DOI:10.7544/issn1000-1239.20210723
59(12):2760-2769,2022

部位级遮挡感知的人体姿态估计

褚 真[1] 米 庆[1] 马 伟[1] 徐士彪[2] 张晓鹏[3]
[1](北京工业大学信息学部 北京 100124)
[2](北京邮电大学人工智能学院 北京 100876)
[3](模式识别国家重点实验室(中国科学院自动化研究所) 北京 100190)
(zhen.chu@foxmail.com)

Part-Level Occlusion-Aware Human Pose Estimation

Chu Zhen[1], Mi Qing[1], Ma Wei[1], Xu Shibiao[2], and Zhang Xiaopeng[3]
[1](*Faculty of Information Technology*, *Beijing University of Technology*, *Beijing* 100124)
[2](*Artificial Intelligence School*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876)
[3](*National Laboratory of Pattern Recognition* (*Institute of Automation*, *Chinese Academy of Sciences*), *Beijing* 100190)

Abstract With the rapid development of deep learning, human pose estimation technology has made remarkable progress in recent years, but the existing methods are still difficult to deal with the common occlusion problem. To address this problem, a human pose estimation method based on keypoint-level occlusion inference is proposed in this paper. Firstly, a baseline human pose estimation network is used to obtain the noisy representation of each keypoint of human body from images with occlusion noises. Then, the occluded keypoints are estimated through the occlusion part prediction module to obtain the visibility vector. The occlusion part prediction module is proposed in this study, which consists of two submodules: occlusion part classification network and visibility encoder. The occlusion part classification network predicts the occlusion state of each keypoint of the human body. Based on the channel attention mechanism, the visibility encoder converts the predicted occlusion state into a set of weight parameters. Finally, the visibility vector and noise features are fused by channel re-weighting method to obtain the keypoint-level occlusion aware features, which are used to calculate the heatmaps of the keypoints. Experimental results on MPII and LSP(leeds sports pose) datasets show that, compared with the baseline human pose estimation network, the proposed method can better deal with the occlusion problem at a small extra computational cost, and achieve better results than existing state-of-the-art methods.

Key words human pose estimation; human keypoint detection; occlusion inference; channel attention mechanism; multi-task learning

摘 要 随着深度学习的快速发展,人体姿态估计技术近年来取得显著进步,但是现有方法仍难以较好地处理普遍存在的遮挡问题.针对此问题,提出一种部位级遮挡感知的人体姿态估计方法.首先,采用基准人体姿态估计网络从含遮挡噪声的图像中获得各人体部位的带噪声特征表达.然后,通过遮挡部位

收稿日期:2021-06-29;**修回日期**:2021-12-01
基金项目:国家自然科学基金项目(61771026,62176010);北京市教委重点项目(KZ201910005008)
This work was supported by the National Natural Science Foundation of China (61771026, 62176010) and the Key Project of Beijing Municipal Education Commission, China (KZ201910005008).
通信作者:马伟(mawei@bjut.edu.cn)

预测模块估计人体被遮挡部位,从而获得可见性向量.遮挡部位预测模块由遮挡部位分类网络和可见性编码器组成,前者预测关节点的遮挡状态,后者利用注意力机制将遮挡状态转换为一组权重.最后,通过通道重加权方式融合可见性向量和带噪声特征,获得部位级遮挡感知的人体部位相关特征,用于计算关节点热图.在 MPII 和 LSP(leeds sports pose)数据集上的实验结果表明,相比基准姿态估计网络,该方法能够在较小的额外计算代价下更好地应对遮挡问题,并且取得了比目前先进方法更佳的结果.

关键词　人体姿态估计;人体关节点检测;遮挡推理;通道注意力机制;多任务学习

中图法分类号　TP391.41

人体姿态估计即定位图像或视频中的人体关节点,是计算机视觉中一项基本但极具挑战性的任务,在运动康复、人机交互、自动驾驶[1]等方面有着广泛应用.近年来,深度学习的发展使得人体姿态估计技术取得了突飞猛进的进步.然而,现有方法仍难以较好地处理现实环境中普遍存在的遮挡问题.如何有效应对遮挡问题,进而提升人体姿态估计方法的实用价值,是目前研究的重点和难点.

数据增强是处理遮挡问题的常用方法之一.例如,Ke 等人[2]通过从关节点周围图像背景中裁剪出正方形区域粘贴到关节点位置来模拟遮挡.Bin 等人[3]提出语义数据增强方法,通过生成网络动态地预测增强后的图像,进而利用增强后的数据训练人体姿态估计网络,以提升这些网络在遮挡环境下的鲁棒性.但是,数据增强方法干扰了网络对于关节点表观属性的认知.为此,Zhou 等人[4]提出 OASNet,利用注意力机制预测遮挡感知的注意力图,删除遮挡区域噪声特征,进而重建出因遮挡而缺失的人体区域特征.相比于数据增强方法,遮挡感知方法能够有效去除噪声干扰.然而,目前此类方法只关注遮挡区域在图像空间中的位置,对所有被遮挡关节点等同对待.由于不同关节点表观和上下文关系差异性大,等同对待难以充分利用关节点之间的关系.

本文提出部位级遮挡感知的人体姿态估计方法,以提高人体姿态估计在遮挡下的鲁棒性.所提出方法在基准人体姿态估计网络框架中引入遮挡部位预测模块,该模块由遮挡部位分类网络和可见性编码器组成.其中,遮挡部位分类网络用于预测每个关节点的遮挡状态,记作关节点可见性向量.关节点可见性向量描述了人体各个部位是否被遮挡这一关键信息,可以作为先验知识指导人体姿态估计任务.基于通道注意力思想,可见性编码器将可见性向量转换为一组权重,与基准姿态估计网络提取的卷积特征进行通道重加权,从而迫使网络学习到被遮挡和可见关节点之间的差异,感知遮挡部位,利用相关关节点的上下文修正错误的预测.所提出遮挡部位预测模块具有通用性,适合任何人体姿态估计基准网络,且参数量低,能够以较低的计算代价有效地减轻遮挡的影响.

