 β,并通过 1×1 卷积生成得到关节点编码集 $H^i_{joint}=\{\beta_1,\beta_2,\beta_3\}\cup\{\beta_4,\beta_5,\cdots,\beta_{17}\}$;最后,根据人体服装配饰在不同位置的分布,将 47 种服装配饰划分为头部、手部、上半身和下半身 4 个子集,促进模型学习服装配饰与对应人体区域的语义关系,提升解析精度. 每种服装配饰被定义为一个服装配饰编码 γ,并通过 1×1 卷积得到服装配饰编码集 $H^i_{clothing}=\{\gamma_1,\gamma_2,\cdots,\gamma_{14}\}\cup\{\gamma_{15},\gamma_{16},\cdots,\gamma_{30}\}\cup\{\gamma_{31},\gamma_{32},\cdots,\gamma_{42}\}\cup\{\gamma_{43},\gamma_{44},\cdots,\gamma_{47}\}$.

每个人体编码集都包含多个人体部位或服装配饰信息,因此将 H^i_{body}、H^i_{joint}、$H^i_{clothing}$ 作为输入,并利用可变形注意力机制中添加的可学习空间位置嵌入矩阵,保留不同尺度人体编码之间的相对位置,该过程定义为:

$$\begin{cases}A^i_{body}=Deform(H^i_{body})+E_{body}\\A^i_{joint}=Deform(H^i_{joint})+E_{joint}\\A^i_{clothing}=Deform(H^i_{clothing})+E_{clothing}\end{cases}\tag{3}$$

其中,$Deform(\cdot)$ 是可变形卷积层,E_{body}、E_{joint}、$E_{clothing}$ 分别表示人体部位、关节点、服装配饰位置嵌入矩阵. 再将 A^i_{body}、A^i_{joint}、$A^i_{clothing}$ 输入到 6 个依次堆叠的可变形编码器,并加入卷积层进行特征提取. 第 1、2 个编码器处理人体部位特征,第 3、4 个编码器处理对应人体关节点特征,第 5、6 个编码器处理对应人体服装配饰特征,以人体部位矩阵为例,其提取过程定义为:

$$F_{body}=Conv(DMSA(A^i_{body}))+EFN(LN(DMSA(A^i_{body})))\tag{4}$$

其中,$DMSA(\cdot)$ 是可变形注意力,$LN(\cdot)$ 为 LayerNorm 归一化层,$EFN(\cdot)$ 为可变形编码器的前馈神经网络. 同理,可得到关节点特征 F_{joint} 和服装配饰特征 $F_{clothing}$,最终将这 3 种尺度下的特征在统一通道维度上进行拼接得到人体多尺度特征 F_{neck}.

2.2　人体实例-类别特征表示

基于 2.1 节得到的人体多尺度特征 F_{neck} 作为后续实例-类别特征表示模块的共享语义信息输入,并将最终的输出统一为像素级分割结果.

2.2.1　人体实例特征表示

由于实例信息与每个实例的位置重心密切相关,为了准确定位人体实例的重心,本文采用一种基于实例的动态位置分布策略. 具体地,首先通过双线性插值法将人体多尺度特征 F_{neck} 调整至固定大小 S×S(其中 S=40). 然后,将调整后的特征与相对坐标连接,并输入到实例特征解码器中. 经过 5 个具有 3×3 内核的卷积层处理后,生成形状为 S×S×1 的重心热图 $H^{S\times S\times1}_{center}$,其过程定义为:

$$H_{center}^{S\times S\times 1}=Sigmoid(g_{CL}(Resize(F_{neck}^{C\times H\times W},(S,S)))) \quad (5)$$

其中 $g_{CL}(\cdot)$ 表示堆叠的卷积层函数，S 为特征比例大小，$Sigmoid(\cdot)$ 表示激活函数.

