基于图结构引导和位置信息强化的人体姿态估计

关 欣，周子健，李 锵

（天津大学 微电子学院，天津 300072）

摘 要：高自由度的肢体常构成各种复杂的姿态，极易产生关键点被遮挡的现象，定位遮挡关键点是人体姿态估计的难点之一，针对上述难点，提出了一种图结构引导并强化关键点位置信息的人体姿态估计方法。首先该方法在高分辨率网络中融入位置信息强化模块，用于提升可见关键点空间位置信息的表征精度。然后，在主干网络并行支路中引入视觉图神经模块，引导网络提取包含人体关键点的相关特征，在像素坐标空间中挖掘关键点之间局部和全局的拓扑连接关系，以便推测被遮挡关键点的位置信息。最后，结合关键点热图聚合单元和语义图卷积网络，在语义空间中更新各关键点间的亲和力权重，表示躯干结构约束下关键点之间的拓扑依赖关系，进一步优化被遮挡关键点的估计。本文模型在COCO2017测试集上的平均精度达到78.1%，能够精准估计复杂姿态中易被遮挡的关键点。

关键词：计算机视觉；人体姿态估计；关键点；图卷积

中图分类号：TP391.4 **文献标志码**：A **文章编号**：1671-5497(2025)10-3283-13

DOI：10.13229/j.cnki.jdxbgxb.20240086

Human pose estimation based on graph structure guidance and location information enhancement

GUAN Xin，ZHOU Zi-jian，LI Qiang

（*School of Microelectronics*，*Tianjin University*，*Tianjin* 300072，*China*）

Abstract：The high degree of freedom of human limbs often constitutes complex poses in which the key points are prone to occluded，and locating the occluded key points is one of the difficulties in human pose estimation. To this end，this paper proposed a method with a guided graph structure and enhanced key points location information. The method incorporates a location information enhancement module in the HRNet，which can improve the representation of the spatial location information of visible key points. A visual graph neural module is integrated into backbone network to extract relevant features containing key points and exploit the local and global topological connectivity relationships between key points in pixel coordinate space to infer the location information of the occluded key points. Finally，a heatmap

收稿日期：2024-01-23.

基金项目：国家自然科学基金项目(62071323)；超声医学工程国家重点实验室开放课题项目(2022KFKT004)；天津市自然科学基金项目(23JCZDJC00020).

作者简介：关欣(1977-)，女，副教授，博士. 研究方向：智能信息处理. E-mail：guanxin@tju.edu.cn

通信作者：李锵(1974-)，男，教授，博士. 研究方向：智能信息处理. E-mail：liqiang@tju.edu.cn

aggregation unit and a semantic graph convolutional network are employed to update the affinity weights between key points in the semantic space, which can represent the topological dependencies between key points under the constraints of the skeleton structure and further optimize the estimation of the occluded key points. The proposed model achieves an average accuracy of 78.1% on the COCO2017 test set, and can accurately estimate the occluded key points prone to occlusion in complex poses.

Key words: computer vision; human pose estimation; key points; graph convolution

0 引 言

二维人体姿态估计是计算机视觉领域热门且极具挑战的研究方向之一,它在人机交互、动作识别、智能医疗、虚拟现实等领域有着广泛的应用[1,2]。受人体肢体结构的复杂性、各类关键点的灵活差异性,以及不同程度环境干扰等因素影响,关键点易被遮挡,定位复杂姿态中易被遮挡的关键点是人体姿态估计的难点之一。例如,基于视觉的乐器演奏姿势校准中常存在乐器遮挡关键点的问题;分析运动员动作姿态时,关键点自遮挡会对数据采集产生干扰。这些情况均对模型预测遮挡关键点的精度提出了更高的要求。

得益于深度学习的迅速发展,基于卷积神经网络(Convolutional neural network, CNN)[3]的人体姿态估计领域也取得了巨大进展。早期工作[4]结合CNN,直接从图像中回归关键点的二维坐标,通过全连接网络将编码后的关键点信息直接转换为对应坐标值。然而,这种单一的坐标值转换在训练过程中缺乏空间位置信息的监督,导致模型空间泛化能力差,在遮挡情形下预测精度不高。不同于直接回归坐标的方式,Tompson等[5]最早提出基于热图预测的关键点建模方式,将坐标真值转换为以关键点二维坐标为中心的高斯热图,其中的每个像素表示所属关键点的置信度得分。热图通过在训练阶段对关键点标签施加软约束,提供空间维度更多的监督信息,进而提高了模型的空间泛化能力。在部分基于热图预测的CNN模型中,如堆叠沙漏网络[6]、级联金字塔网络(Cascaded pyramid network, CPN)[7]、简单基线网络[8]以及高分辨率网络(High-resolution network, HRNet)[9],均旨在通过卷积获得网络不同阶段多尺度特征,融合空间位置信息和高级语义信息来实现更精准的热图位置估计。然而,这些方法聚焦于关键点多尺度特征的提取和利用,无法自适应地提取包含人体结构的关键点特征,且未进一步挖掘热图中包含的关键点隐式联系。

近年来,结合Transformer[9]的姿态估计模型发展迅速。基于视觉的Transformer[10]姿态估计方法将人体图像切块并转换为图像块序列,通过序列式建模引入全局敏感性和长距离依赖性,进而学习关键点之间复杂的分布联系。例如,TokenPose[11]作为该领域典型代表,通过随机初始化的方式建立关键点与不同图像块之间的相关性,从而学习可靠的关键点表征。HRFormer[12]采用多尺度特征并行策略,通过局部窗口自注意力机制实现不同图像块之间的信息交互。基于Transformer的姿态估计方法侧重于建立关键点与不同图像区域之间的相关性,但忽略了特征提取阶段对人体拓扑结构特征的表征。由于人体的骨骼结构对关键点存在约束,无论是不同尺度特征的提取还是网络输出关键点的热图阶段,关键点之间均不是孤立存在的。如何自适应地表征不同感受野下关键点的拓扑依赖关系,对应对各种遮挡时的姿态估计任务至关重要[13]。

为此,本文提出图结构引导和位置信息强化的人体姿态估计,既考虑了人体关键点分布与不同图像区域之间的相关性,又关注了关键点热图特征之间的连接依赖关系,多方面丰富了关键点的特征表示。首先,采用HRNet作为主干网络提取多尺度特征,并融入位置信息强化模块(Location information enhancement module, LIEM),通过在横、纵坐标轴方向对关键点位置特征进行独立编码,实现嵌入位置信息的高精度关键点表征。其次,在主干网络不同分辨率支路中引入视觉图神经模块(Vision graph neural module, VGNM),在特征提取阶段以特征图块为节点,结合图卷积算法拟合人体关键点分布的拓扑关系,引导网络在不同感受野下自适应地提取包含人体关键点的相关特征。然后,在热图聚合单元(Heatmap aggregation unit, HAU)中进一步挖掘热图特征信息并学习关键点热图间的隐式关系,然后将其中

的复合热图特征进行维度转换得到关键点特征向量。最后,采用语义图卷积网络(Semantic graph convolutional network, SGCN)更新关键点特征向量的亲和力权重,结合邻接矩阵强化受躯干约束下关键点之间的拓扑依赖关系,优化遮挡关键点位置估计并应对关键点错连问题,最终实现精准高效的复杂姿态估计。