本文工作的主要贡献有 3 个方面:

1) 提出部位级遮挡感知人体姿态估计方法,通过关节点级别的遮挡推测、知识编码和使用,提升遮挡状态下的人体姿态估计准确度;

2) 构建遮挡部位预测模块,由遮挡部位分类网络和可见性编码器组成.前者预测关节点遮挡状态,后者将遮挡状态编码为人体姿态估计所用先验知识.所构建的遮挡部位预测模块能够兼容不同的基准姿态估计网络;

3) 在合成和实际数据集上的实验均表明,所提出方法能够有效地提升遮挡状态下的人体姿态估计性能.

1　相关工作

首先,回顾近年来人体姿态估计相关工作.其次,由于本文重点解决遮挡问题,在此也将对相似任务中如何处理遮挡问题进行介绍.

1.1　人体姿态估计

传统姿态估计方法[5-7]使用手工构建的特征提取器,往往仅仅考虑小范围的局部特征,特征的丰富度也非常有限,因此很难对姿态做出准确的判断.目前先进的人体姿态估计方法都是基于深度卷积神经网络进行的.DeepPose[8]把深度学习引入到人体姿态估计任务中,它基于卷积神经网络直接回归关节点的坐标.由于直接回归法相对困难,基于热图的方法是目前的主流.CPM(convolutional pose machines)[9]能够提取不同尺度的局部区域的关节点概率,再利用多阶段的方式逐步修正提取的结果.Hourglass[10]使用了 U 型的网络结构,把设计的残差模块作为该网络的基本单元,通过反复的上下采样和同尺度特征

的跨层连接来获取更有效的多尺度信息,并且使用多阶段的网络架构实现逐步优化前一阶段的预测热图的“由粗到精”的学习策略.在 Hourglass 的基础上,PyraNet[11]把残差模块替换为金字塔残差模块,目的是捕捉到细粒度多尺度特征.Tang 等人[12]提出一种复合模型,利用神经网络学习人体的层级结构.Hua 等人[13]在 Hourglass 基础上引入精炼模块和残差注意力模块,以提高上采样效果.Lin 等人[14]提出基于结构化空间学习和中间估计,以保持视频估计结果的时序一致性.SBN(simple baseline network)[15]把 ResNet[16]的全连接层替换为几层反卷积用来增大输出特征图的分辨率,虽然结构简单,但是性能更好.HRNet[17]全程保持高分辨率的表征,并逐渐增加更低分辨率的子网,同时,在并行的子网之间反复交换信息来实现多尺度融合,它超越了以往所有的网络模型,在其他计算机视觉任务中也有着广泛的应用.

尽管取得显著进展,现有人体姿态估计网络仍难以应对遮挡问题.本文提出部位级遮挡感知的人体姿态估计方法,以较低的额外计算代价提升现有网络应对遮挡的鲁棒性,所提出方法能够兼容任何主流人体姿态估计基准网络.

1.2 遮挡处理

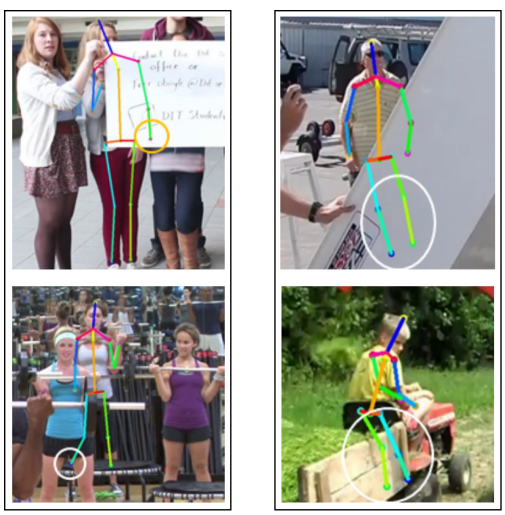
CPN[18]采用 2 阶段的网络结构,利用 GlobalNet 提取的特征帮助 RefineNet 优化被遮挡的困难的关节点的检测结果.Chu 等人[19]利用基于条件随机场的注意力机制来处理遮挡问题.Ke 等人[2]提出的 keypoint masking 技术,通过从关节点周围图像背景中裁剪出正方形区域粘贴到关节点位置来模拟遮挡.Chen 等人[20]利用生成对抗网络预测遮挡部位,通过对抗式学习不断修正预测结果.Bin 等人[3]提出语义数据增强方法,利用生成网络粘贴不同语义粒度的身体部位来模拟挑战性更高的图像.OASNet[4]在人体姿态估计网络上添加了额外的分支,通过监督学习的方式预测图像中遮挡区域的空间位置,然后删除被遮挡区域的特征,再利用孪生网络更好地重建特征图上被遮挡区域的特征,从而降低遮挡的干扰,依靠周边信息恢复被遮挡部位的特征.前述工作尝试感知遮挡所在图像空间位置.本文提出遮挡部位感知的人体姿态估计方法.人体姿态结构性强,感知遮挡部位相比感知遮挡位置更加有助于姿态估计时抹除遮挡对相关部位估计的影响和利用相关部位作为上下文线索对遮挡部位进行更有效推断.