为了进一步增强实例之间的区分性，本文对属于同一实例的像素特征进行聚类处理. 将人体多尺度特征输入到实例特征编码器中，经过 4 个具有 3×3 内核的卷积层后，再将得到的特征图沿通道维度进行归一化操作，使得每个像素特征的长度为 1，从而生成归一化特征 F_{nor}，其过程表示为：

$$F_{nor}=Normalize(g_{ins}(F_{neck})) \quad (6)$$

其中，$g_{ins}(\cdot)$ 表示多个卷积层函数，$Normalize(\cdot)$ 是归一化操作.

通过重心热图获取重心坐标 (x_i,y_i)，并从归一化的特征图 F_{nor} 中提取对应像素的重心特征 f_{x_i,y_i}，其过程表示为：

$$f_{x_i,y_i}=F_{nor}[x_i,y_i] \quad (7)$$

其中，$[x_i,y_i]=WHERE(H_{center}^{S\times S\times 1}>\theta_c)$，$WHERE(\cdot)$ 函数旨在选择特定实例的坐标，$i\in\{1,\cdots,N_c\}$，N_c 表示实例坐标数量.

最后将重心特征作为卷积核，与归一化特征图进行卷积操作得到人体实例特征 F_{ins}，其过程可定义为：

$$F_{ins}=Concat(f_{x_i,y_i})\times F_{nor} \quad (8)$$

其中，$Concat(\cdot)$ 表示特征拼接，× 是卷积操作.

2.2.2 人体类别特征表示

实例特征因实例而异，但类别特征是独立于实例的信息，因此，本文在类别特征解码器中预定义了一组通用的人体类别潜在特征 $f_S\in R^{K\times D}$，其中 K 代表部位类别的数量，D 是特征的维度. 这些人体类别潜在特征在初始化时通过随机生成，并在训练过程中不断更新，以适应不同的人体类别特征的变化.

为了将通用的类别特征转化为特定实例的类别特征，首先将人体多尺度特征输入到类别特征解码器中，并利用各部位边界框的中心点作为参考点，这些参考点提供了人体部位的位置先验信息. 随后，基于这些参考点，使用可变形注意力机制构建每个部位的特定表征. 具体而言，通用类别潜在特征会根据各部位的参考点提取相关特征点，并通过自适应聚合机制对这些特征点进行聚合，从而生成特定的类别特征 $f'_{S,i}$，其过程定义为：

$$f'_{S,i}=\sum_{j=1}^{M}A_{ij}\cdot F_{neck}(p_i+\Delta p_{ij}) \quad (9)$$

其中，M 为采样点的数量，A_{ij} 表示参考点 p_i 与特征点 $p_i+\Delta p_{ij}$ 的相似性，$F_{neck}(p_i+\Delta p_{ij})$ 表示从多尺度特征中采用的特征点，Δp_{ij} 是参考点 p_i 的偏移量. 最后，将特定的部位类别潜在特征与多尺度特征 F_{neck} 进行 3×3 内核卷积操作，并进行归一化处理，最终生成人体类别特征 F_{cate}，其过程定义为：

$$F_{cate}=Normalize(Conv_{3\times 3}(f'_{S,i}\times F_{neck})) \quad (10)$$

其中，$Conv_{3\times 3}(\cdot)$ 表示 3×3 卷积，$Normalize(\cdot)$ 表示归一化处理.

2.3 人体实例-类别特征融合

为了得到更精准且鲁棒的多人解析结果，本文基于 2.2 节的实例特征 $F_{ins}\in\mathbb{R}^{C\times H\times W}$ 和类别特征 $F_{cate}\in\mathbb{R}^{C\times H\times W}$ 进行实例-类别特征的深度融合，并将沿通道维度拼接后的特征通过 1×1 内核的卷积层得生成新的融合特征 F_{fusion}. 接着，将 F_{fusion} 和 F_{ins} 分别通过一个 1×1 内核的卷积层，生成查询特征矩阵 Q 和键特征矩阵 K.

