复杂场景下的多人人体姿态估计算法

石　磊[1,2]，　王天宝[1]，　孟彩霞[2,3]，　王清贤[1]，　高宇飞[1]，　卫　琳[1]

(1. 郑州大学 网络空间安全学院　河南 郑州 450002；2. 郑州大学 计算机与人工智能学院　河南 郑州 450001；3. 郑州警察学院 图像与网络侦查系　河南 郑州 450053)

摘要：复杂场景下人员的交叉遮挡,导致现有的人体姿态估计算法存在准确度不高和人体骨架错连的问题。为此,提出一种复杂场景下的多人人体姿态估计优化算法。首先,使用分组分块级联卷积替换普通卷积,结合特征融合促进特征通道之间的信息交互,在不引入额外计算成本的前提下提高算法精度;其次,引入空间注意力机制挖掘与人体姿态估计任务相关的空间语义特征,将网络结构并行化处理以提高算法性能;最后,对大卷积核和空间注意力机制的嵌入位置进行轻量化处理,减少时间开销。与现有的自底向上的姿态估计算法 OpenPifPaf++相比,所提算法在 COCO 2017 数据集上平均准确率提高 0.8 个百分点;在 CrowdPose 数据集上平均准确率比 OpenPifPaf 算法提高 1.2 个百分点,复杂场景下对应的准确率提高 1.5 个百分点。

关键词：复杂场景；多人人体姿态估计；分组卷积；空间注意力机制；轻量化

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1671-6841(2025)04-0001-07

DOI：10.13705/j.issn.1671-6841.2024027

Multi-person Pose Estimation Algorithm in Complex Scenes

SHI Lei[1,2], WANG Tianbao[1], MENG Caixia[2,3], WANG Qingxian[1], GAO Yufei[1], WEI Lin[1]

(1. *School of Cyber Science and Engineering*, *Zhengzhou University*, *Zhengzhou* 450002, *China*;
2. *School of Computer and Artificial Intelligence*, *Zhengzhou University*, *Zhengzhou* 450001, *China*;
3. *Department of Image and Network Investigation*, *Zhengzhou Police University*, *Zhengzhou* 450053, *China*)

Abstract: The cross-obscuration of individuals in complex scenes led to low accuracy and incorrect skeleton connections in existing human pose estimation algorithms. Therefore, a multi-person pose estimation optimization algorithm in complex scenes was proposed. Firstly, the ordinary convolution was replaced with the grouped cascade convolution, which was combined with feature fusion to promote the exchange of information between channels. The accuracy of the algorithm was improved without incurring additional computational costs. Secondly, the spatial attention mechanism was introduced to mine the spatial semantic features related to the human pose estimation task, and the network structure was parallelized to enhance the performance of the algorithm. Finally, the embedding positions of the large convolutional kernel and the attention mechanism were lightweighted to reduce temporal overhead. Compared to the existing bottom-up pose estimation algorithm OpenPifPaf++, the proposed algorithm improved the average accuracy by 0.8 percentage points on the COCO 2017 dataset. Compared with the OpenPifPaf algorithm, the proposed algorithm improved the average accuracy by 1.2 percentage points on the CrowdPose dataset, and the corresponding accuracy for complex scenes by 1.5 percentage points.

Key words: complex scene; multi-person pose estimation; group convolution; spatial attention mechanism; lightweight

收稿日期:2024-02-22
基金项目:国家自然科学基金项目(62006210);国家重点研发计划项目(2020YFB1712401-1)
第一作者:石磊(1967—),男,教授,主要从事计算机视觉研究,E-mail:shilei@zzu.edu.cn。
通信作者:孟彩霞(1982—),女,教授,主要从事计算机视觉研究,E-mail:mengcaixia@rpc.edu.cn。