处理遮挡也是其他计算机视觉任务中研究的重点之一.在行人检测中,Zhang 等人[21]发现对于基于卷积网络的行人检测器,不同的通道对与人体不同部位有不同的响应,为此提出了作用于通道上的注意力机制.OR-CNN[22]设计了 AggLoss 最小化建议与对象的距离,并且用部件遮挡感知的 RoI 池化单元替换原有的 RoI 层.Pang 等人[23]提出了 Mask 引导的注意力网络,在增强人体可见区域权重的同时抑制被遮挡的区域.针对遮挡下的人脸关节点进行检测.Zhu 等人[24]提出了遮挡自适应的网络,它可以在高维空间上过滤掉遮挡区域的特征的同时根据上下文恢复出相应的几何信息.与前述工作不同,本文研究结构性更强的人体姿态的估计问题,并提出了部位级遮挡感知的人体姿态估计方法.

2 本文方法

2.1 设计动机

本文以当前性能优秀的 HRNet 和 SBN 为例,测试现有方法在被遮挡节点上的预测效果,结果如图 1 所示,圆圈用于标识预测错误的位置.其中,图 1(a)中遮挡影响了未被遮挡的关节点(左手腕、右脚踝)的检测.图 1(b)中由于遮挡存在,导致预测姿态不自然.简言之,遮挡不仅影响被遮挡的部位,也对与遮挡部位相邻的未被遮挡关节点的定位有一定程度的影响.



(a) HRNet (b) SBN

Fig. 1 Failure examples of existing methods to deal with occlusion problems

图 1 现有方法处理遮挡问题的失败案例

关节点被遮挡也将对其他关节点的预测产生负面影响.为了对比不同部位遮挡对其他关节点估计

的影响,首先基于 MPII 数据集分别在头部、躯干(包含肩膀、髋在内的关节点)、上肢、下肢添加黑色的遮挡;然后排除遮挡部位的关节点,分别计算遮挡下的结果与原始结果的差值,得到其他关节点在遮挡影响下的下降值,再对这些下降值求平均,最终得到遮挡对总体的影响程度 $PCKh@0.5$,在第 i 个关节点上的 $PCKh@0.5$ 定义为

$$PCKh@0.5=\sum_s\delta\left(\frac{d_s^i}{d_s^h}<0.5\right)\Big/\sum_s 1, \quad (1)$$

其中,d_s^i 表示在第 s 个人体的第 i 个关节点上的预测结果与真值之间的距离. d_s^h 表示第 s 个人体的头部尺寸,用于归一化 d_s^i.$\delta(*)$函数在 $*$ 条件成立时为 1,否则为 0,此处用于指示预测误差 d_s^i 是否小于一定阈值,即 $0.5d_s^h$.结果如表 1 所示:

Table 1　Influence of Different Parts of Occlusion on *PCKh*@0.5 of Other Keypoints

表 1　不同部位遮挡对其他关节点 *PCKh*@0.5 的影响

方法	被遮挡部位	头	肩膀	肘	手腕	髋	膝盖	脚踝	平均结果
HRNet[17]	无	97.1	95.9	90.7	86.1	88.8	87.0	82.1	
	头部		95.6	89.8	85.4	88.5	85.5	81.9	87.8(↓0.65)
	躯干	97.8		86.9	79.4		86.7	82.3	86.6(↓1.98)
	上肢	97.3	95.4			87.0	83.2	80.1	88.6(↓1.58)
	下肢	97.3	95.9	90.4	84.8	90.1			91.7(↓0.02)
SBN[15]	无	96.4	95.3	89.0	83.2	88.4	84.0	79.6	
	头部		94.8	87.9	82.3	87.6	83.4	78.7	85.8(↓0.8)
	躯干	96.8		84.3	75.6		83.5	78.2	83.7(↓2.76)
	上肢	96.4	94.8			85.7	80.1	75.8	86.6(↓2.18)
	下肢	96.4	95.5	88.0	82.1	89.5			90.3(↓0.2)

注:"↓"表示下降程度;$PCKh@0.5$ 表示以 0.5 为阈值时的 PCKh 值.

从表 1 中可以看出,遮挡躯干对上肢关节点的检测影响较大.在 HRNet 和 SBN 上的平均 PCKh(head-normalized probability of correct keypoint)@0.5 分别下降了 1.98 和 2.76.究其原因,一方面是由于躯干与上肢直接相连,关联度高;另一方面是由于躯干面积较大且人体上肢灵活,上肢经常与躯干重叠,形成人体自遮挡.同理,遮挡上肢对其他关节点的影响也较大,在 HRNet 和 SBN 上平均 $PCKh@0.5$ 分别下降了 1.58 和 2.18.此外,遮挡头部对检测其他关节点有一定影响,在 HRNet 和 SBN 上平均 $PCKh@0.5$ 分别下降了 0.65 和 0.8.而由于 MPII 数据集中人体姿态多为站立,与其他部位距离较远,因而遮挡下肢对其他关节点的检测影响较小.

综上,人体部位遮挡对自身以及与之相关的其他部位均有一定程度的影响.如果获得关节点级别遮挡线索,则可通过上下文更好地优化被遮挡关节点的定位,同时减少其对其他关节点的影响,提高人体姿态估计模型应对遮挡的能力.