为了计算 Q 和 K 之间的相似度，针对特征矩阵 Q 中的每一个像素点 $p(i_p,i_p)$，将其表示为向量 $Q_p\in\mathbb{R}^{C'}$. 同时，从特征矩阵 K 中提取与像素点 $p(i_p,i_p)$ 位于同一行或同一列的所有其他像素点的特征矩阵 $K_p\in R^{(H+W-1)\times C'}$，具体过程定义如下：

$$K_p=\{(K^{i,j_p}\mid i\in[1,H],i\neq i_p)\cup(K^{i_p,j}\mid j\in[1,W],j\neq j_p)\} \quad (11)$$

其中，K^{i,j_p} 为特征矩阵 K 中像素点 $p(i,i_p)$ 对于的特征向量，$K^{i_p,j}$ 则为像素点 $p(i_p,j)$ 对应的向量.

为了计算特征矩阵 Q 中像素点 p 与特征向量 K_p 中第 m 个像素点的相关性，首先将特征向量 Q_p 与 K_p^m 进行点积运算，再对结果应用 Softmax 操作，最终得到 Q_p 与 K_p^m 之间的相关性 S_p^m，其过程定义如下：

$$S_p^m=Softmax(Q_pK_p^m) \quad (12)$$

其中，$S_p^m\in S$，S 代表特征矩阵 Q 与特征矩阵 K 之间的注意力权重矩阵.

同时，将人体实例特征 F_{ins} 通过一个 1×1 内核的卷积层生成新的特征矩阵 $V\in\mathbb{R}^{C\times H\times W}$. 与特征矩阵 K_p 的提取方法类似，对于特征矩阵 V 中的每一个像素点 p，生成一个包含其交叉方向像素点信息的特征矩阵 $V_p\in R^{(H+W-1)\times C}$.

最后，将特征向量 V_p 中第 m 个像素点的解析特征 V_p^m 与对应的注意力权重 S_p^m 逐元素相乘，得到加权后的注意力输出. 接着，将该输出进一步与实例特征融合，生成最终的人体解析结果 $F_{parsing}$. 深度融合得到多人解析结果 $F_{parsing}$ 的具体过程定义如下：

$$F_{parsing}=\sum_{m\leqslant(H+W-1)}S_p^mV_p^m+F_{ins} \quad (13)$$

为了优化实例-类别特征深度融合过程，本文定义了 4 个损失函数：重心热图损失 $\mathcal{L}_{\text{center}}$、实例特征损失 $\mathcal{L}_{\text{instance}}$、类别特征损失 $\mathcal{L}_{\text{category}}$ 和解析损失 $\mathcal{L}_{\text{parsing}}$. 其中，$\mathcal{L}_{\text{center}}$ 用于计算人体实例重心热图与真实值之间的差距，$\mathcal{L}_{\text{instance}}$ 用于优化实例特征的表示，$\mathcal{L}_{\text{category}}$ 用于优化类别特征的表示，$\mathcal{L}_{\text{parsing}}$ 用于监督多人解析结果的像素级准确性. 通过对以上 4 个损失函数进行加权求和，得到最终的总损失 $Loss$，其计算公式为：

$$Loss=\alpha_1\mathcal{L}_{center}+\alpha_2\mathcal{L}_{instance}+\alpha_3\mathcal{L}_{category}+\alpha_4\mathcal{L}_{parsing} \quad (14)$$

其中 α_1、α_2、α_3、α_4 表示各损失对应的权重.

3 实 验

3.1 数据集与评估指标

为了验证本文所提算法的有效性，在两个具有挑战性的公用数据集 MHP V2.0[18] 和 CIHP[19] 上进行了实验. MHP V2.0[18] 数据集专注于多人解析，包含 25403 幅图像，每幅图像平均包含 3 个人体实例. 该数据集提供 11 种人体部位类别标签和 47 种服装配饰标签，能够有效评估本文方法在解析小尺度人体部位和服装配饰方面的有效性和准确性；CIHP[19] 数据集是一个大型综合性多人解析数据集，专注于复杂的野外场景任务，共包括 38280 幅图像，其中约 74% 用于训练，13% 用于验证，13% 用于测试. 该数据集能够全面评估多人解

析模型在实际应用中的鲁棒性. 本文通过对图像短边长度在640 ~ 800 像素范围内随机采样的方式进行尺度抖动,从而生成具有不同分辨率的图像.