0 引言

近年来,深度学习的飞速发展使得基于图像、视频的人体姿态估计技术取得了日新月异的进步。在简单清晰的场景下,现有的人体姿态估计算法在保证实时性的同时还拥有优异的准确度[1]。然而,当面临诸如火车站台、候车大厅等人群密集的复杂拥挤场景时,人体骨架丢失、遮挡、错连等问题使得现有的人体姿态估计算法的性能下降[2]。如何有效地提高复杂场景下人体姿态估计算法的准确度,是目前此类问题研究的重点和难点。

人体姿态估计算法大多采用基于卷积神经网络的方法,它们在性能上优于基于图形结构和可变部件模型的传统方法[3]。基于卷积神经网络的2D多人人体姿态估计技术分为自顶向下和自底向上2种方法。

自顶向下的方法利用人体检测器构建出人体边界框,然后在人体边界框内估计目标关键点的位置以及关键点之间的关联。该方法依靠检测器的更新优化以及大量人为标记的边界框,展现出优异的准确度和效率。但当面对复杂场景时,自顶向下的方法中人体边界框会出现重叠,进而导致不同关键点之间匹配混乱,性能大打折扣。

自底向上的方法首先估计出人体中的每个关键点,然后将预测关键点分组组合,构成多个人体姿势。凭借全局关键点关联匹配的姿态估计方式,在面对复杂场景时展现出较好的抗干扰能力,但是存在方法整体精度不高、关键点冗余以及不同个体之间关键点错连的问题。

为了解决上述问题,本文提出一种面向复杂场景的多人人体姿态估计优化算法。该算法充分关注复杂场景下人体关键点的定位和关联,有效地缓解了复杂场景对多人人体姿态估计任务的干扰。主要贡献如下:

1) 采用分组分块的卷积方式[4],结合特征融合获取不同特征通道间的关键点语义信息,促进特征通道之间的信息交互。

2) 引入CC Attention机制[5],并串联组成CCA模块,获取更加全面的关键点语义特征,提高算法性能。

3) 对算法进行轻量化处理,采用轻量型卷积,同时改变CCA模块的嵌入位置,降低算法额外的参数量。

1 相关工作

复杂场景下自底向上的多人人体姿态估计算法可以保持较好的鲁棒性。Pishchulin等[6]首次提出了一种自底向上的算法DeepCut,使用一个整数线性程序将属于同一个体的关键点关联起来,但处理时间需要数小时。为此,Cao等[7]提出了OpenPose,采用将贪婪解码器与其他定义场相结合的思路,使用多阶段反复迭代的卷积神经网络结构,结合部分置信图和部分关系场,大幅提高了多人人体姿态估计算法的效率。

上述方法在高分辨率图像中表现优异,不仅可以提高多人人体姿态估计的准确度,还减少了预测时间。但在分辨率有限、人员拥挤等复杂场景下,这些方法的表现往往不尽如人意。Kreiss等[8]提出了PifPaf,首次引入了级联场的概念。与OpenPose中的部分关系场[7]、PersonLab中的中间域[9]相比,级联场在复杂场景下可以产生更加精确的关键点关联。此外,PifPaf还可以解决不同人员之间骨架交叉的问题。后续,Kreiss等[10]又提出了OpenPifPaf,该算法主要由基础网络、2个级联场网络和解码器构成,很好地解决了个体之间关键点错连的问题,但在人体预测关键点完整度方面还存在改进的空间。

2 多人人体姿态估计算法

2.1 整体架构

本文提出轻量型卷积,使用3个小卷积核以级联表示的形式代替7×7大卷积核。姿态估计算法改进前后对比如图1所示。改进后的姿态估计算法由CC-ResNest基础网络和CIF、CAF级联场网络组成编码器,获取图像的关键点特征信息。其中CC-ResNest用于提取图像的高级语义信息,CIF用于表征语义关键点的强度,CAF用于表征不同关键点间的关联强度。最后,利用解码器将CIF字段和CAF字段转换为一个包含17个关键点的人体骨架姿态,每个语义关键点最终由坐标(x, y)和置信度分数表示。