2.2　方法整体架构

本文方法的整体架构如图 2 所示.首先,将输入图像同时输入基准姿态估计网络和遮挡部位预测模块.然后,使用遮挡部位预测模块的输出对基准姿态

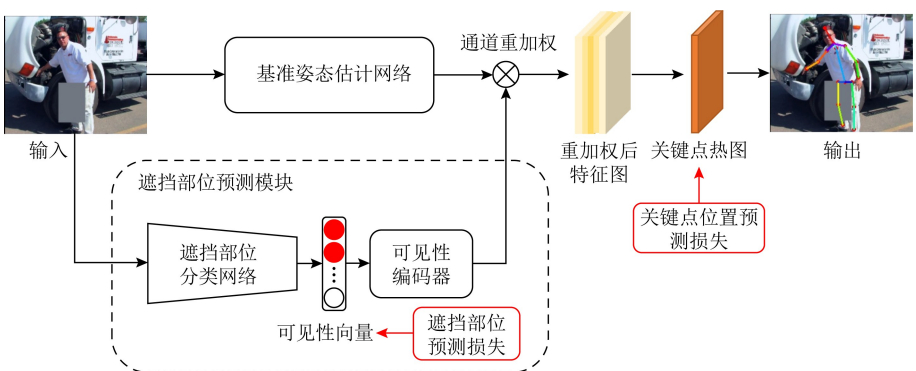


Fig. 2　The overall architecture of proposed method

图 2　本文方法整体架构

估计网络提取的特征施加通道重加权操作,得到优化后的特征.最后,使用 1×1 卷积获得最终结果.其中,基准姿态估计网络可以是现有任何人体姿态网络.所提出遮挡部位预测模块由遮挡部位分类网络和可见性编码器(visibility encoder, VE)组成.下面分别对其进行介绍.

2.3 遮挡部位分类网络

为了获得关节点级别的遮挡线索,所提出遮挡部位分类网络将根据输入图像预测人体每个关节点的遮挡状态.而关节点仅有被遮挡和可见 2 种状态,因而相比于人体姿态估计的回归任务,遮挡部位分类任务更简单,模型也更容易收敛.因此,权衡计算量和精度,遮挡部位分类网络将使用轻量级网络 MobileNetV2[25] 作为主干网络,用于提取适合遮挡部位分类任务的特征,获得每个关节点可见性向量,作为可见性编码器的输入.可见性向量表示为

$$\boldsymbol{o}=(v_0p_0,v_1p_1,\cdots,v_kp_k), \tag{2}$$

其中,p_i 表示人体每个关节点,v_i 是一个二值变量,表示第 i 个关节点是否被遮挡,$v_i\in\{0,1\}$,$i\in[0,k]$,0 表示被遮挡,1 表示可见.

本文对 MobileNetV2 做出适当修改以适应关节点的遮挡分类任务.将 MobileNetV2 末尾用于图像分类的 1000 维全连接层分类器替换为输出通道数为 n 的 1×1 卷积.

在训练阶段,采用二分类交叉熵损失监督遮挡分类网络训练过程,以最小化在每个关节点上的遮挡状态预测误差.遮挡分类预测损失定义为:

$$L_{\text{occ}}=\sum_{j=0}^{n}-(y_j\ \text{lb}\ (y'_j)+(1-y_j)\ \text{lb}\ (1-y'_j)), \tag{3}$$

其中,y'_j 是模型预测的第 j 个关节点可见的概率,y_j 表示真值,即数据集中提供的关节点可见性标签,n 为关节点总数.

2.4 可见性编码器

为了将可见性向量与带噪声特征融合,首先利用可见性编码器扩展可见性向量的维度,然后利用通道注意力机制对带噪声特征进行重加权.前述过程可表示为

$$f_{\text{occ}}=\boldsymbol{\Omega}^{\text{T}}f_{\text{ch}}, \tag{4}$$

其中,f_{ch} 为需要被通道重加权的特征,$\boldsymbol{\Omega}$ 为权重参数向量.

本文选择通道重加权而非其他特征融合方式的原因如下:首先,基于热图的人体姿态估计方法将人体关节点转化为以关节点位置为中心的 2 维高斯热图,网络末尾使用 1×1 卷积将高维特征转化为与关节点数量相等的热图,关联了不同部位间的影响关系,说明关节点的信息与通道相关.其次,深层的网络能够学习出人体整体的结构,建模关节点之间的关系.而遮挡部位分类网络预测到的可见性向量仅表达了关节点独自的遮挡状态信息,缺乏关节点之间的关联信息.因此,通道重加权能够更好地利用关节点之间的上下文信息,并在本文所提出的可见性编码器的帮助下,利用注意力机制区分被遮挡与未被遮挡部位直接的差异,利用相关部位的上下文线索克服遮挡的干扰.

为了获得权重参数向量 $\boldsymbol{\Omega}$,利用可见性编码器把可见性向量编码到更高维度的特征上.具体而言,利用可见性编码器把可见性向量转换为一组维度与基准姿态估计网络提取的卷积特征通道数相等的权重,其值小于 1.然后对卷积特征进行通道重加权.该过程的公式表示为

$$\boldsymbol{\Omega}=F(\boldsymbol{o}), \tag{5}$$

$$F=Sigmoid(F_2(F_1(\boldsymbol{o}))), \tag{6}$$

其中,F 表示可见性编码器,其结构如图 3 所示.输入为遮挡部位分类网络的输出,即可见性向量 $\boldsymbol{o}$.经过 2 个全连接层 F_1 和 F_2 使得向量的维度和基准姿态估计网络提取的卷积特征通道数相同,再经过 $Sigmoid$ 函数使该模块输出向量每个元素的值调整为 0 和 1 之间,得到权重参数向量 $\boldsymbol{\Omega}$.再与基准人体姿态估计网络提取的卷积特征 f_{ch} 进行对应通道上相乘,得到重加权后的特征 f_{occ}.

当基准姿态估计网络为 HRNet 时,2 个全连接层输出通道数分别为 64 和 32.此时,将可见性编码器模块添加在 HRNet 的 stage 4 之后、1×1 卷积之前;当基准姿态估计网络为 SBN 时,2 个全连接层输出通道数分别为 64 和 256.此时,将可见性编码器添加在最后一层反卷积后.