本文在实验中采用多人解析单阶段方法中常见的 3 个评估指标. 基于部位的平均精度(Average Precision based on Part,AP^p)用于评估模型在特定人体部位上的分割性能,主要使用 AP^p_{vol} 和 AP^p_{50}. 其中,AP^p_{vol} 表示在多个真实值的交并比(Intersection over Union,IoU)阈值下计算的平均精度;AP^p_{50} 是指在 IoU 阈值设置为 50% 时,预测结果被认为是正确的解析;正确解析的语义部位百分比(Percentage of Correctly parsed semantic Parts,PCP)衡量模型正确解析出的语义部位在所有部位中的占比. 以上 3 个评估指标均用于衡量解析结果的准确性,以百分比(%)为单位进行度量. 指标数值越大,表示模型性能越优.

3.2　实验环境及设置

本文实验运行在以下硬件平台:32 核 CPU Intel(R)Xeon(R)Platinum 8358 CPU @2.60GHz,内存容量为 512GB. 采用 PyTorch 作为深度学习框架. 模型训练与测试均基于一张 24GB 显存的 RTX 3090 GPU 完成.

在实验过程中,本文选择 ResNet-101 作为模型的主干网络,初始学习率设为 0.005. 训练过程共进行 36 个 epoch,每 12 个 epoch 作为一轮. 在第 27 个和第 33 个 epoch 时,将学习率缩减至原来的 1/10,以进行性能对比分析.

3.3　实验结果及对比分析

3.3.1　定量结果对比分析

为了验证本文方法在解析精准性和鲁棒性方面的优势,首先在最具挑战性的多人解析数据集 MHP V2.0[18] 上,与现有基于自顶向下、自底向上双阶段方法以及单阶段方法进行了定量对比,结果如表 1 所示.

表 1　在 MHP 2.0 数据集上的定量结果

Table 1　Quantitative results on MHP v2.0 dataset

方法	主干网络	训练轮次	AP^p_{vol}	AP^p_{50}	PCP_{50}
双阶段自顶向下方法					
M-CE2P[1]	ResNet-101	150	42.7	34.5	43.7
SNT[20]	ResNet-101	–	42.5	34.4	43.5
RP-RCNN[11]	ResNet-50	150	46.8	45.3	43.8
AIParsing[3]	ResNet-101	75	46.6	43.2	47.3
双阶段自底向上方法					
PGN[19]	ResNet-101	–	35.5	17.6	26.9
MHParser[21]	ResNet-101	–	36.1	18.0	27.0
NAN[18]	–	80	41.8	25.1	32.3
DSPF[4]	ResNet-101	150	44.3	39.0	42.3
单阶段方法					
ReSParse[5]	ResNet-101	75	42.7	34.3	43.7
CIParsing[22]	ResNet-101	75	46.3	40.9	52.0
SMP[6]	ResNet-101	36	48.2	47.1	51.5
UniParser[17]	ResNet-101	36	49.3	**51.2**	52.9
Ours	ResNet-101	36	**51.2**	<u>50.8</u>	**55.6**

注:粗体表示各列最优值,下划线表示次优值.

从表中可以看出,本文方法在多个评估指标上均展现了显著的性能提升. 具体而言,与现有单阶段方法中表现最好的 UniParser[17] 相比,本文方法在基于部位的平均精度指标 AP^p_{vol} 上提升了 1.9%(从 49.3% 提升到了 51.2%),表明本文方法在多个 IoU 阈值条件下,能够保持一致且高效的分割性能.

此外,在人体语义部位解析正确性指标 PCP_{50} 上,本文方法提升了 2.7%(从 52.9% 提升到 55.6%),证明其在区分和解析人体不同语义部位(包括小尺度部位和服装配饰)方面更加准确. 虽然在指标 AP^p_{50} 上仅获得次优结果,但仍达到了 50.8% 的高水平,展现了在解析精准性上的显著优势. 以上对比方法的实验数据均来源于原论文.