2.2 ResNest基础网络

ResNest块内架构如图2所示。为获取不同特征通道间的关键点语义信息,促进通道间的信息交互,对ResNet网络进行分组分块,并引入Split Attention以特征融合的方式构成ResNest。借鉴ResNext[11]中分组卷积的原理,将特征图沿着通道维度依次进

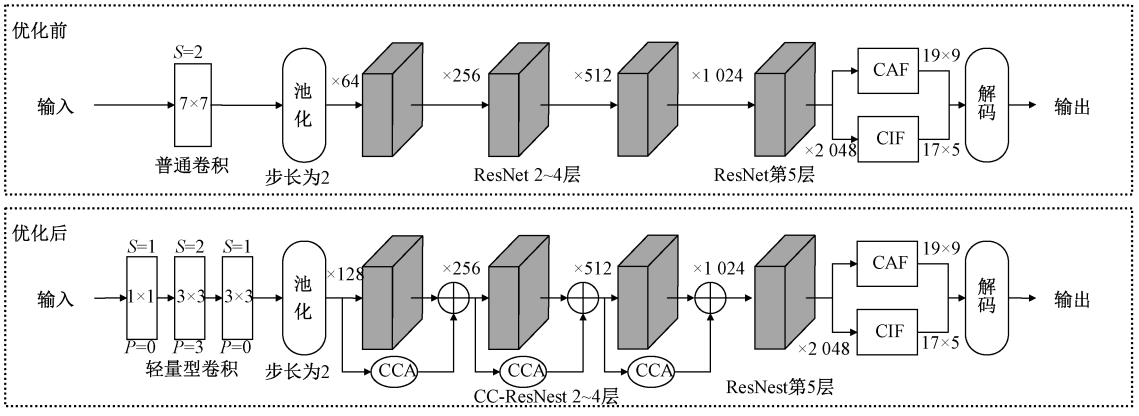

图 1　姿态估计算法改进前后对比

Figure 1　Comparison of pose estimation algorithms before and after improvement

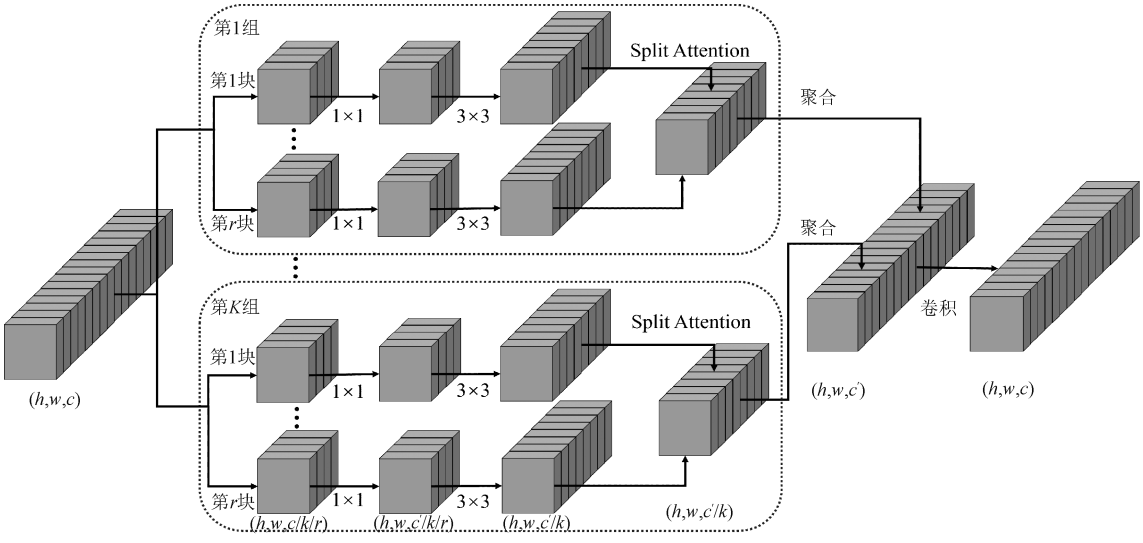

图 2　ResNest 块内架构

Figure 2　The intra block architecture of ResNest

行分组和分块处理。每组的特征图表示都是由组内各个 Split 加权确定,利用 Split Attention 实现特征融合,构成一个跨通道信息交互表示的 ResNest 基础网络。