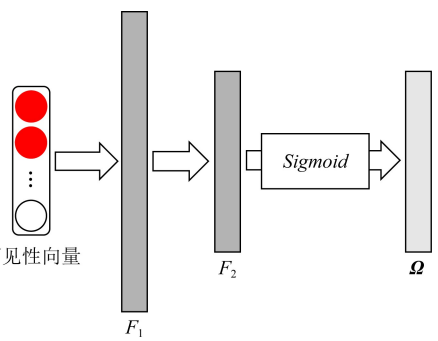


Fig. 3 The network sturcture of the visibility encoder

图 3 可见性编码器的网络结构

2.5　整体架构训练损失函数

本文对人体姿态估计网络和部位级遮挡分类网络进行联合端到端训练.为此,对人体姿态估计网络预测的关节点热图和遮挡分类模块预测的关节点可见性向量的整体损失进行度量,整体损失函数为

$$L=L_{\mathrm{hm}}+\lambda L_{\mathrm{occ}}, \tag{7}$$

其中,L_{hm}表示人体姿态估计关节点热图检测的 L2 损失函数,L_{occ}由式(2)给出,表示用于遮挡部位分类的二分类交叉熵损失函数.λ 为平衡 2 个损失函数的超参数.鉴于遮挡分类任务优化较快,本文设 $\lambda=0.01$.基准姿态估计模块和遮挡分类网络均使用 ImageNet 预训练模型进行参数初始化.

3　实验结果与分析

3.1　数据集和评价指标

MPII 数据集是一个用于 2 维人体姿态估计任务的数据集,包含约 25 000 张从真实场景中采集的图像和超过 40 000 个人体关节点标注,其中每人共有 16 个关节点被标注,是单人姿态估计任务的主流数据集.

LSP(leeds sports pose)数据集由 2 000 个样本原始数据集和 10 000 个样本的扩展数据集组成.其中,原始数据集中的 1 000 个样本用于测试,其余 11 000 个样本用于训练.每人有 14 个标注的关节点.

PCKh 是 MPII 和 LSP 数据集的评价指标,用于计算检测的关节点与其真值的归一化距离小于预设阈值(头部长度)的比例.

本文分别基于 MPII 和 LSP 数据集,构建合成随机矩形遮挡的图像数据集,包括训练集和验证集.随机矩形遮挡的高是人体目标框高度的[1/4,1/2]之间的随机值,宽是人体目标框宽度的[1/2,1]倍之间的随机值.遮挡区域的位置在人体的包围框内,颜色是图像的平均值.

3.2　实验设置

训练阶段实验设置.实验基于 PyTorch 框架在 GTX 1080Ti GPU 上训练,并使用了 ImageNet 的预训练参数.参考 Xiao 等人[15]、Sun 等人[17]的实验设置,输入图像大小调整为 256×256,批大小为 32,优化器为 Adam,初始学习率为 0.001.基于 HRNet 基准网络,迭代训练到 170 和 200 轮时,学习率分别下降至 0.000 1 和 0.000 01,总共训练 210 轮;基于 SBN 基准网络,迭代训练到 90 和 120 轮时,学习率分别下降至 0.000 1 和 0.000 01,总共训练 150 轮.数据增强的策略包括−45°～45°随机旋转,0.65～1.35 随机尺度变换和左右随机翻转.

测试阶段实验设置.输入图像经过网络推理得到热图后,对该热图和翻转后的热图对应位置求平均,得到最终的热图.在后处理时,参考 Hourglass[10],将热图上值最高的一点向次高点的 1/4 像素的偏移作为最终的关节点预测位置.

3.3　消融实验

本文首先设计消融实验确定方法的最终结构.所有消融实验均基于 HRNet 和所构建合成 MPII 数据集训练和测试.以下分别介绍通道重加权位置、可见性编码器结构、遮挡部位分类网络和遮挡部位预测模块通用性的消融实验.

3.3.1　通道重加权位置

本文基于 HRNet 设计消融实验,对比在 3 个位置(A,B,C)施加通道重加权(如图 4 所示)对结果的影响.其中:A 表示在 HRNet 前执行重加权;B 表示在主干网络提取到特征之后执行重加权;C 表示经过最后一个 1×1 卷积,得到的 16 个关节点热图后再执行重加权;1,2,3,4 表示 HRNet 四个阶段网络结构组成.实验结果如表 2 所示,在位置 B 施加通道重加权操作的效果最好.

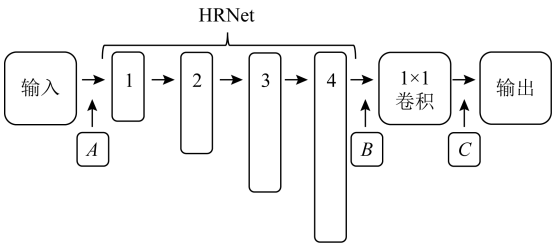


Fig. 4　Indication of the positions where channel weighting is applied on HRNet

图 4　在 HRNet 上施加通道重加权位置示意图

Table 2　Ablation Study of HRNet with Channel Re-weighting at Different Positions

表 2　在 HRNet 不同位置施加通道重加权消融实验

位置	平均 $PCKh@0.5$
A	87.5
B	**87.9**
C	87.8

注:黑体数字表示最优值.

3.3.2　可见性编码器结构

在基于 HRNet 验证可见性编码器结构的消融实验中,比较该模块不同数量的全连接层之间的差异.

实验结果如表 3 所示,当全连接层数量为 2 时,可见性编码器的结果最好,因此本文方法最终选择 2 层全连接的可见性编码器.

Table 3 Ablation Study on the Number of Fully Connection Layers in the Visibility Encoder

表 3 可见性编码器全连接层数量消融实验

全连接层数量	平均 *PCKh*@0.5
1	87.5
2	**87.9**
3	87.8

注:黑体数字表示最优值.