1　相关工作

在当前人体姿态估计领域中,基于热图预测的CNN方法和基于Transformer架构的方法备受关注。一方面,研究人员设计了众多CNN架构的姿态估计模型。例如,Newell等[6]设计了由多个沙漏模块堆叠而成的Hourglass网络,迭代提取人体图像的多尺度特征。但随着沙漏模块堆叠数量的增加,模型精度提升有限。Chen等[7]提出了一个两阶段CPN,在融合不同尺度特征的同时,分别检测容易和困难两种类型的关键点,在预测遮挡关键点时具有较好的表现。Xiao等[8]仅使用单阶段编解码网络Simple Baseline,对深度低分辨率特征进行反卷积上采样获得关键点热图,网络构造简单但参数量较大,计算效率有待提升。上述多个基于CNN的模型为了获取关键点高级的语义信息,需要连续多次下采样提取特征,降低特征图尺寸的同时会造成关键点空间位置信息遗失。为此,Sun等[9]提出HRNet,通过高分辨率支路来保留原始图像中的空间位置信息,同时采用多分辨率特征并行和融合策略增强关键点的表征能力。另一方面,在基于Transformer架构的方法中,Yang等[14]提出Transpose架构,将卷积提取的特征图分割为图像块并展平为一维向量,通过自注意力机制捕捉人体各部位特征之间的全局依赖关系。Li等[12]设计的TokenPose架构同样将特征图转换为一维向量,但额外引入了随机初始化得到的各个关键点特征,将一维向量和关键点特征共同输入自注意力单元,从而学习关键点特征与一维向量所对应不同图像块之间的映射关系。基于Transformer的架构多以高级语义特征作为输入,自注意力机制虽侧重于学习高级特征与关键点之间的对应关系,但在低级特征的提取阶段,无法自适应提取与人体结构相关的特征表示,进而难以表征复杂的关键点拓扑结构。

人体关键点的结构分布可以看作以节点和边构成的图数据来处理,结合图卷积网络(Graph convolutional network, GCN)[15]表征人体关键点之间的拓扑关系,从而使模型更灵活地捕获骨架约束下的关键点分布关系。例如,Qiu等[16]提出基于特征图引导的渐进式GCN,利用骨架姿态结构信息和图像上下文特征,实现对二维人体遮挡关键点的预测校准。Bin等[17]设计平行双支路GCN结构,通过双支路的特征交互学习局部关键点的特征相关性,同时利用长程关系估计较难定位的关键点。Wang等[18]对网络预测热图进行采样,通过采样点引导特征优化定位结果,同时结合关键点之间的拓扑关系来细化姿态回归特征,从而提高姿态估计的性能。Banik等[19]以邻接矩阵的形式编码人体骨骼的结构信息,利用GCN更新关键点的拓扑连接关系,进而回归预测出三维人体坐标。然而,上述图卷积方法在热图表征后进一步约束关键点的分布,忽略了特征提取阶段图卷积对人体结构建模的重要性。由于人体关键点组成的骨架模型是一种铰链式结构,邻近关键点之间具有更强的相关性,需要在特征提取阶段引导网络自适应地学习与人体结构相关的特征,同时在热图表征后进一步建立语义特征在躯干链式结构约束下的关键点拓扑关系。

2　本文方法

2.1　整体网络框架

本文提出的姿态估计网络整体架构如图1所示。以HRNet为主干网络,网络整体分为4个阶段,每个阶段先通过下采样产生一个新的低分辨率支路,新支路分辨率减半的同时通道数翻倍;网络采用多分辨率并行机制,同时结合上、下采样和融合机制实现不同分辨率特征的信息交互。为了充分利用多尺度并行特征,本文在主干网络中融入LIEM,强化坐标轴方向位置信息的提取,以实现可见关键点高精度的位置表征。同时,在每个并行支路开始阶段引入VGNM,引导网络在不同感受野下学习与人体躯干相关的特征,进一步在拓扑空间推测被遮挡关键点的表示。考虑到最高分辨率支路包含丰富的空间位置信息,本文在HAU中提取不同阶段高分辨率支路的特征,将其与输出的热图特征拼接后进行特征提取和维度转换,最后在SGCN中对关键点特征向量迭代更新,学习受躯干约束下关键点间的拓扑连接关系,优化

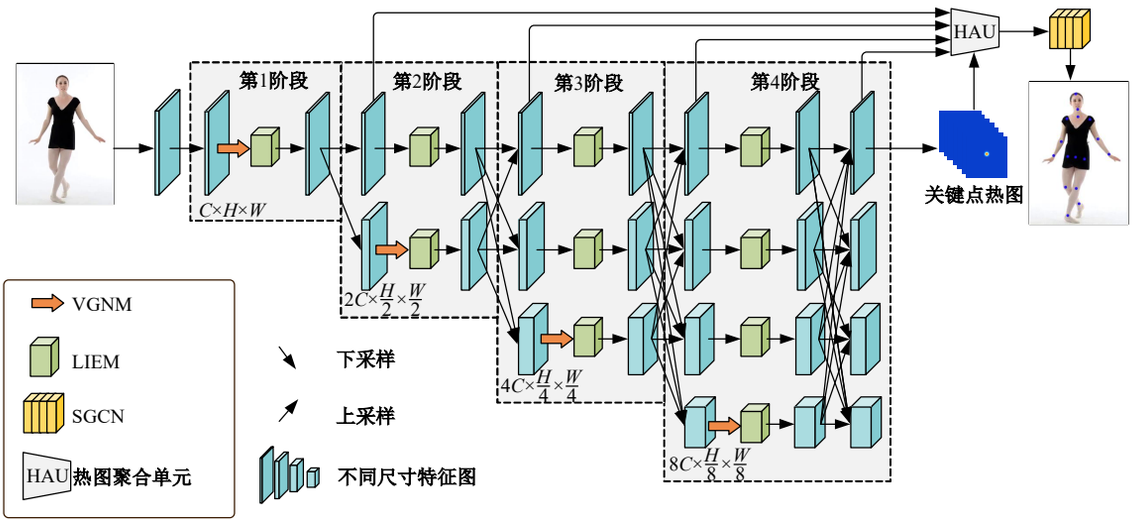

图1 姿态估计网络整体架构

Fig. 1 Overall architecture of pose estimation networks

被遮挡关键点的位置估计。

2.2 LIEM

为了提升可见关键点坐标位置信息的表征精度,受坐标注意力[20]中独立方向编码思想的启发,本文提出LIEM。如图2所示,将输入特征沿两个垂直的空间坐标方向分别进行一维特征编码:沿一个空间方向捕捉关键点特征长距离的依赖性,以学习该方向关键点位置的内在联系;沿另一个空间方向保留关键点精确的位置信息,缓解二维全局池化导致的关键点位置偏移问题。最终形成一对方向感知和位置敏感的特征图,以增强通道中位置信息的表征。对编码后的特征沿空间维度分别进行全局最大池化和全局平均池化,以进一步更新、筛选包含高精度位置表征的通道特征,并通过多层感知机(Multi-layer perception, MLP)实现通道维度缩放生成对应的权重信息。

具体来说,将输入特征$\boldsymbol{X}_{\text{in}}\in\mathbb{R}^{C\times H\times W}$逐通道沿两个垂直的空间坐标方向分别进行全局自适应平均池化[20],得到一对具备空间方向感知能力的一维向量。对上述两个方向特征沿空间维度进行拼接操作,再通过1×1卷积运算实现对通道维度的缩放变换,具体可表示为:

$$\boldsymbol{I}=\delta_{\text{h}}(f_{1\times1}([\boldsymbol{U}^{\text{h}},\boldsymbol{V}^{\text{w}}])) \tag{1}$$

式中:$\boldsymbol{U}^{\text{h}}\in\mathbb{R}^{C\times H\times1}$和$\boldsymbol{V}^{\text{w}}\in\mathbb{R}^{C\times1\times W}$分别为全局平均池化后的一维特征向量;[•,•]为空间维度拼接操作;$f_{1\times1}(\cdot)$为1×1卷积运算;$\delta_{\text{h}}(\cdot)$为h_swish激活函数和批归一化;$\boldsymbol{I}$为输出的中间特征。