为进一步验证本文方法在复杂野外场景下的解析优势,选取了 CIHP[19] 数据集进行评估与分析. 基于多人解析方法的 3 个评估指标 AP^p_{vol}、AP^p_{50} 和 PCP_{50} 的定量对比结果如表 2 所示.

表 2　在 CIHP 数据集上的定量结果

Table 2　Quantitative results on CIHP dataset

方法	主干网络	训练轮次	AP^p_{vol}	AP^p_{50}	PCP_{50}
双阶段自顶向下方法					
P-RCNN[23]	ResNet-101	75	55.9	69.1	66.2
M-CE2P[1]	ResNet-101	150	48.9	54.7	–
SNT[20]	ResNet-101	–	52.0	58.9	–
RP-RCNN[11]	ResNet-50	150	58.3	71.6	62.2
AIParsing[3]	ResNet-101	75	60.3	75.2	68.5
双阶段自底向上方法					
PGN[19]	ResNet-101	–	35.5	17.6	26.9
单阶段方法					
ReSParser[5]	ResNet-101	75	56.4	69.2	65.0
CIParsing[22]	ResNet-101	75	59.6	74.6	69.3
SMP[6]	ResNet-101	36	57.3	71.7	64.5
UniParser[17]	ResNet-101	36	60.4	**75.9**	69.0
Ours	ResNet-101	36	**62.1**	<u>75.7</u>	**71.5**

注:粗体表示各列最优值,下划线表示次优值.

从表中可以看出,本文方法在其中两个评估指标上取得了显著的提升. 具体地,与现有方法相比,本文方法在评估指标 AP^p_{vol} 和 PCP_{50} 上分别提升了 1.7%(从 60.4% 提升至 62.1%)和 2.5%(从 69.0 提升到了 71.5%),表明本文方法在分割精度和语义部位解析能力方面具有更强的优势. 此外,在指标 AP^p_{50} 上取得了次优结果,达到了 75.7%,进一步验证了本文方法在野外场景中的鲁棒性和解析的精准性. 以上对比方法的实验数据均来源于原论文.

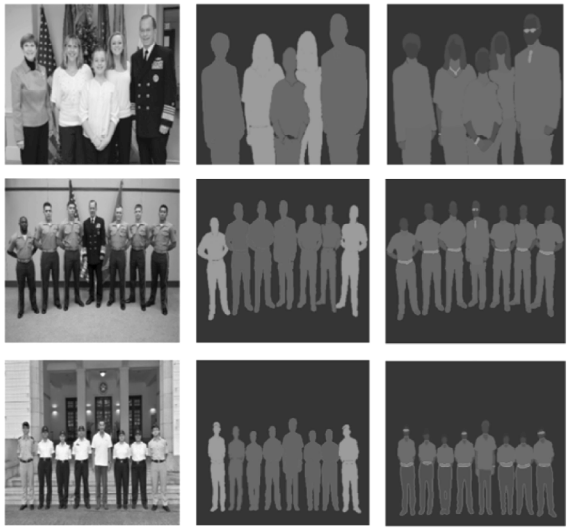
3.3.2　定性结果对比分析

本文分别在 MHP V2.0[18] 和 CIHP[19] 两个多人解析数据集上进行了定性实验,结果如图 2 和图 3 所示. 从图中可以看出,本文方法在多种复杂场景下表现出较高的准确性.

在 MHP V2.0[18] 数据集中,本文方法能够更准确地解析人体的细节部分,如脚部、颈部等小部位. 此外,对于不同形状、尺寸大小的服装配饰(如眼镜、领带、项链等),本文方法的分割边界更加精确,显示其在细粒度解析中的强大能力. 在 CIHP[19] 数据集中,本文方法展现了对多样化野外场景的鲁棒性. 无论是光照变化还是背景复杂情况下,本文方法均能保持高质量的解析结果.