3.3.3 遮挡部位分类网络

遮挡部位预测模块所使用的遮挡分类网络可以是现有的任何轻量级分类网络.本文选择有代表性的轻量级网络 MobileNetV2[25],ShuffleNetV2[26],GhostNet[27]进行对比实验,结果如表 4 所示.从表 4 中可看出,选择不同的分类网络对最终姿态估计结果的影响极小,因此本文选择经典的 MobileNetV2 作为遮挡部位分类网络.

Table 4 Ablation Study of the Occlusion Classification Network

表 4 遮挡分类网络消融实验

遮挡分类网络	平均 *PCKh*@0.5
MobileNetV2[25]	**87.887**
ShuffleNetV2[26]	87.869
GhostNet[27]	87.875

注:黑体数字表示最优值.

3.3.4 遮挡部位预测模块通用性

鉴于本文方法兼容所有基准人体姿态估计网络,为了使其性能最优,设计验证遮挡部位预测模块通用性的消融实验,结果如表 5 所示.在 HRNet 和 SBN 中引入遮挡部位预测模块后,平均指标分别提升了 0.3 和 0.5.实验结果说明所提出的遮挡部位预测模块能广泛提升现有方法在合成遮挡下的性能.综上,最终选择 HRNet 作为本文方法的基准姿态估计网络,用于和现有方法横向比较.

Table 5 Ablation Study for Verifying the Universality of the Visibility Encoder

表 5 验证可见性编码器通用性的消融实验

方法	头	肩膀	肘	手腕	髋	膝盖	脚踝	平均 *PCKh*@0.5
HRNet[17]	94.7	94.4	88.6	81.6	**87.6**	82.0	79.0	87.6
本文方法(HRNet-w32)	**95.0**	**94.7**	**88.7**	**82.1**	87.4	**82.3**	**80.7**	**87.9**
SBN[15]	94.1	92.9	84.7	76.4	84.1	76.5	72.4	83.9
本文方法(ResNet-50)	93.7	**93.1**	**85.0**	**77.0**	**84.8**	**77.3**	74.0	**84.4**

注:黑体数字表示最优值.

3.4 横向对比实验——效率分析

表 6 给出了各方法的参数量、计算量和在 RTX 3090 显卡上的推理速度(输入图片的尺寸为 256×256)的横向对比.从表 6 中可看出,本文方法设计的遮挡预测模块的参数量和计算量分别为 30.8 MB 和 9.85 GFLOPS,相比基准网络 HRNet 分别仅增加 8.0%和 3.8%,且推理速度仅慢 5.9%,达到 143 fps.

Table 6 Efficiency Comparison Between the Proposed Method and Existing Methods

表 6 本文方法与现有方法的效率对比

方法	参数量/MB	计算量/GFLOPS	推理速度/fps
SBN[15]	34.0	11.99	79
HRNet[17]	28.5	9.49	152
本文方法	30.8	9.85	143
本文方法与 HRNet 相比/%	8.0	3.80	−5.9

进一步说明本文方法在维持较低计算代价的同时,有效降低遮挡对人体姿态估计的影响.

3.5 横向对比实验——量化对比与分析

通过 3.3 节的消融实验确定了本文方法的最终结构,即基准姿态估计网络为 HRNet,通道重加权施加在 HRNet 尾部和 1×1 卷积之间,可见性编码器使用 2 层全连接层.以下分别在 MPII 与 LSP 数据集上进行横向对比实验与分析.

在 MPII 数据集上,利用在合成遮挡 MPII 训练集上训练得到本文模型.在实际 MPII 验证集上测试该模型,并将其与多种先进的方法做横向对比,结果如表 7 所示.所有对比方法的结果数值取自原文献.其中 SBN,PyraNet,DLCM,Hourglass,HRNet 为在 MPII 数据集上的原始结果,而 OASNet 和本文方法都使用了构建遮挡的数据增强策略,显式地利用遮挡信息.

从表 7 中可看出,本文方法平均准确度优于其他方法,尤其在人体四肢等灵活度大、挑战性高的关节点上优势更明显.相比对比方法中表现最好的 OASNet,本文方法在头部、肘部、手腕、髋和膝盖关节点上,*PCKh*@0.5 得分值分别领先 0.1,0.2,0.2,0.9,0.4,平均 *PCKh*@0.5 得分值为 91.0,领先 OASNet 方法 0.3.

综上可看出,本文所提出的部位级遮挡感知的人体姿态估计方法推测关节点级别的遮挡线索,在此基础上利用上下文优化被遮挡关节点的定位,同时减小了被遮挡关节点对未被遮挡关节点的影响,能够显著提升人体姿态估计模型在应对遮挡问题上的性能.

表 8 给出了在 LSP 数据集上的测试结果.从表 8 中可看出,本文方法在多数关节点上,尤其是灵活度高的四肢上,准确度高于现有方法.

Table 7　Comparison Between the Proposed Method and Existing Methods on the MPII Valid Set

表 7　本文方法与现有方法在 MPII 验证集上的横向对比

方法	头	肩膀	肘	手腕	髋	膝盖	脚踝	平均 *PCKh*@0.5
SBN[15]	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5
PyraNet[11]	96.8	96.0	90.4	86.0	89.5	85.2	82.3	89.6
DLCM[12]	95.6	95.9	90.7	86.5	89.9	86.6	82.5	89.8
Hourglass[10]	97.2	95.8	90.3	86.0	89.8	85.8	82.7	90.1
HRNet[17]	97.1	95.9	90.7	86.1	88.8	87.0	82.1	90.2
OASNet[4]	97.2	**96.5**	91.1	86.8	89.6	87.1	**83.6**	90.7
本文方法	**97.3**	**96.5**	**91.3**	**87.0**	**90.5**	**87.5**	83.5	**91.0**

注:黑体数字表示最优值.