对$\boldsymbol{I}$沿不同空间方向分离为独立的特征向量$\boldsymbol{I}^{\text{h}}$和$\boldsymbol{I}^{\text{w}}$,再分别经过对应的1×1卷积恢复到与输入特征通道一致,并利用sigmoid函数生成注意力权重,可分别表示为式(2)和式(3)。

$$\boldsymbol{g}^{\text{h}}=\sigma(f_{\text{h}}(\boldsymbol{I}^{\text{h}})) \tag{2}$$

$$\boldsymbol{g}^{\text{w}}=\sigma(f_{\text{w}}(\boldsymbol{I}^{\text{w}})) \tag{3}$$

式中:$f_{\text{h}}(\cdot)$和$f_{\text{w}}(\cdot)$分别为不同方向的1×1卷积运算;$\sigma(\cdot)$为sigmoid函数;$\boldsymbol{g}^{\text{h}}$和$\boldsymbol{g}^{\text{w}}$分别为不同方向更新后的权重。不同方向权重更新可以表示为:

$$\boldsymbol{X}_{\text{m}}=\boldsymbol{X}_{\text{in}}\times\boldsymbol{g}^{\text{h}}\times\boldsymbol{g}^{\text{w}} \tag{4}$$

式中:$\boldsymbol{X}_{\text{m}}\in\mathbb{R}^{C\times H\times W}$为中间特征图。对位置编码后的特征分别再沿空间维度进行池化聚合特征处理,并通过MLP进行通道尺度缩放,最后利用sigmoid函数生成通道权重$\boldsymbol{g}^{\text{c}}$,如式(5)所示:

$$\boldsymbol{g}^{\text{c}}=\sigma(W(\text{AvgPool}(\boldsymbol{X}_{\text{m}}))+W(\text{MaxPool}(\boldsymbol{X}_{\text{m}}))) \tag{5}$$

式中:AvgPool(•)为全局平均池化;MaxPool(•)为全局最大池化;W为MLP的共享参数;$\sigma(\cdot)$为sigmoid函数。LIEM最终的输出表示为:

$$\boldsymbol{X}_{\text{out}}=\boldsymbol{g}^{\text{c}}\times\boldsymbol{X}_{\text{m}} \tag{6}$$

式中:$\boldsymbol{g}^{\text{c}}$为上一阶段生成的通道权重;$\boldsymbol{X}_{\text{out}}\in\mathbb{R}^{C\times H\times W}$为最终的输出特征。

2.3 VGNM

为了充分利用图卷积算法拟合人体关键点拓扑结构的能力,同时结合多尺度特征的空间信息推测不可见的遮挡关键点,本文在主干网络每个并行支路引入VGNM[21]。如图3所示,该模块由图卷积单元和前向传播单元两个部分构成。

图卷积单元作为核心组成部分,通过全连接层将输入特征进行维度转换形成N个特征向量

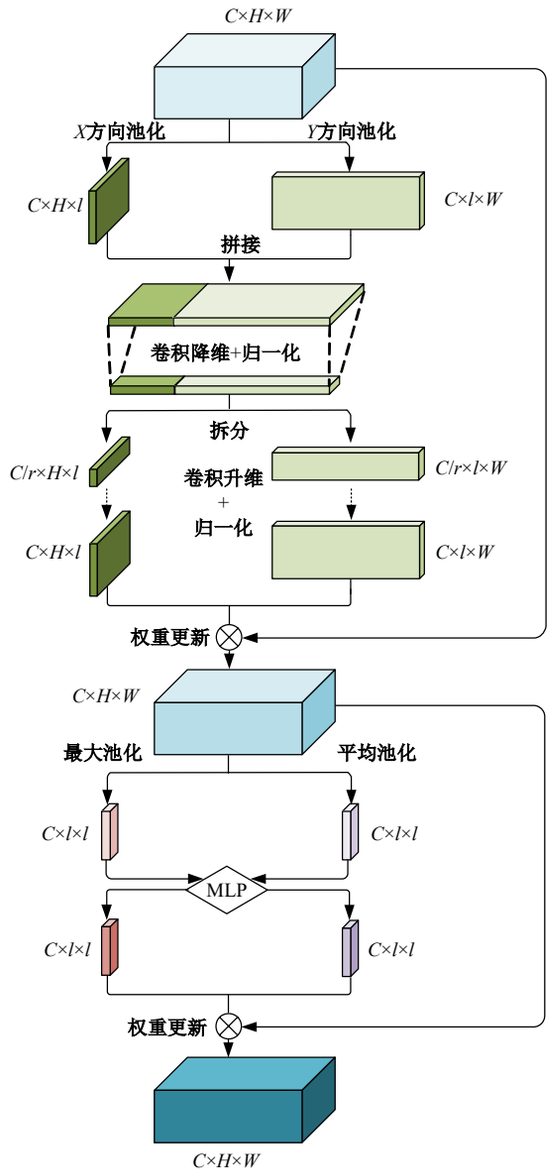

图 2　LIEM 结构

Fig. 2　Structure of LIEM

$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$,其中 $\boldsymbol{x}_i\in\mathbb{R}^D$,$D$ 为向量的特征维度。VGNM 将特征向量视为一系列无向节点 $\boldsymbol{\Lambda}=\{\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2,\cdots,\boldsymbol{\lambda}_N\}$。对于每个节点 $\boldsymbol{\lambda}_i$,结合 K 近邻算法捕获与每个节点具备相似特征的邻居节点 $N(\boldsymbol{\lambda}_i)$,并为每个邻居节点 $\boldsymbol{\lambda}_j\in N(\boldsymbol{\lambda}_i)$ 添加从 $\boldsymbol{\lambda}_j$ 到 $\boldsymbol{\lambda}_i$ 的边 $\boldsymbol{e}_{ji}$,组成边的集合 ε。通过构建特征的节点和边的表示,可以得到包含人体特征的拓扑图结构 $\boldsymbol{\zeta}=(\boldsymbol{\Lambda},\varepsilon)$,从而实现从特征向量 $\boldsymbol{X}\in\mathbb{R}^{N\times D}$ 到图结构 $\boldsymbol{\zeta}=G(\boldsymbol{X})$ 的转变。对于构建得到的图结构,图卷积层通过聚合函数汇聚邻近节点的信息计算节点的特征表示,并进一步更新汇聚后的节点特征。聚合和更新后的节点 $\boldsymbol{\lambda}_i'$ 可以表示为:

$$\boldsymbol{\lambda}_i'=h(\boldsymbol{\lambda}_i,g(\boldsymbol{\lambda}_i,N(\boldsymbol{\lambda}_i);\ \boldsymbol{W}_{\text{aggregate}});\ \boldsymbol{W}_{\text{update}}) \tag{7}$$