为了进一步验证本文方法在人体小尺度部位和服装配饰表征方面的有效性,以及对复杂多样野外场景的鲁棒性,本文基于公开源码,与 UniParser[17] 单阶段方法进行了定性对比,

结果如图 4 和图 5 所示.



(a)原始图像 (b)实例分割结果 (c)多人解析结果

图 2 在 MHP V2.0 数据集上的定性结果

Fig. 2 Qualitative results on MHP V2.0 dataset

(a)原始图像 (b)实例分割结果 (c)多人解析结果

图 3 在 CIHP 数据集上的定性结果

Fig. 3 Qualitative results on CIHP dataset

从对比图中可以看出,UniParser[17]会出现如下问题:1)实例过多导致错误分割:当图像中人体实例数量较多时,容易出现分割错误. 如图 4 第 1 列中,人体实例头顶被错误地解析出多余配饰信息;2)遮挡下的服装配饰解析困难:当服装配饰被人体部位遮挡时,难以准确解析被遮挡的配饰. 如图 4 第 3 列中,耳环被头发遮挡时,UniParser[17]未能准确识别出耳环;3)野外场景中小尺度部位和配饰丢失:在复杂的野外场景下,难以有效解析部分小尺度部位和服装配饰. 如图 5 中人体所穿的衬衫、腰带和眼镜等小尺度配饰未被识别并解析.

为了验证所提算法在多人解析任务中的有效性,本文在 MHP V2.0[18]数据集上对性能指标 AP^{p}_{vol}、AP^{p}_{50} 和 PCP_{50} 进行了详细评估与分析,结果如表 3 所示. "Baseline"表示仅使用主干网络的基准模型,表中的"M"、"IC"和"F"分别是人体多