Table 8　Comparison Between the Proposed Method and Existing Methods on the LSP Test Set

表 8　本文方法与现有方法在 LSP 测试集上的横向对比

方法	头	肩膀	肘	手腕	髋	膝盖	脚踝	平均 *PCKh*@0.2
Wei 等人[9]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Chu 等人[19]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.5
Yang 等人[11]	98.3	94.5	92.2	86.0	89.8	85.8	82.7	90.1
Tang 等人[28]	98.6	**95.4**	**93.3**	89.8	94.3	**95.7**	94.4	94.5
Zhou 等人[4]	98.8	95.2	92.3	89.8	**95.2**	95.5	94.7	94.5
本文方法	**98.9**	95.3	92.5	**89.9**	**95.2**	**95.7**	**94.8**	**94.6**

注:黑体数字表示最优值.

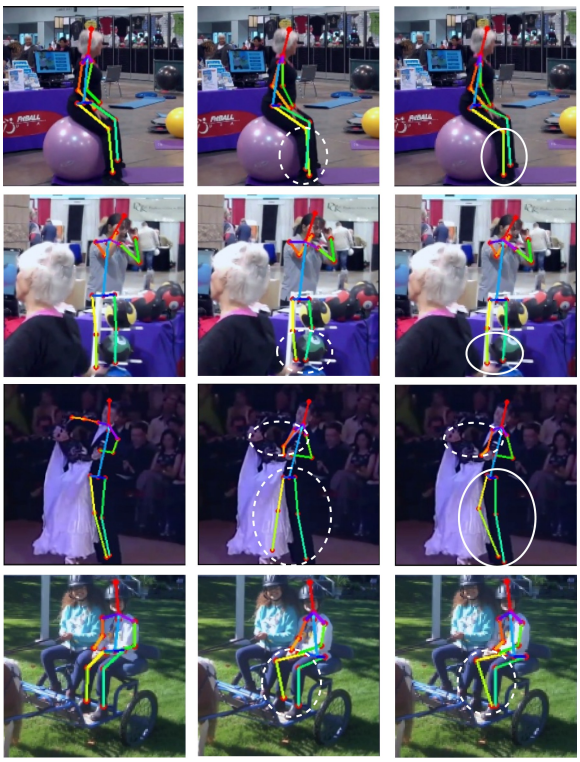
3.6　横向对比实验——可视化对比与分析

图 5 展示了本文方法与 HRNet 在原始 MPII 验证集上的可视化结果.图 5 中 3 列分别为真值、HRNet 和本文方法在相同图像上的可视化结果.实线圆圈和虚线圆圈分别标识了 HRNet 和本文方法预测正确和预测失败的例子.

从图 5 可看出,第 1 行图像中人的双脚距离近且互相遮挡,导致 HRNet 错误地预测了 2 只脚的位置,所估计的双腿的姿态与真值相比出现明显偏差.而本文方法通过对遮挡部位的预测,避免了脚关节点遮挡对于腿部其他关节点的影响,同时借助其他可见关节点成功预测了脚关节点位置;第 2 行图像中双脚表观较为模糊,且与之相邻的膝盖关节点被遮挡,干扰了 HRNet 对双脚关节点的准确定位.本文方法能够在提升被遮挡关节点检测精度的同时,减少其对双脚关节点预测的干扰,令网络对姿态的估计更加合理;第 3 行中 HRNet 完全错误地预测图像中男士被遮挡的右脚位置,使得估计到的姿态为右脚翘起的错误状态,而本文方法结果合理、更加接近真值.

综上可看出,本文方法够有效克服遮挡对自身部位和相关部位的影响.

本文方法仍有不足之处,尚难以处理如复杂背景导致的挑战性高的情形.如图 5 第 3 行例子中所示,本文方法对右臂关节点的预测相比 HRNet 来说没有改进,结果仍然错误.第 4 行例子中,本文方法错将旁人相近关节点当作主体对象关节点,且右脚位置有一定偏移.



(a) 真值 (b) HRNet (c) 本文方法

注:实线圆圈和虚线圆圈分别标识预测正确和失败的例子.

Fig. 5 Visual comparison of proposed method and HRNet on the MPII valid set

图 5 本文方法与 HRNet 在 MPII 验证集上的可视化对比

4 结论与展望

本文提出部位级遮挡感知的人体姿态估计方法,通过在基准人体姿态估计网络中引入所提出遮挡部位预测网络,有效降低遮挡对人体姿态估计任务的影响.实验表明,本文方法在较小的计算代价下能够增强多种基准方法应对遮挡的能力,尤其对于四肢等灵活度高的部位较为明显.

本文方法对复杂背景下的人体关节点预测能力仍然有限.原因在于本文方法依赖所学习的遮挡线索处理遮挡问题,未进一步考虑关节点之间更全局的关系.现有方法大多根据人类经验设计关节点之间的关系模型,仅关注局部信息,而忽视了潜在的全局关联.为了从全局视角下建模关节点之间的关系,在未来的工作中,将考虑设计遮挡状态下基于数据驱动的关节点影响关系建模.同时,探索基于图神经网络融合全局关节点关系的人体姿态优化算法,以提升本任务在遮挡状态下的准确性.

作者贡献声明:褚真提出研究思路、设计方案,进行实验、起草论文;米庆、马伟负责对文章内容进行指导及修订;徐世彪、张晓鹏负责论文的指导.