Split Attention 块内架构如图 3 所示,由 r 个 Split 组成一个组,融合多个 Split 并按元素进行求和。随后,沿着空间维度进行全局平均池化,获得 1×1×c 的通道表示 $S^k \in \mathbf{R}^{C/K}$,以此收集通道的全局上下文信息,即

$$S_c^k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c'^k(i,j), \tag{1}$$

其中:$U_c'^k(i,j)$ 表示像素点数据;H 和 W 分别表示三维矩阵的行和列。

将输出特征图按通道进行软注意力聚合得到基数组表示 $V^k \in \mathbf{R}^{H \times W \times C/K}$,加权在初始的输入块上实现通道间的信息交互,即

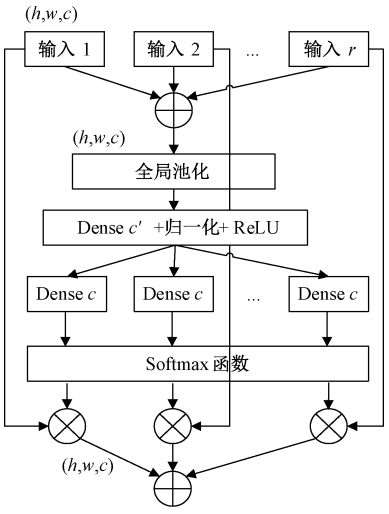

图 3　Split Attention 块内架构

Figure 3　The intra block architecture of Split Attention

$$V_c^k = \sum_{i=1}^{R} a_i^k(c) U_i, \tag{2}$$

其中：$a_i^k(c)$ 表示软分配权重；$\boldsymbol{U}_i$ 表示初始的特征矩阵。

相较于 ResNet，ResNest 利用 Split Atention 将特征图感受野覆盖到不同的特征图组，更加关注不同特征图组和不同通道之间的信息交互。凭借简单、模块化、不引入额外计算成本等优点，ResNest 更有助于下游任务性能的提升，如姿态估计、目标检测、语义分割等任务。

2.3 CCA 模块

复杂场景下易出现不同人体骨架之间遮挡的问题，单个骨架左右两侧区域的关键点之间仍然可能存在语义联系，为此引入 CCA 模块。CCA 模块由 2 个 CC Attention 机制串联组成，用于收集空间内像素点附近以及远处的各种语义信息，缓解骨架遮挡对多人人体姿态估计的干扰。CCA 模块的架构如图 4 所示。

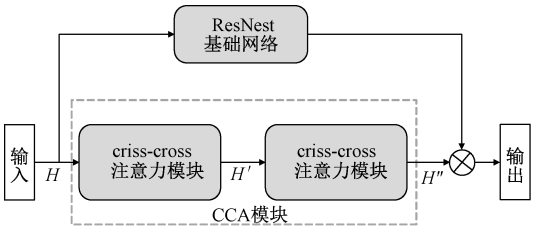

图 4 CCA 模块的架构

Figure 4 The architecture of CCA module

相较于 CC Attention，CCA 模块可以更有效地从远程依赖项中捕获上下文信息。输入特征图 H 经过 CC Attention 生成新的特征图 H'，该注意力机制仅在水平和垂直方向上聚合空间信息。将特征图 H' 再次输入 CC Attention 中生成 H''，确保结果特征图中每个像素点都可以收集空间中所有的像素信息，提取更加丰富的空间信息。此外，2 个 CC Attention 共享相同的参数，避免了添加过多的参数。