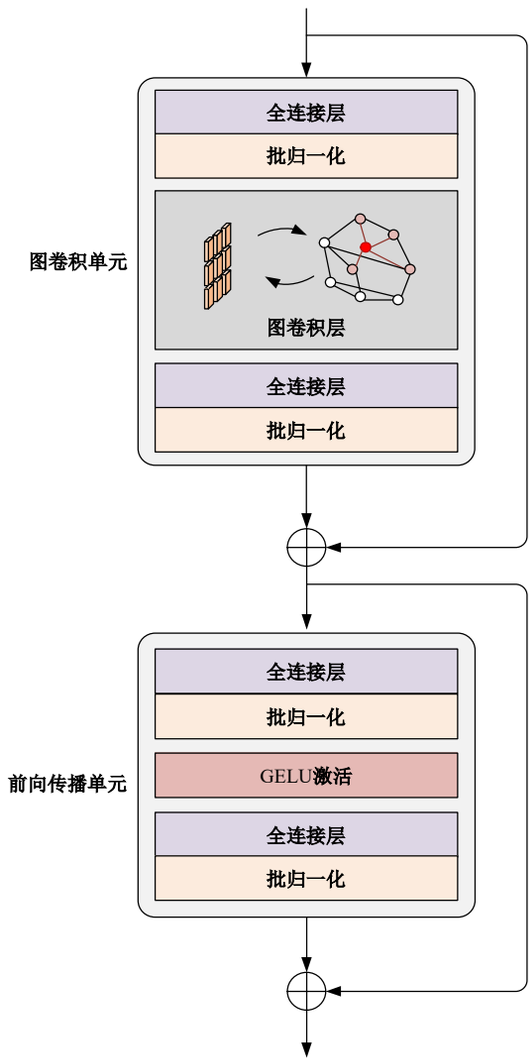

图 3　VGNM 结构

Fig. 3　Structure of VGNM

式中:$\boldsymbol{W}_{\text{aggregate}}$ 和 $\boldsymbol{W}_{\text{update}}$ 分别为聚合和更新节点特征时可学习的权重参数,聚合和更新函数分别表示为式(8)和式(9)。

$$g(\cdot)=\boldsymbol{\lambda}_i''=[\boldsymbol{\lambda}_i',\max(\{\boldsymbol{\lambda}_j-\boldsymbol{\lambda}_i|j\in N(\boldsymbol{\lambda}_i)\})] \tag{8}$$

$$h(\cdot)=\boldsymbol{\lambda}_i'=\boldsymbol{W}_{\text{update}}\boldsymbol{\lambda}_i''+\boldsymbol{b}_{\text{h}} \tag{9}$$

式中:$g(\cdot)$ 为聚合节点的最大相对图卷积算法[15];$h(\cdot)$ 为节点特征的更新函数;$\boldsymbol{b}_{\text{h}}$ 为对应更新偏置。综上所述,图卷积单元可以表示为:

$$\boldsymbol{Y}=\text{GraphConv}(\boldsymbol{X}\boldsymbol{W}_{\text{in}})\boldsymbol{W}_{\text{out}}+\boldsymbol{X} \tag{10}$$

式中:$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别为图卷积单元的输入和输出;GraphConv(·)为图卷积层,其前后的全连接层用于将节点特征映射到同一特征域并增加特征的多样性:$\boldsymbol{W}_{\text{in}}$ 和 $\boldsymbol{W}_{\text{out}}$ 为全连接层中可学习的权重参数。

前向传播单元由相应的全连接层、批归一化

和GELU激活函数组成,可以表示为:

$$Z=\delta_g(YW_1)W_2+Y \tag{11}$$

式中:Y和Z分别为前向传播单元的输入和输出,其中包含了残差连接结构,全连接层用于进一步提升节点特征的转换能力;W_1和W_2为全连接层中可学习的权重系数;$\delta_g(\cdot)$为GELU非线性激活函数,公式中省略了批归一化。

2.4 HAU

HAU结构如图4所示。由于主干网络的最高分辨率支路包含原始图像丰富的位置信息,并融合了各个阶段其他低分辨支路多尺度的关键点特征信息。因此,本文结合特征复用[22]的理念,充分挖掘网络不同深度层的关键点特征。通过抽取最高分辨率支路每个阶段多尺度融合后的特征,将其与关键点热图在通道维度拼接得到复合热图特征,作为HAU的输入。

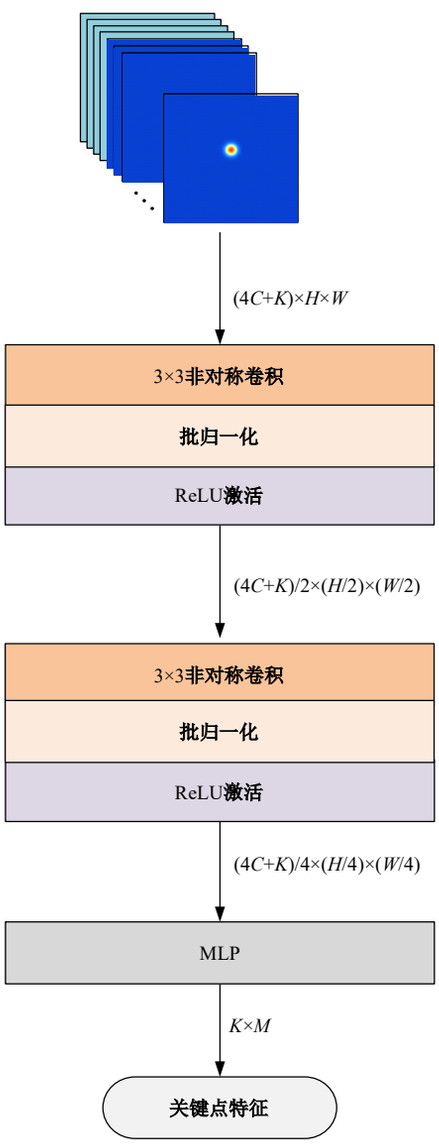

图4 HAU结构

Fig. 4 Structure of HAU

研究表明,卷积核中不同位置参数的重要程度不同,处于中心交叉位置即骨架方向上的权重参数贡献度更高,而角点位置的参数影响较小[23]。为了强化模型在语义阶段的表征能力,本文在HAU中主要通过两个非对称卷积[23]下采样提取输入的复合热图特征。非对称卷积结构(见图5)在训练阶段由1×3、3×1和3×3这3组不同类型的卷积核并行提取特征,同时更新对应卷积核权重。在推理时采用上述3组卷积核叠加融合而成的复合卷积核,可以缓解下采样导致的关键点空间位置信息遗失问题。下采样后的热图特征通过MLP进行特征维度变换,得到K个M维的关键点特征$J\in\mathbb{R}^{K\times M}$,可以表示为:

$$J=\mathrm{MLP}(f_{ac}(f_{ac}([X_1,X_2,X_3,X_4,X_h]))) \tag{12}$$

式中:$X_1,X_2,X_3,X_4\in\mathbb{R}^{C\times H\times W}$分别为最高分辨支路不同阶段的特征图;$X_h\in\mathbb{R}^{K\times H\times W}$为主干网络输出的$K$张尺寸为$H\times W$的关键点热图,每张热图包含对应关键点特征的位置信息和语义信息;

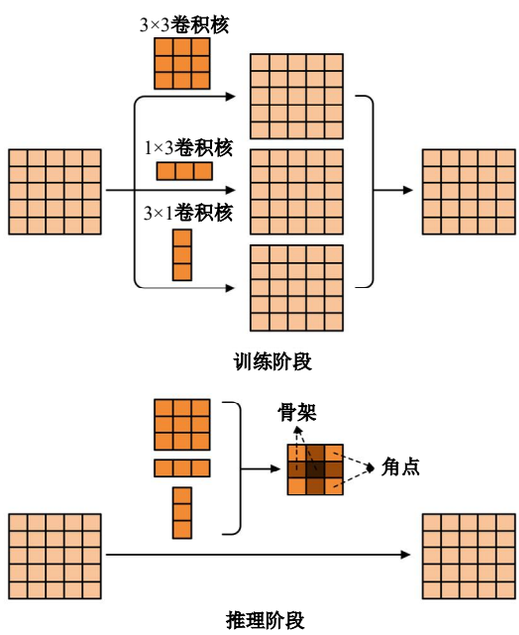

图5 非对称卷积结构

Fig. 5 Structure of asymmetric convolution

$[\cdot,\cdot]$为通道维度拼接操作;$f_{ac}(\cdot)$为步长为2、填充为1的非对称卷积下采样,公式中省略了批归一化和ReLU激活函数;MLP$(\cdot)$为MLP函数,用于将下采样后的特征图转换为关键点特征向量