参 考 文 献

[1] Zhang Yanyong, Zhang Sha, Zhang Yu, et al. Multi-modality fusion perception and computing in autonomous driving [J]. Journal of Computer Research and Development, 2020, 57(9): 1781-1799 (in Chinese)
(张燕咏, 张莎, 张昱, 等. 基于多模态融合的自动驾驶感知及计算[J]. 计算机研究与发展, 2020, 57(9): 1781-1799)

[2] Ke Lipeng, Chang Ming-Ching, Qi Honggang, et al. Multi-scale structure-aware network for human pose estimation [C] //Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 713-728

[3] Bin Yanrui, Cao Xuan, Chen Xinya, et al. Adversarial semantic data augmentation for human pose estimation [C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 606-622

[4] Zhou Lu, Chen Yingying, Gao Yunze, et al. Occlusion-aware siamese network for human pose estimation [C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 396-412

[5] Tian Yuandong, Zitnick C L, Narasimhan S G. Exploring the spatial hierarchy of mixture models for human pose estimation [C] //Proc of the 12th European Conf on Computer Vision. Berlin: Springer, 2012: 256-269

[6] Chen Yaodong, Li Renfa, Li Shiying, et al. A combined grammer for object detection and pose estimation [J]. Chinese Journal of Computers, 2014, 37(10): 2206-2217 (in Chinese)
(陈耀东, 李仁发, 李实英, 等. 面向目标检测与姿态估计的联合文法模型 [J]. 计算机学报, 2014, 37(10): 2206-2217)

[7] Yang Yi, Ramanan D. Articulated human detection with flexible mixtures of parts [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(12): 2878-2890

[8] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks [C] //Proc of the 27th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 1653-1660

[9] Wei S, Ramakrishna V, Kanade T, et al. Convolutional pose machines [C] //Proc of the 29th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4724-4732

[10] Newell A, Yang Kaiyu, Deng Jia. Stacked hourglass networks for human pose estimation [C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 483-499

[11] Yang Wei, Li Shuang, Ouyang Wanli, et al. Learning feature pyramids for human pose estimation [C] //Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1281-1290

[12] Tang Wei, Yu Pei, Wu Ying. Deeply learned compositional models for human pose estimation [C] //Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 190-206

[13] Hua Guoguang, Li Lihong, Liu Guang. Multi-path affinage stacked-hourglass networks for human pose estimation [J]. Frontiers of Computer Science, 2020, 14(4): 1-12

[14] Liu Shiguang, Li Yang, Hua Guoguang, et al. Human pose estimation in video via structured space learning and halfway temporal evaluation [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(7): 2029-2038

[15] Xiao Bin, Wu Haiping, Wei Yichen. Simple baselines for human pose estimation and tracking [C] //Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 466-481

[16] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the 29th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[17] Sun Ke, Xiao Bin, Liu Dong, et al. Deep high-resolution representation learning for human pose estimation [C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5693-5703

[18] Chen Yilun, Wang Zhicheng, Peng Yuxiang, et al. Cascaded pyramid network for multi-person pose estimation [C] //Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7103-7112

[19] Chu Xiao, Yang Wei, Ouyang Wanli, et al. Multi-context attention for human pose estimation [C] //Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1831-1840

[20] Chen Yu, Shen Chunhua, Wei Xiushen, et al. Adversarial posenet: A structure-aware convolutional network for human pose estimation [C] //Proc of the 16th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1212-1221

[21] Zhang Shanshan, Yang Jian, Schiele B. Occluded pedestrian detection through guided attention in CNNs [C] //Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6995-7003

[22] Zhang Shifeng, Wen Longyin, Bian Xiao, et al. Occlusion-aware R-CNN: Detecting pedestrians in a crowd [C] //Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 637-653

[23] Pang Yanwei, Xie Jin, Khan M H, et al. Mask-guided attention network for occluded pedestrian detection [C] //Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4967-4975

[24] Zhu Meilu, Shi Daming, Zheng Mingjie, et al. Robust facial landmark detection via occlusion-adaptive deep networks [C] //Proc of the 32nd Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3486-3496

[25] Sandler M, Howard A, Zhu Menglong, et al. MobileNetV2: Inverted residuals and linear bottlenecks [C] //Proc of the 31st Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4510-4520

[26] Ma Ningning, Zhang Xiangyu, Zheng Haitao, et al. ShuffleNetV2: Practical guidelines for efficient CNN architecture design [C] //Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 116-131

[27] Han Kai, WangYunhe, Tian Qi, et al. GhostNet: More features from cheap operations [C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 1580-1589

[28] Tang Wei, Wu Ying. Does learning specific features for related parts help human pose estimation? [C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 1107-1116

Chu Zhen, born in 1995. Master. His main research interests include human pose estimation and deep learning.

褚　真,1995 年生.硕士.主要研究方向为人体姿态估计和深度学习.

Mi Qing, born in 1987. PhD, lecturer. Member of CCF. Her main research interests include code style and code readability analysis, data mining and analytics, deep learning, etc.

米　庆,1987 年生.博士,讲师.CCF 会员.主要研究方向为代码风格及代码可读性分析、数据挖掘与分析、深度学习等.

Ma Wei, born in 1980. PhD, associate professor. Member of CCF. Her main research interests include computer vision, image management.

马　伟,1980 年生.博士,副教授.CCF 会员.主要研究方向为计算机视觉、图像处理.

Xu Shibiao, born in 1986. PhD, professor, master supervisor. Member of CCF. His main research interests include computer vision and graphics.

徐士彪,1986 年生.博士,教授,硕士生导师.CCF 会员.主要研究方向为计算机视觉与图形学.

Zhang Xiaopeng, born in 1963. PhD, researcher, PhD supervisor. Member of CCF. His main research interests include computer graphics and virtual reality.

张晓鹏,1963 年生.博士,研究员,博士生导师.CCF 会员.主要研究方向为计算机图形学、虚拟现实.