CC Attention 机制在像素点的水平和垂直方向上收集空间信息，每个像素点都会获取从其他位置收集的语义信息，达到增强空间语义信息的目的，CC Attention 机制的架构如图 5 所示。使用 2 个 1×1 卷积对输入特征图 H 进行降维后，得到空间大小为 $C'\times W\times H$ 的 Q 和 K，将 Q 和 K 通过仿射变换生成 A，具体过程为

$$d_{i,u} = Q_u \boldsymbol{M}_{i,u}^{\mathrm{T}},\ i \in [1, W+H-1], \tag{3}$$

其中：Q_u 表示位置 u 处的通道特征；$\boldsymbol{M}_u$ 表示位置 u 对应路径上的特征向量。最后，对 A 和 V 进行聚合操作并加权到初始的 H 上生成 H'，

$$H' = \sum_{i \in |\boldsymbol{\Phi}_u|} A_{i,u} \boldsymbol{\Phi}_{i,u} + H_u, \tag{4}$$

其中：$\boldsymbol{\Phi}_u$ 表示 V 中位置 u 的十字特征向量。上下文信息被添加到局部特征图 H，以增强空间特征表示。

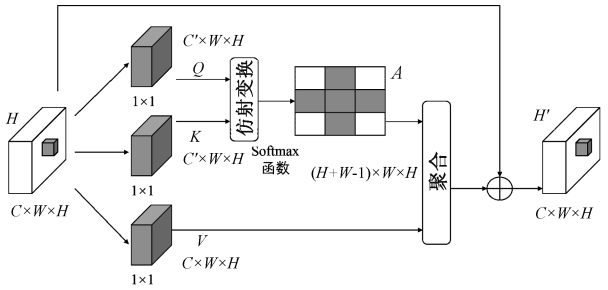

图 5 CC Attention 机制的架构

Figure 5 The architecture of CC Attention mechanism

3 实验

3.1 数据集

借助公开的 COCO 2017[12] 数据集来确定 CC-ResNest 基础网络的具体分组分块方案，并验证本文算法的场景普适性。此外，利用 CrowdPose[13] 数据集进行准确度测评，通过准确度和时间指标来体现本文算法在复杂场景下实现多人人体姿态估计的优势。CrowdPose 数据集中包含许多具有挑战性的复杂图像，符合复杂场景下多人人体姿态估计的数据集要求。

3.2 训练细节

在基础网络部分，首先使用 ImageNet 对模型进行预训练，其中 SGD 优化器的 Nesterov 动量为 0.95，批次为 32，权重衰减为 10^{-5}。学习率初始值为目标值的 10^{-3}，学习率每经过 10 个批次以指数形式衰减 1 次，衰减系数为 10。在优化步骤中更新模型参数的指数加权，衰减常数为 10^{-2}。

3.3 实验对比

在 COCO 2017 数据集上以 CC-ResNest 50 为基础网络训练 20 个批次，依据不同的分组分块方案进行消融实验，结果如表 1 所示。其中，4s2g 表示分为 2 组，组内分为 4 块。实验结果表明，分组数和组内分块数的增加可以提高姿态估计的准确率，同时增加时间和内存开销。综合考虑算法性能以及准确率和内存指标，将后续实验中 Split Attention 的参数设置为 2s2g。

利用 COCO 2017 数据集评估本文算法的性能，并与现有的多人人体姿态估计算法进行对比。不同算法在 COCO 2017 数据集上的实验结果如表 2 所示，其中 $AP^{0.50}$ 和 $AP^{0.75}$ 分别表示 IoU 阈值为 0.50 和 0.75 时的准确率，AP^{M} 和 AP^{L} 分别表示中等目标

表 1　Split Attention 在 COCO 2017 数据集上的消融实验结果

Table 1　The ablation experimental results of Split Attention on the COCO 2017 dataset

方案	时间复杂度	准确率/%	t/ms
1s2g	27.10	59.6	17.502
2s2g	27.11	61.5	17.725
4s2g	27.13	62.4	18.291
2s4g	27.24	62.0	18.087