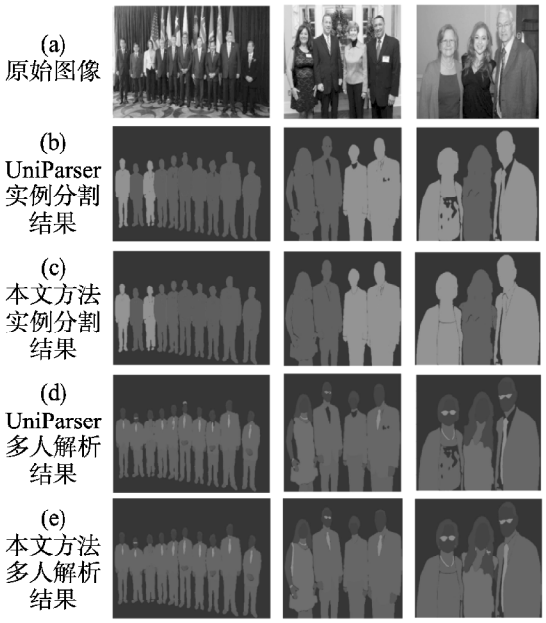


图 4 在 MHP V2.0 数据集上的定性对比结果

Fig. 4 Qualitative comparison results on MHP V2.0 dataset

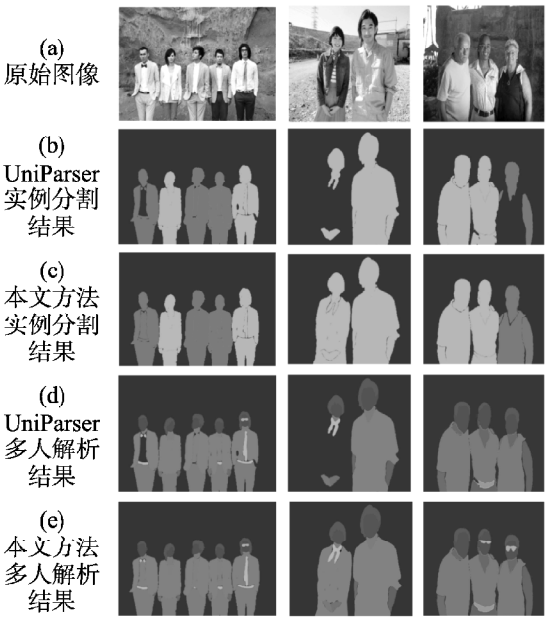


图 5 在 CIHP 数据集上的定性对比结果

Fig. 5 Qualitative comparison results on CIHP dataset

尺度特征模块、人体实例-类别特征表示模块和人体实例-类别特征融合模块. 从表中可以看出,将人体多尺度特征模块加入 Baseline 后,AP^{p}_{vol}、AP^{p}_{50} 和 PCP_{50} 的值分别提升了 2.0%、4.9% 和 0.8%,这一结果表明,引入可变形注意力机制能够增强模型对人体小尺度部位和服装配饰的捕捉,从而提高解析的精准性;在此基础上,进一步加入人体实例-类别特征表示模块后,AP^{p}_{50} 和 PCP_{50} 分别提升了 27.2% 和 24.3%,表明以

人体多尺度特征为统一输入生成的人体重心热图以及实例-类别特征能够有效区分不同实例,从而大幅提升人体解析精度;最后,加入人体实例-类别特征融合模块,AP^{p}_{vol}、AP^{p}_{50} 和 PCP_{50} 的值分别提升了 3.7%、4.6% 和 3.2%,进一步验证了实例-类别特征融合模块在提高解析的准确性和鲁棒性方面的作用,使模型能够生成高精度的多人解析结果.

表 3　在 MHP 2.0 数据集上验证各模块的有效性

Table 3　Validity of each module on MHP v2.0 dataset

模　型	AP^{p}_{vol}	AP^{p}_{50}	PCP_{50}
Baseline	34.6	15.9	27.3
Baseline + M	36.6	19.0	28.1
Baseline + M + IC	47.5	46.2	52.4
Baseline + M + IC + F	**51.2**	**50.8**	**55.6**

本文在 MHP V2.0 数据集上对最终解析结果进行了 Matrix-NMS 对比实验,结果如表 4 所示,无论是否加入非极大抑制(Non-Maxinum Suppression,NMS)过程,模型的性能指标均不会发生变化. 结果进一步验证了本文方法的有效性,同时表明其具备 NMS-free 特性,即无需依赖后处理步骤即可实现高精度的多人解析.

表 4　在 MHP 2.0 数据集上验证 NMS-free 的有效性

Table 4　Validity of NMS-free on MHP v2.0 dataset

NMS	AP^{p}_{vol}	AP^{p}_{50}	PCP_{50}
N	51.2	50.8	55.6
Y	51.2	50.8	55.6

注:"N":表示未加入非极大抑制过程,"Y"表示加入非极大值抑制过程.

表 5　在 MHP v2.0 验证集的模型运行时间结果

Table 5　Model runtime results on the MHP v2.0 val set

	方法	运行时间(ms)
自顶向下方法	SNT[20]	3546
	M-CE2P[1]	1023
	RP-RCNN[11]	341
自底向上方法	MHParser[21]	1224
	NAN[18]	997
	PGN[19]	524
单阶段方法	ReSParser[5]	110
	UniParser[17]	**102**
	本文方法	154

传统的自顶向下和自底向上双阶段方法通常依赖于两阶段策略,这种设计可能导致计算效率低下. 如表 5 所示,本文不仅在运行速度上优于自底而上的方法,同时在自顶向下方法中也表现出显著优势,其速度约为当前最快的自顶而下方法 RP-RCNN 的两倍. 与单阶段方法 ReSParser 相比,本文通过消除对 NMS 过程的依赖,使解析速度提升了 7.3%. 因此,本方法不仅在解析精度方面表现优异,还显著减少了计算开销. 以上对比方法的实验数据均来源于原论文.