$\boldsymbol{J}\in\mathbb{R}^{K\times M}$,作为后续语义图卷积网络的输入。

2.5　SGCN

为了进一步强化躯干约束下关键点之间的拓扑依赖关系,优化被遮挡关键点的精确位置估计,本文采用SGCN[24]学习关键点拓扑结构中的显示关系,同时捕获关键点特征中隐式的非物理连接,更新HAU提取到的关键点特征向量$\boldsymbol{J}\in\mathbb{R}^{K\times M}$。SGCN结构如图6所示,主体包含4个重复的残差连接单元,每个单元主要由两个语义图卷积层、批归一化和ReLU非线性激活函数组成,并结合非局部层[25]进一步扩大感受野,捕捉长距离关键点间的隐式联系。先前提出的GCN网络,如aGCN[26]和GAT[27]以邻居节点来计算整个图的隐藏表示,通过不同输入边的权重调节整个拓扑图中的信息流。相比之下,SGCN可以动态调整全局拓扑图中每个独立边的输入权重,这些权重代表了人体骨架图结构中隐含的先验信息,用于学习可见关键点如何隐式影响其他被遮挡关键点。语义图卷积的公式为:

$$\boldsymbol{J}'=\mathop{/\!/}_{m=1}^{M}\delta_{\mathrm{r}}(\rho(\boldsymbol{M}_m\odot\boldsymbol{A})\boldsymbol{\omega}_m\boldsymbol{J}) \tag{13}$$

式中:邻接矩阵$\boldsymbol{A}\in\mathbb{R}^{K\times K}$用于表征关键点特征的物理连接关系;$\boldsymbol{M}_m$为不同通道对应可学习的权重矩阵,$\odot$操作根据邻接矩阵激活$\boldsymbol{M}_m$相应位置的权重参数;$\rho(\cdot)$为Softmax非线性函数;$\boldsymbol{\omega}_m$为对应权重矩阵的第$m$行;$\delta_r(\cdot)$为ReLU非线性激活函数;$/\!/$为通道维度级联操作;$\boldsymbol{J}'\in\mathbb{R}^{K\times M}$为经过SGCN迭代更新后输出的关键点特征向量。结合躯干信息在语义空间进一步更新关键点向量,以实现对遮挡关键点的优化校准。

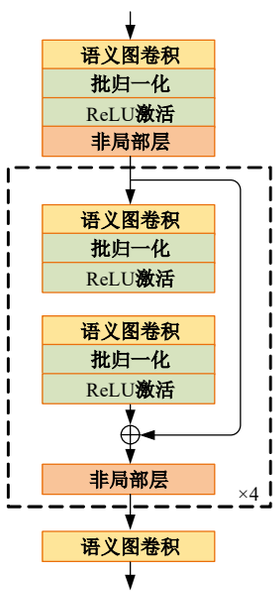

图6　SGCN结构

Fig. 6　Structure of SGCN

2.6　损失函数

本文采用联合均方误差损失函数L进行计算,其由热图损失函数L_{h}和坐标损失函数L_{l}构成。其中,L_{h}用于计算关键点预测热图和目标热图之间的损失,定义为:

$$L_{\mathrm{h}}=\sum_{k=1}^{K}\|h_k-\hat{h}_k\|^2 \tag{14}$$

式中:h_k为网络预测的第k个关键点热图;$\hat{h}_k$为第k个关键点真值热图。

对于网络最终回归得到的关键点坐标位置,采用L_{l}计算关键点预测坐标和真值坐标之间的损失,L_{l}定义为:

$$L_{\mathrm{l}}=\sum_{k=1}^{K}\|l_k-\hat{l}_k\|^2 \tag{15}$$

式中:l_k为网络预测的第k个关键点位置坐标;$\hat{l}_k$为第k个关键点真值坐标。

综上所述,整个网络的总体损失函数L可以表示为:

$$L=L_{\mathrm{h}}+L_{\mathrm{l}} \tag{16}$$

3　实验验证

3.1　数据集与评价指标

将本文模型分别在人体姿态估计公开数据集MPII[28]和COCO2017[29]上进行验证和测试。

MPII数据集是人体姿态估计领域的经典数据集之一,图像提取自流媒体视频,包含大量日常活动的真实场景和丰富的背景信息,其中约有25 000张图像和40 000多个注释信息。每个人体实例标签包含16个关键点的信息标注。数据集中约28 000个样本用于训练,11 000个样本用于测试。

正确关键点百分比(Percentage of correct keypoints,PCK)作为MPII评估模型关键点估计准确度的标准,统计了被准确检测的关键点所占比例。目前,姿态估计领域普遍采用头部尺寸因子的50%即PCKh@0.5作为归一化指标,并通过设定不同的阈值计算最终的PCK平均值,用于统计正确预测的关键点占总数的比例。PCK平均值的计算公式为:

$$\mathrm{PCK}_{\mathrm{mean}}=\frac{\sum_p\sum_i\delta\left(\frac{d_{pi}}{d_p^{\mathrm{def}}}\leqslant T_n\right)}{\sum_p\sum_i 1} \tag{17}$$

式中:d_{pi}为第p个行人中序号为i的关键点预测值与标注真值的欧氏距离;d_p^{def}为用于归一化的第p个行人的头部尺度因子;T_n为人工设定的第n个阈值,$T_n\in[0:0.01:0.1]$;$\mathrm{PCK}_{\mathrm{mean}}$为不同$T_n$阈值下的平均值;$\delta(\cdot)$为狄拉克函数,满足条件时为1,用于计算符合标准的关键点数量。

COCO2017数据集是由微软提供的人体姿态估计大型数据集,包含约220 000张标注图像和约250 000个标注的人体实例,每个人体有17个关键点信息标注。其中,训练集train2017包含57 000张图像,验证集val2017包含5 000张图像,测试集test-dev2017包含20 000张图像。本实验采用train2017进行训练,在val2017上验证和评估模型,并给出test-dev2017的测试结果。

目标关键点相似度(Object keypoint similarity,OKS)用于衡量预测关键点与标注真值之间的相似性,计算公式为:

$$\mathrm{OKS}=\frac{\sum_i\exp\left[\frac{-d_i^2}{2s^2k_i^2}\right]\delta(v_i>0)}{\sum_i\delta(v_i>0)} \tag{18}$$

式中:d_i为关键点i的预测值与标注真值之间的欧氏距离;s为尺度因子,其值为人体检测框面积的平方根;k_i为第i个关键点的归一化因子;$v_i\in(0,1,2)$为第i个关键点的可见性,其中:0为未标记的关键点,1为无遮挡的标记关键点,2为有遮挡的标记关键点;$\delta(\cdot)$为狄拉克函数,满足条件时为1,此处指仅计算已标注的关键点。

给定阈值T_s,当$\mathrm{OKS}>T_s$表示该阈值下模型预测结果正确。平均精度(Average precision,AP)作为COCO2017数据集上的评价指标,是指在不同OKS阈值$T_s\in[0.5:0.05:0.95]$下取得的平均值;$\mathrm{AP}^{0.5}$和$\mathrm{AP}^{0.75}$分别指T_s为0.5和0.75时对应的精度:AP^{M}、AP^{L}是根据数据集中人体检测框面积设定的指标,AP^{M}用于面积为$32^2\sim96^2$,AP^{L}用于面积$>96^2$。平均召回率(Average recall,AR)为辅助指标,计算过程和AP类似。

3.2 实验设置

本文实验使用的操作系统为64位Ubuntu 16.04,服务器配置包括Intel CPU Core i9-9900X @ 3.5GHz、GPU Nvidia RTX2080Ti(11 GB)×4,采用PyTorch深度学习框架。