和大目标对应的准确率。由表 2 可以看出,本文算法的平均准确率(AP)达到 72.6%,处理一幅图像的平均时间为 85 ms,帧数为 23 帧/s。准确度方面虽不及一些自顶向下的姿态估计算法,但均优于自底向上的姿态估计算法。其中,相比 OpenPifPaf++算法,本文算法的平均准确率提高 0.8 个百分点。表明本文算法具有不错的场景普适性,满足通用场景下姿态估计的任务需求。

表 2　不同算法在 COCO 2017 数据集上的实验结果

Table 2　Experimental results of different algorithms on the COCO 2017 dataset

算法	自顶向下方法的准确率/%					t/ms
	AP	$AP^{0.50}$	$AP^{0.75}$	AP^{M}	AP^{L}	
AlphaPose[14]	61.8	83.7	69.8	58.6	67.6	—
CPN[15]	72.1	90.5	78.9	67.9	79.1	—
HRNet[16]	75.5	92.5	83.3	71.9	81.5	—
DarkPose[17]	77.4	92.6	84.6	73.6	83.7	—
算法	自底向上方法的准确率/%					t/ms
	AP	$AP^{0.50}$	$AP^{0.75}$	AP^{M}	AP^{L}	
OpenPose[7]	61.8	84.9	67.5	57.1	68.2	100
PifPaf[8]	66.7	—	—	62.4	72.9	—
OpenPifPaf[10]	68.1	87.8	74.4	65.4	73.0	53
HigherHRNet[18]	70.5	89.3	77.2	66.6	75.8	>999
OpenPifPaf++[10]	71.8	89.4	78.1	68.5	77.4	81
本文	72.6	89.8	78.3	69.5	77.8	85

随后,针对算法中的各优化模块进行了消融实验研究,结果如表 3 所示。可以看出,相较于 CC Attention 机制,CCA 模块可以捕获更加丰富的空间语义特征。CCA 模块和 ResNest 伴随着额外的时间开销,提高了多人人体姿态估计算法在通用场景下的准确度。轻量化处理后的多人人体姿态估计算法虽牺牲少量准确度,但减少了 35%的额外时间开销,并且在某些准确率指标上(如 $AP^{0.50}$)还有细微的提升。这表明轻量化处理带来的梯度回溯在某些指标上出现了更加合适的参数选择,同时缓解了 CCA 模块过拟合图像空间特征信息的问题。

表 3　各优化模块在 COCO 2017 数据集上的消融实验结果

Table 3　The ablation experimental results of each optimization module on the COCO 2017 dataset

模型	准确率/%					t/ms
	AP	$AP^{0.50}$	$AP^{0.75}$	AP^{M}	AP^{L}	
基线	68.1	87.8	74.4	65.4	73.0	81
基线+CCA 模块	71.2	88.9	77.2	67.5	75.3	84
基线+CCA 模块+ResNest	72.7	89.4	78.8	69.7	78.2	87
最终模型	72.6	89.8	78.3	69.5	77.8	85

将本文算法与现有的多人人体姿态估计算法在 CrowdPose 数据集上进行定量对比,结果如表 4 所示。其中 AP、$AP^{0.50}$、$AP^{0.75}$ 指标与 COCO 2017 数据集中的含义相同,AP^{E}、AP^{M}、AP^{H} 指标分别表示算法在简单、中等难度、复杂场景下的准确率。从表 4 可以看出,本文算法的平均准确率达到 71.7%,相比 OpenPifPaf 算法提高 1.2 个百分点;在简单、中等、复杂场景下分别获得 78.6%、73.8%、65.3%的准确率。相比其他的多人人体姿态估计算法,本文算法在多个指标上均获得不同程度的提升。对于中等难度和复杂场景下的指标 AP^{M} 和 AP^{H},相比 OpenPifPaf 算法,分别提升了 1.7 个百分点和 1.5 个百分点,这是本文算法的主要贡献点。

表 4　不同算法在 CrowdPose 数据集上的实验结果

Table 4　Experimental results of different algorithms on the CrowdPose dataset　　单位:%