4　结　论

针对多人解析中存在阶段间强耦合性、多尺度表征不足、解析精准性受限的问题,本文提出一种深度融合实例-类别特征的多人解析算法. 首先在特征空间中定义人体部位、关节点和服装配饰 3 种编码集,并通过可变形注意力引导的特征提取,来学习丰富的人体多尺度特征,提高模型对人体的表征能力;然后将人体多尺度特征作为共享语义信息,对人体实例-类别特征进行统一表示,以提高解析的精准性;最后引入交叉注意力机制对实例-类别特征进行深度融合,提高模型对野外场景下的鲁棒性,得到精准的多人解析结果. 然而,本文方法在面对严重遮挡场景时,解析准确性仍有待提高,此外,引入交叉注意力机制会导致模型复杂度较高. 未来的研究将致力于进一步提升多人解析的精准性和鲁棒性,同时优化网络结构以降低模型复杂性.

References:

[1] Ruan T, Liu T, Huang Z, et al. Devil in the details: towards accurate single and multiple human parsing[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 4814-4821.

[2] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[3] Zhang S, Cao X, Qi G J, et al. AIParsing: anchor-free instance-level human parsing[J]. IEEE Transactions on Image Processing, 2022, 31: 5599-5612, doi: 10.1109/TIP.2022.3192989.

[4] Zhou T, Wang W, Liu S, et al. Differentiable multi-granularity human representation learning for instance-aware human semantic parsing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 1622-1631.

[5] Dai Y, Chen X, Wang X, et al. Resparser: fully convolutional multiple human parsing with representative sets[J]. IEEE Transactions on Multimedia, 2023, 26: 1384-1394, 10.1109/TMM.2023.3281070.

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

[7] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5693-5703.

[8] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[9] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

[10] Ren S, He K, 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, 2016, 39(6): 1137-1149.

[11] Yang L, Song Q, Wang Z, et al. Renovating parsing R-CNN for accurate multiple human parsing[C]//European Conference on Computer Vision, Cham: Springer International Publishing, 2020: 421-437.

[12] LUO W J, NI P, ZHANG H. Human parsing method for multi-level

deep feature exchange[J]. Journal of Chinese Computer Systems, 2020,41(1):149-154.

[13] Wei Y,Liu L,Fu X,et al. Crowded pose-guided multi-task learning for instance-level human parsing[J]. Machine Vision and Applications,2023,34(4):46,10.1007/S00138-023-01392-4.

[14] Truong Q T,Nguyen D T,Hua B S,et al. Self-supervised video object segmentation with distillation learning of deformable attention [J]. arxiv preprint arxiv:2401.13937,2024.

[15] Mao A,Liang Y,Jiao J,et al. Mask-guided deformation adaptive network for human parsing[J]. ACM Transactions on Multimedia Computing,Communications,and Applications(TOMM),2022,18(1):1-20.

[16] Cheng B,Xiao B,Wang J,et al. Higherhrnet:scale-aware representation learning for bottom-up human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020:5386-5395.

[17] Chu J,Jin L,Teng Y,et al. UniParser:multi-human parsing with unified correlation representation learning[J]. IEEE Transactions on Image Processing,2024,33:5159-5171,doi:10.1109/TIP.2024.3456004.

[18] Zhao J,Li J,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,2018:792-800.

[19] Gong K,Liang X,Li Y,et al. Instance-level human parsing via part grouping network[C]//Proceedings of the European Conference on Computer Vision(ECCV),2018:770-785.

[20] Ji R,Du D,Zhang L,et al. Learning semantic neural tree for human parsing[C]//Computer Vision-ECCV 2020,16th European Conference,2020:205-221.

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

[22] Chen X,Wang X,Gao L,et al. CIParsing:unifying causality properties into multiple human parsing[J]. arXiv preprint arXiv:2308.12218,2023.

[23] Yang L,Song Q,Wang Z,et al. Parsing r-cnn for instance-level human analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:364-373.

附中文参考文献:

[12] 罗文劼,倪 鹏,张 涵. 多层次深度特征交换的人体解析方法[J]. 小型微型计算机系统,2020,41(1):149-154.