实验预处理阶段以4:3的高宽比设定人体检测框,从输入中裁剪出人体图像后进行对应尺寸调整。MPII数据集尺寸调整为256×256,COCO2017数据集尺寸调整为256×192和384×288。在数据增强中,MPII数据集包括随机左右翻转、随机旋转±30°和随机缩放±0.25比例;COCO2017数据集包括随机左右翻转、随机旋转±45°和随机缩放±0.35比例。主干网络HRNet共有W32和W48两个版本,数字代表最大分辨率支路的通道数。HAU输出的关键点特征维度M设定为128,关键点个数K根据不同数据集设定为16或17。训练过程使用Adam优化器,模型的基础学习率为0.001,分别在第170和200轮时以0.1的比例降低学习率,模型一共训练230轮。

3.3 结果分析

为了验证本文方法的有效性,本节在不同数据集上通过消融实验分析各个模块的性能,并将本文网络与人体姿态估计的经典主流网络进行对比,最后对估计结果进行可视化。

3.3.1 消融实验

本节通过一系列消融实验对本文网络模型的有效性进行验证。以HRNetW32版本为基础模型Baseline进行实验分析。将所提出的LIEM、VGNM、SGCN和HAU依次引入基础模型,分别在MPII和COCO2017数据集下对比不同网络配置的估计结果,对比结果如表1和表2所示。

在MPII数据集中以PCK平均值作为评价指标。由表1可知,在Baseline的基础上引入LIEM,模型在头部、肩膀和臀部关键点的预测结果有所提升,除脚踝精度略有下降外,其他关键点精度持平,模型精度平均值维持在90.3%。这表明坐标轴方向独立编码强化了躯干主体关键点的定位能力,但仅在坐标轴方向强化关键点特征的表征精度,还不足以定位躯干末端较远的脚踝关键点。此外,MPII数据集样本量较小,融入LIEM后的模型在小数据集上效果不明显。进一步引入VGNM后,所有关键点精度均有不同幅度提升,其中肘部和膝盖提升较为明显,模型精度平均值提升至90.4%,表明空间维度拓扑结构学习可以固定躯干近端易被遮挡的关键点。在此基础上,采用MLP对网络输出的复合热图进行维度转换并引入SGCN,精度平均值达到90.5%,其中头部、肘部和脚踝的预测精度得到提升,模型初步

表 1　在 MPII 验证集上 PCKh 阈值为 0.5 时不同网络配置的实验结果

Table 1　Experimental results for different network configurations with a PCKh threshold of 0.5 on MPII validation set

%

Baseline	LIEM	VGNM	SGCN	HAU&SGCN	头部	肩部	肘部	腕部	臀部	膝盖	脚踝	平均值
√					97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
√	√				97.3	96.0	90.3	86.4	89.3	87.1	83.2	90.3
√	√	√			97.4	96.1	90.7	86.5	89.4	87.4	83.4	90.4
√	√	√	√		97.6	96.1	90.8	86.5	89.4	87.4	83.5	90.5
√	√	√		√	97.6	96.3	90.8	86.7	89.5	87.4	83.6	90.6

表 2 在 COCO2017 验证集上不同网络配置的实验结果

Table 2　Experimental results for different network configurations on COCO2017 validation set

Baseline	LIEM	VGNM	SGCN	HAU&SGCN	参数量/M	运算量/G	AP/%	$AP^{0.5}$/%	$AP^{0.75}$/%	AP^{M}/%	AP^{L}/%	AR/%
√					28.5	7.1	74.4	90.5	81.9	70.8	81.0	79.8
√	√				28.7	7.2	74.9	91.1	82.0	71.6	81.4	79.9
√	√	√			29.4	7.5	75.6	91.8	82.3	72.2	81.8	80.2
√	√	√	√		30.1	7.8	76.2	92.5	82.7	72.4	82.0	80.6
√	√	√		√	30.3	7.9	76.5	92.7	82.8	72.5	82.3	80.8

强化了语义空间中关键点的拓扑分布关系。进一步引入结合 HAU 的 SGCN(HAU&SGCN)后,关键点的精度平均值达到 90.6%,其中肩部、腕部、臀部和脚踝得到提升。这表明 HAU 可以保留原始图像的空间分布并挖掘热图关键点间的分布联系,同时结合 SGCN 在语义空间学习躯干约束下的关键点拓扑连接关系,进一步在拓扑空间优化了躯干远端易被遮挡的关键点。

在 COCO2017 数据集中以 AP 为主要评价指标。由表 2 可知,本文模型在 COCO2017 数据集上的提升更为显著。引入 LIEM 后不同尺度下精度均有所提升,模型 AP 值提升 0.5%,网络在不同感受野下记录了可见关键点位置的高精度表征,坐标位置信息得以保留,同时强化了长距离依赖关系的建立。在此基础上增加 VGNM 后,模型强化了在不同分辨率空间下捕获与人体结构相关特征的能力,在拓扑空间实现对不可见关键点的初步预测,AP 值进一步提升了 0.7%。采用 MLP 对复合热图进行维度转换并引入 SGCN 后,AP 值提升 0.6%,表明模型初步在语义空间迭代学习关键点之间的拓扑依赖关系。最后引入 HAU&SGCN,模型的 AP 值提升 0.3%,最终达到 76.5% 的平均精度。这表明引入 HAU 可以深入挖掘复合热图特征之间的拓扑分布信息和空间位置信息,生成的特征进一步结合 SGCN 在语义空间中强化躯干结构约束下关键点间的拓扑依赖关系。

综合上述不同数据集上的实验结果和分析,本文提出的 LIEM、VGNM、SGCN 和 HAU 均被验证具有效性和泛化性,在不同数据集对应的评价指标上均有一定提升,在姿态估计最常用的 COCO2017 数据集上效果更为显著,模型完成了有效精准的姿态估计任务。

3.3.2　对比实验

由 COCO2017 验证集上的对比结果(见表 3)可知,当输入图像尺寸为 256×192 时,本文模型两个版本(W32 和 W48)的 AP 值分别达到 76.5% 和 77.2%,优于其他具有相同输入尺寸的经典模型,如 Hourglass、CPN,Simple Baseline 以及 HRFormer-B。RAM-GPRNet(W48)与本文提出模型 W32 版本的 AP 值均达到 76.5%,但本文模型的参数量为其 43.3%,运算量为其 50%,能够更加高效地完成姿态估计任务;与 AMHRNet 对应的两个版本相比,本文模型 AP 值分别提升 0.4% 和 0.8%,且运算量和参数量更低;在运算量和参数量相接近的情况下,本文提出的模型 W32 版本相比 EMF-HRNet 和 TokenPose-L/D24,AP 值分别提升了 0.9% 和 0.7%;与 SCC-Net 相比,模型 W48 实现 3.8 百分点的 AP 值提升。当输入图像尺寸扩大为 384×288 时,本文提出的模型 AP 值分别达到 78.0% 和 78.4%,优于具有相同输入尺寸的所列其他模型。综合而言,本文模型的估计精度优于多数网络,在参数量和运算量上也有相应优势,能够高效精准地完成姿态估计任务。

表3 与其他姿态估计网络在COCO2017验证集上的对比结果

Table 3 Comparison results with different pose estimation networks on COCO2017 validation set