算法	自顶向下方法的准确率					
	AP	$AP^{0.50}$	$AP^{0.75}$	AP^{E}	AP^{M}	AP^{H}
AlphaPose[14]	61.0	81.3	66.0	71.2	61.4	51.1
AlphaPose++[14]	66.0	84.2	71.5	75.5	66.3	57.4
OPEC-Net[19]	70.6	86.8	75.6	—	—	—
算法	自底向上方法的准确率					
	AP	$AP^{0.50}$	$AP^{0.75}$	AP^{E}	AP^{M}	AP^{H}
OpenPose[7]	—	—	—	62.7	48.7	32.3
HigherHRNet[18]	67.6	87.4	72.6	75.8	68.1	58.9
OpenPifPaf[10]	70.5	89.1	76.1	78.4	72.1	63.8
本文	71.7	89.5	77.3	78.6	73.8	65.3

以 AlphaPose[14] 和 OpenPifPaf[10] 分别代表自顶向下和自底向上的方法,场景复杂化带来的准确率波动如表 5 所示。结果表明,场景复杂化致使 AlphaPose 和 OpenPifPaf 算法整体的平均准确率分别下降了 6.3%和 1.3%,并且对 AP^{M}、AP^{H} 指标的影响更大。实验结果证明,随着场景复杂度的提高,姿态估计算法的准确率也会出现明显降低。在复杂场景下,相较于自顶向下的算法,自底向上的算法具有更

加优异的鲁棒性和准确率,场景复杂化对算法准确度的干扰更小,更加适配于复杂场景下的多人人体姿态估计任务。

表5 场景复杂化带来的准确率波动

Table 5 Accuracy fluctuations caused by scene complexity

单位:%

算法	准确率			
	AP	AP^{E}	AP^{M}	AP^{H}
AlphaPose	-6.3	-4.3	-6.0	-14.9
OpenPifPaf	-1.3	-2.3	-4.1	-8.0

在CrowdPose数据集上对基础网络进行消融实验研究,重点关注AP、AP^{M}和AP^{H} 3个指标的数据表现。经过40个轮次训练后,ResNet不同变体网络在CrowdPose数据集上的定量评估结果如表6所示。相较于经典的ResNet变体网络,CC-ResNest表现出了更高的准确度,可以捕获更加丰富的特征信息,更加适配于复杂环境下的多人人体姿态估计任务。

表6 ResNet不同变体网络在CrowdPose数据集上的定量评估结果

Table 6 Quantitative evaluation results of different variant networks of ResNet on the CrowdPose dataset

单位:%

ResNet	准确率		
	AP	AP^{M}	AP^{H}
ResNet 50	58.7	58.5	53.3
ResNet 101	63.4	62.5	59.7
ResNext 50	59.6	60.7	58.4
SE-ResNet 50	60.5	58.8	54.7
ResNest 50	62.1	62.5	59.3
CC-ResNest 50	63.6	63.8	60.9

在COCO 2017和CrowdPose数据集上的实验结果表明,本文算法面向复杂场景和通用场景均表现出较高的准确度和场景普适性。在复杂场景下对不同的多人人体姿态估计算法进行可视化数据对比,结果表明,OpenPose算法可以快速有效地实现简单场景下的多人人体姿态估计,但在复杂场景下往往会产生许多冗余、错连、纠缠的关键点。OpenPifPaf算法可以很好地区分不同人体之间的关键点,与复杂场景下的多人人体姿态估计任务相适配,但在姿态估计关键点的完整度方面[20]还有提升的空间。本文算法在OpenPifPaf算法的基础上有了较为明显的提升,人体关键点的预测完整度平均提高10%,有效缓解了复杂场景对多人人体姿态估计算法的干扰。

4 结语

针对复杂场景下多人人体姿态估计存在骨架遮挡、丢失、纠缠的问题,本文提出了基于双注意力机制的多人人体姿态估计算法。设计Split Attention和CCA模块分别关注通道域和空间域的关键点信息,结合分组分块卷积,有效降低了复杂场景对多人人体姿态估计任务的干扰。在COCO 2017和CrowdPose数据集上的实验结果表明,所提算法在较小的时间开销下能够增强图像中语义关键点相关信息的获取能力,在复杂场景下对姿态估计的准确度有明显的提升。未来的工作将关注特定场景下基于姿态估计的异常行为识别研究,通过减缓背景噪声干扰以及利用图卷积网络挖掘异常行为识别相关的深层语义信息。

参考文献:

[1] 张国平, 马楠, 贯怀光, 等. 深度学习方法在二维人体姿态估计的研究进展[J]. 计算机科学, 2022, 49(12): 219-228.
ZHANG G P, MA N, GUAN H G, et al. Research progress of deep learning methods in two-dimensional human pose estimation[J]. Computer science, 2022, 49(12): 219-228.

[2] 褚真, 米庆, 马伟, 等. 部位级遮挡感知的人体姿态估计[J]. 计算机研究与发展, 2022, 59(12): 2760-2769.
CHU Z, MI Q, MA W, et al. Part-level occlusion-aware human pose estimation[J]. Journal of computer research and development, 2022, 59(12): 2760-2769.

[3] LIU W, BAO Q, SUN Y, et al. Recent advances of monocular 2D and 3D human pose estimation: a deep learning perspective[J]. ACM computing surveys, 2021, 55(4): 80.

[4] ZHANG H, WU C R, ZHANG Z Y, et al. ResNest: split-attention networks[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2022: 2735-2745.

[5] HUANG Z L, WANG X G, HUANG L C, et al. CCNet: criss-cross attention for semantic segmentation[C]// IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 603-612.

[6] PISHCHULIN L, INSAFUTDINOV E, TANG S Y, et al. DeepCut: joint subset partition and labeling for multi person pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 4929-4937.

[7] CAO Z, HIDALGO G, SIMON T, et al. OpenPose: real-time multi-person 2D pose estimation using part affinity fields[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(1): 172-186.

[8] KREISS S, BERTONI L, ALAHI A. PifPaf: composite fields for human pose estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2019: 11969-11978.

[9] PAPANDREOU G, ZHU T, CHEN L C, et al. PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]// European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 282-299.

[10] KREISS S, BERTONI L, ALAHI A. OpenPifPaf: composite fields for semantic keypoint detection and spatio-temporal association[J]. IEEE transactions on intelligent transportation systems, 2022, 23(8): 13498-13511.

[11] 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. Piscataway:IEEE Press, 2017: 5987-5995.

[12] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2014: 740-755.

[13] LI J F, WANG C, ZHU H, et al. CrowdPose: efficient crowded scenes pose estimation and a new benchmark[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 10855-10864.

[14] FANG H S, XIE S Q, TAI Y W, et al. RMPE: regional multi-person pose estimation[C]// IEEE International Conference on Computer Vision. Piscataway:IEEE Press, 2017: 2353-2362.

[15] CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7103-7112.

[16] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5686-5696.

[17] ZHANG F, ZHU X T, DAI H B, et al. Distribution-aware coordinate representation for human pose estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 7091-7100.

[18] CHENG B W, XIAO B, WANG J D, et al. HigherHRNet: scale-aware representation learning for bottom-up human pose estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 5385-5394.

[19] QIU L T, ZHANG X Y, LI Y R, et al. Peeking into occluded joints: a novel framework for crowd pose estimation[C]//European Conference on Computer Vision. Berlin: Springer Press , 2020: 488-504.

[20] 王珂, 陈启腾, 陈伟, 等. 基于深度学习的二维人体姿态估计综述[J]. 郑州大学学报(理学版), 2024, 56(4): 11-20.
WANG K, CHEN Q T, CHEN W, et al. Review of 2D human pose estimation based on deep learning[J]. Journal of Zhengzhou university (natural science edition), 2024, 56(4): 11-20.