方法	输入尺寸	参数量/M	运算量/G	AP/%	$AP^{0.5}$/%	$AP^{0.75}$/%	AP^{M}/%	AP^{L}/%	AR/%
8-stage Hourglass[6]	256×192	25.1	14.3	66.9	—	—	—	—	—
CPN50[7]	256×192	27.0	6.2	68.6	—	—	—	—	—
CPN50+OHKM[7]	256×192	27.0	6.2	69.4	—	—	—	—	—
Simple Baseline152[8]	256×192	68.6	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNetW32[9]	256×192	28.5	7.1	74.4	90.5	81.9	70.8	81.0	79.8
HRNetW48[9]	256×192	63.6	14.6	75.1	90.6	82.2	71.5	81.8	80.4
TokenPose-L/D24[12]	256×192	27.5	11.0	75.8	90.3	82.5	72.3	82.7	80.9
HRFormer-B[13]	256×192	43.2	12.2	75.6	90.8	82.8	71.7	82.6	80.8
RAM-GPRNet(W32)[30]	256×192	31.4	7.7	76.0	—	—	—	—	—
RAM-GPRNet(W48)[30]	256×192	70.0	15.8	76.5	—	—	—	—	—
EMF-HRNet[31]	256×192	28.8	9.5	75.6	90.4	82.6	72.0	82.4	80.8
AMHRNet(W32)[32]	256×192	36.4	—	76.1	91.0	82.7	71.5	82.9	81.2
AMHRNet(W48)[32]	256×192	71.8	—	76.4	91.1	83.1	72.2	83.3	81.4
SCC-Net[33]	256×192	58.9	10.5	73.4	92.6	81.5	70.4	77.5	76.2
Ours(W32)	256×192	30.3	7.9	**76.5**	92.7	82.8	72.5	82.3	80.8
Ours(W48)	256×192	66.2	17.6	**77.2**	93.0	83.3	72.9	82.7	81.3
CPN50[7]	384×288	—	13.9	70.6	—	—	—	—	—
CPN50+OHKM[7]	384×288	—	13.9	71.6	—	—	—	—	—
Simple Baseline152[8]	384×288	68.6	35.6	74.3	89.6	81.1	70.5	79.7	79.7
HRNetW32[9]	384×288	28.5	16.0	75.8	90.6	82.7	71.9	82.8	81.0
HRNetW48[9]	384×288	63.6	32.9	76.3	90.8	82.9	72.3	83.4	81.2
HRFormer-B[13]	384×288	43.2	26.8	77.2	91.0	83.6	73.2	84.2	82.0
RAM-GPRNet(W32)[30]	384×288	31.4	17.2	77.3	—	—	—	—	—
RAM-GPRNet(W48)[30]	384×288	70.0	35.6	77.7	—	—	—	—	—
EMF-HRNet[31]	384×288	28.8	—	76.5	90.7	83.1	72.7	83.6	81.5
Ours(W32)	384×288	30.3	18.6	**78.0**	93.1	83.5	73.1	82.9	81.4
Ours(W48)	384×288	66.2	37.4	**78.4**	93.3	83.6	73.4	83.7	81.7

在COCO2017测试集上的对比结果如表4所示。结果表明,本文模型能够高效实现姿态估计。模型AP值最高达到78.1%,参数量和运算量方面也有一定优势,在测试集上优于其他主流经典模型,并且与COCO2017验证集上的结果呈现一致性。综合而言,本文模型能够在高精度表征位置信息的同时,结合图卷积算法构建不规则的人体结构模型,建立关键点之间的拓扑连接关系,完成精准高效的姿态估计任务。

表4 与其他姿态估计模型在COCO2017测试集上的对比结果

Table 4 Comparison results with different pose estimation networks on COCO2017 dataset

方 法	输入尺寸	参数量/M	运算量/G	AP/%	$AP^{0.5}$/%	$AP^{0.75}$/%	AP^{M}/%	AP^{L}/%	AR/%
CPN50[6]	384×288	—	—	72.6	86.1	69.7	78.3	64.1	—
Simple Baseline152[7]	384×288	68.6	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet(W32)[8]	384×288	28.5	16.0	74.9	92.5	82.8	71.3	80.9	80.1
HRNet(W48)[8]	384×288	63.6	32.9	75.5	92.5	83.3	71.9	81.5	80.5
TokenPose-L/D24[12]	384×288	29.8	22.1	75.9	92.3	83.4	72.2	82.1	80.8
HRFormer-B[13]	384×288	43.2	26.8	76.2	92.7	83.8	72.5	82.3	81.2
RAM-GPRNet(W32)[30]	384×288	31.4	17.2	76.5	—	—	—	—	—
RAM-GPRNet(W48)[30]	384×288	70.0	35.6	77.0	—	—	—	—	—
Ours(W32)	384×288	30.3	18.6	**77.6**	92.9	83.4	72.8	82.5	81.2
Ours(W48)	384×288	66.2	37.4	**78.1**	93.0	83.6	73.2	83.1	81.5

3.3.3　实验结果可视化

为了直观展示模型的估计效果,验证其有效性,本文选取了部分COCO2017数据集中的估计可视化结果,并与基准模型进行对比分析。

如图7所示,第一行为基准模型HRNet的估计结果,第二行为本文模型的估计结果,通过黄色圆圈标注了二者估计差异较大的关键点。通过对比可知,在一些自遮挡严重(见图7(a)(b))的场景中,本文模型可以根据体位信息精准估计被遮挡的关键点。在一些较复杂场景中形成的其他人物遮挡场景(见图7(c)~(e)),本文模型也能对被遮挡且易混淆的关键点进行精准识别。

此外,针对乐器演奏场景,本文测试了人体动作识别数据集UCF101[34]中不同乐器演奏动作的单帧图像,对比结果展示在图7(f)~(h)中。可以明显看出,针对不同乐器演奏的姿态,本文模型能够精准估计演奏者各个关键点的位置(见图7(f)),当乐器对演奏者产生明显遮挡时(见图7(g)和图7(h)出精准估计,具有较好的鲁棒性。

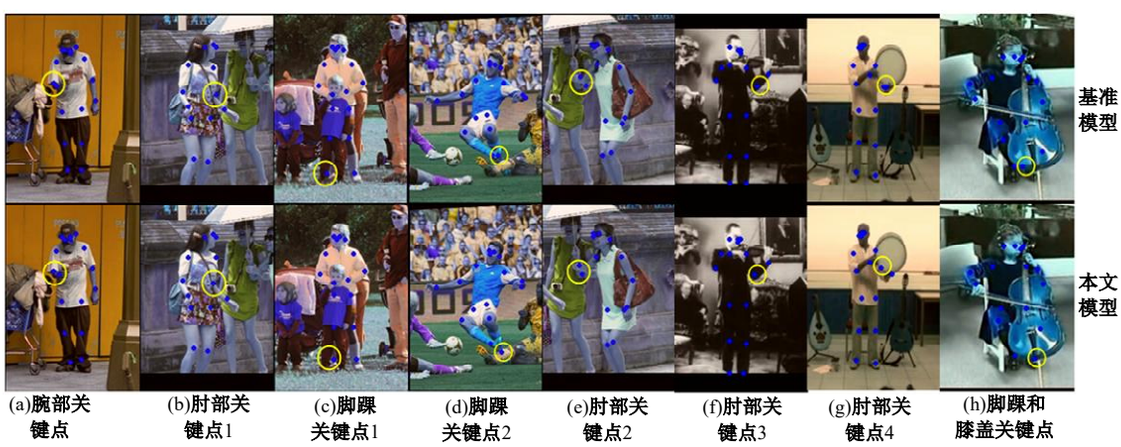

图7　基准模型与本文方法的可视化结果对比

Fig. 7　Comparison of visualization results between baseline model and proposed method

在图8中进一步展示了部分COCO2017数据集中将关键点连接组合为完整人体骨架的对比结果,方便更加清晰地观察不同姿态下关键点之间的对应连接关系。图8中第一行为基准模型HRNet的估计结果,第二行为本文模型的估计结果。通过对比可以明显看出,在一些动作幅度较大的运动姿态中,本文模型的预测表现更好,能够精准定位不同程度遮挡下的关键点,同时避免了关键点错连的问题(见图8(a)(f)),能够实现各种复杂姿态下对易遮挡关键点位的精准估计。

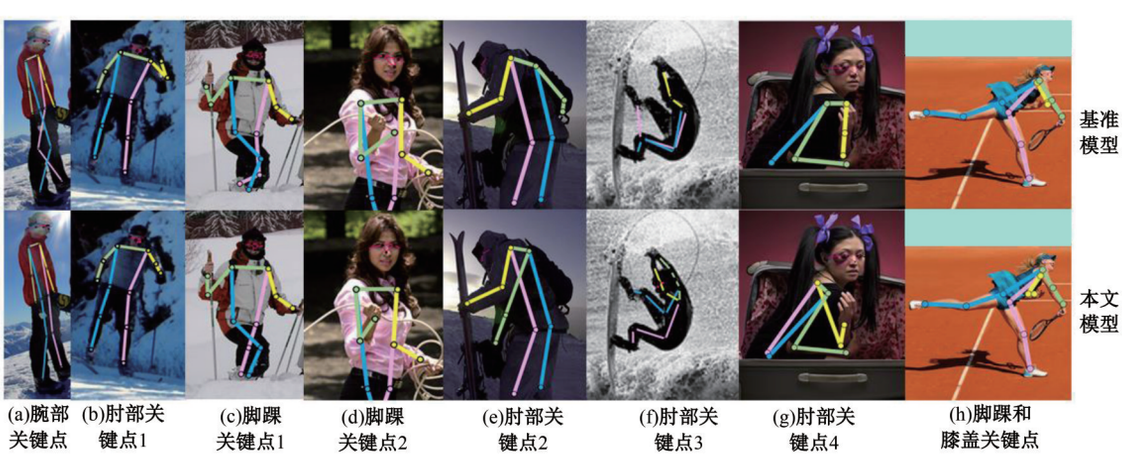

图8　基准模型与本文方法的骨架可视化结果对比

Fig. 8　Comparison of skeleton visualization results between baseline model and proposed method

4　结束语

本文提出了一种基于图结构引导和位置信息强化的人体姿态估计算法,以有效估计复杂姿态中易被遮挡的关键点。该方法能够在特征提取阶段充分利用多尺度特征,实现对可见关键点高精

度的位置信息表征;同时结合图卷积算法构建复杂的人体结构模型,并进一步挖掘热图关键点特征的语义信息,学习关键点间的拓扑连接关系,优化被遮挡关键点的估计。在公开数据集MPII和COCO2017上的实验结果表明,本文姿态估计方法对比其他网络配置具有精度高且参数量低的优势。模型在COCO2017测试集上AP值达到78.1%,具有较高的检测精度。未来工作中,可以考虑在现有优势的基础上,结合Transformer等自注意力机制,在语义层探索各类骨骼热图和对应关键点热图之间的特征交互,进一步强化人体骨骼特征对关键点的约束,实现复杂场景下对遮挡关键点更高精度的估计。

参考文献:

[1] Eduardo RDS, Adams LS, Stoffel R A, et al. Monocular multi-person pose estimation: a survey[J]. Pattern Recognition, 2021, 118: No. 108046.

[2] 田皓宇, 马昕, 李贻斌. 基于骨架信息的异常步态识别方法[J]. 吉林大学学报: 工学版, 2022, 52(4): 725-737.
Tian Hao-yu, Ma Xin, Li Yi-bin. Abnormal gait recognition method based on skeleton information[J]. Journal of Jilin University(Engineering and Technology Edition), 2022, 52(4): 725-737.

[3] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[4] Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1653-1660.

[5] Tompson J, Jain A, Lecun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[C]//Neural Information Processing Systems, Montreal, Canada, 2014: 1799-1807.

[6] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 483-499.

[7] 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, Salt Lake City, USA, 2018: 7103-7112.

[8] Xiao B, Wu H P, Wei Y C. Simple baselines for human pose estimation and tracking[C]//European Conference on Computer Vision, Munich, Germany, 2018: 472-487.

[9] 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, Long Beach, USA, 2019: 5686-5796.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Neural Information Processing Systems(NeurIPS), Long Beach, USA, 2017: 5998-6008.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations, Online, 2021.

[12] Li Y J, Zhang S K, Wang Z C, et al. Tokenpose: Learning keypoint tokens for human pose estimation [C]//Proceedings of the IEEE International Conference on Computer Vision(ICCV), Montreal, Canda, 2021: 11293-11302.

[13] Yuan Y H, Fu R, Huang L, et al. Hrformer: high-resolution transformer for dense prediction[J]. Advances in Neural Information Processing Systems, 2021, 34: 7281-7293.

[14] Yang S, Quan Z B, Nie M, et al. Transpose: Keypoint localization via transformer[C]//Proceedings of the IEEE International Conference on Computer Vision, Montreal, Canda, 2021: 11782-11792.

[15] Li G H, Müller M, Thabet A , et al. DeepGCNs: Can GCNs Go As Deep As CNNs?[C]//IEEE International Conference on Computer Vision, Seoul, South Korea, 2019: 9266-9275.

[16] Qiu L T, Zhang X Y Y, Li Y R, et al. Peeking into occluded joints: A novel framework for crowd pose estimation[C]//European Conference on Computer Vision, Glasgow, UK, 2020: 488-504.

[17] Bin Y R, Chen Z M, Wei X S, et al. Structure-aware human pose estimation with graph convolutional networks[J]. Pattern Recognition, 2020, 106: No. 107410.

[18] Wang J, Long X, Gao Y, et al. Graph-PCNN: Two stage human pose estimation with graph pose refinement[C]//European Conference on Computer Vision, Glasgow, UK, 2020: 492-508.

[19] Banik S, GarcÍa A M, Knoll A. 3D human pose regression using graph convolutional network[C]// IEEE International Conference on Image Processing

(ICIP), Anchorage, USA, 2021: 924-928.

[20] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 13708-13717.

[21] Han K, Wang Y H, Guo J Y, et al. Vision gnn: An image is worth graph of nodes[J]. Advances in Neural Information Processing Systems, 2022, 35: 8291-8303.

[22] Huang G, Liu Z, Laurens V D M, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Honolulu, USA, 2017: 4700-4708.

[23] Ding X H, Guo Y C, Ding G G, et al. ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks[C]//International Conference on Computer Vision, Seoul, South Korea, 2019: 1911-1920.

[24] Zhao L, Peng X, Tian Y, et al. Semantic graph convolutional networks for 3D Human Pose Regression [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 3420-3430.

[25] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 2018: 7794-7803.

[26] Yang J W, Lu J S, Lee S, et al. Graph R-CNN for scene graph generation[C]//European Conference on Computer Vision, Munich, Germany, 2018: 690-706.

[27] Velikovi P, Cucurull G, Casanova A, et al. Graph attention networks[J]. Stat, 2017, 1050(20): No. 10-48550.

[28] Andriluka M, Pishchulin L, Gehler P, et al. 2D human pose estimation: New benchmark and state of the art analysis[C]//IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 3686-3693.

[29] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of the European Conference on Computer Vision(ECCV), Zurich, the Switzerland, 2014: 740-755.

[30] Zhang K, He P, Yao P, et al. Learning enhanced resolution-wise features for human pose estimation [C]//IEEE International Conference on Image Processing, Abu Dhabi, United Arab Emirates, 2020: 2256-2260.

[31] Wang R, Wu W Y, Wang X Y. Enhancing multi-scale information exchange and feature fusion for human pose estimation[J]. The Visual Computer, 2023, 39(10): 4751-4765.

[32] Tran T D, Vo X T, Nguyen D L, et al. High-resolution network with attention module for human pose estimation[C]//Asian Control Conference, Jeju Island, South Korea, 2022: 459-464.

[33] Dong K W, Sun Y J, Cheng X Z, et al. Combining detailed appearance and multi-scale representation: A structure-context complementary network for human pose estimation[J]. Applied Intelligence, 2023, 53 (7): 8097-8113.

[34] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild[J/OL]. [2023-08-16]. https://arxiv. org/abs/1212. 0402.