中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2024)12-3529-14

论文引用格式:Xu L H, Zhao L, Sun X X, Yan K D and Li G Y. 2024. A comprehensive review of progress in deep-learning-based occluded human pose estimation. Journal of Image and Graphics, 29(12):3529-3542(徐琳皓, 赵林, 孙辛欣, 颜克冬, 李广宇. 2024. 基于深度学习的遮挡人体姿态估计进展综述. 中国图象图形学报, 29(12):3529-3542)[DOI:10.11834/jig.230730]

基于深度学习的遮挡人体姿态估计进展综述

徐琳皓[1],赵林[1*],孙辛欣[2],颜克冬[1],李广宇[1]

1. 南京理工大学计算机科学与工程学院,南京　210094; 2. 南京理工大学设计艺术与传媒学院,南京　210094

摘　要: 人体姿态估计(human pose estimation,HPE)是计算机视觉中的一项基本任务,旨在从给定的图像中获取人体关节的空间坐标,在动作识别、语义分割、人机交互和人员重新识别等方面得到了广泛应用。随着深度卷积神经网络(deep convolutional neural network,DCNN)的兴起,人体姿态估计取得了显著进展。然而,尽管取得了不错的成果,人体姿态估计仍然是一项具有挑战性的任务,特别是在面对复杂姿态、关键点尺度的变化和遮挡等因素时。为了总结关于遮挡的人体姿态估计技术的发展,本文系统地概述了自2018年以来的代表性方法,根据神经网络包含的训练数据、模型结构以及输出结果,将方法细分为基于数据增广(data augmentation)的预处理、基于特征区分的结构设计和基于人体先验的结果优化3类。基于数据增广方法通过生成遮挡的数据来增加训练样本;基于特征区分的方法通过利用注意力机制等方式来减少干扰特征;基于人体结构先验的方法通过利用人体结构先验来优化遮挡姿态。同时,为了更好地评测遮挡方法的性能,重新标注了MSCOCO (Microsoft common objects in context)val2017数据集。最后,对各种方法进行了对比和总结,阐明了它们在面对遮挡时性能的优劣。此外,在此基础上总结和讨论了遮挡情况下人体姿态估计困难的原因以及该领域未来的发展趋势。

关键词: 人体姿态估计(HPE);遮挡;数据增广;人体结构先验;遮挡标注数据不足

A comprehensive review of progress in deep-learning-based occluded human pose estimation

Xu Linhao[1], Zhao Lin[1*], Sun Xinxin[2], Yan Kedong[1], Li Guangyu[1]

1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing　210094, China;

2. School of Design Art and Media, Nanjing University of Science and Technology, Nanjing　210094, China

Abstract: Human pose estimation (HPE) is a prominent area of research in computer vision whose primary goal is to accurately localize annotated keypoints of the human body, such as wrists and eyes. This fundamental task serves as the basis for numerous downstream applications, including human action recognition, human-computer interaction, pedestrian re-identification, video surveillance, and animation generation, among others. Thanks to the powerful nonlinear mapping capabilities offered by convolutional neural networks, HPE has experienced notable advancements in recent years. Despite this progress, HPE remains a challenging task, particularly when facing complex postures, variations in keypoint scales, occlusion, and other factors. Notably, the current heatmap-based methods suffer from severe performance degradation

收稿日期:2023-10-27;**修回日期**:2024-01-19;**预印本日期**:2024-01-26

*通信作者:赵林　linzhao@njust.edu.cn

基金项目:国家自然科学基金项目(62172222);国家重点研发计划(国际合作专项)资助(SQ2023YFE0102775)

Supported by: National Natural Science Foundation of China(62172222); National Key R&D Program of China (International Collaboration Special Project)(SQ2023YFE0102775)

when encountering occlusion, which remains a critical challenge in HPE given that diverse human postures, complex backgrounds, and various occluding objects can all cause performance degradation. To comprehensively delve into the recent advancements in occlusion-aware HPE, this paper not only explores the intricacies of occlusion prediction difficulties but also delves into the reasons behind these challenges. The identified challenges encompass the absence of annotated occluded data. Annotating occluded data is inherently complex and demanding. Most of the prevalent datasets for HPE predominantly focus on visible keypoints, with only a few datasets addressing and annotating occlusion scenarios. This deficiency in annotated occluded data during model training significantly compromises the robustness of models in effectively handling situations that involve a partial or complete obstruction of body keypoints. Feature confusion presents a key challenge for top-down HPE methods, where the reliance on detected bounding boxes extracted from the image leads to the cropping of the target person's region for keypoint prediction. However, in the presence of occlusion, these detection boxes may include individuals other than the target person, thereby interfering with the accurate prediction of keypoints. This issue is particularly problematic because the high feature similarity between the target person and the interfering individuals prevents the model from distinguishing features effectively, thereby compromising the accuracy of keypoint predictions and emphasizing the need to develop strategies for addressing feature confusion in occluded scenes. Navigating the intricacies of inference becomes particularly challenging in the presence of substantial occlusion. The expansive coverage of occlusion leads to the loss of valuable contextual and structural information that is essential for accurately predicting the occluded keypoints. Contextual cues and structural insights play pivotal roles in the inference process, and their absence impedes the model's ability to draw precise conclusions. The significant loss of contextual information also hampers the model's capacity to glean necessary details from adjacent keypoints, which is crucial for making informed predictions about occluded keypoints. This, in turn, results in the potential omission of keypoints or the emergence of anomalous pose estimations. Besides, this paper systematically reviews representative methods since 2018. Based on the training data, model structure, and output results contained in neural networks, this paper categorizes methods into three types, namely, preprocessing based on data augmentation, structural design based on feature discrimination, and result optimization based on human body priors. Preprocessing based on data augmentation techniques, which generate data with occlusion, are employed to augment training samples, compensate for the lack of annotated occluded data, and alleviate the performance degradation of the model in the presence of occlusion. These techniques utilize synthetic methods to introduce occlusive elements and simulate occlusion scenarios observed in real-world settings. Through these techniques, the model is exposed to a diverse set of samples featuring occlusion during the training process, thereby enhancing its robustness in complex environments. This data augmentation strategy aids the model in understanding and adapting to occluded conditions for keypoint prediction. By incorporating diverse occlusion patterns, the model can learn a broad range of scenarios, thus improving its generalization ability in practical applications. This method not only helps enhance the model's performance in occluded scenes but also provides comprehensive training to boost its adaptability to complex situations. Feature-discrimination-based methods utilize attention mechanisms and similar techniques to reduce interference features. By strengthening features associated with the target person and suppressing those related to non-target individuals, these methods effectively mitigate the interference caused by feature confusion. These methods rely on mechanisms, such as attention, to selectively emphasize relevant features, thereby allowing the model to focus on distinguishing the keypoint features of the target person from those of interfering individuals. By enhancing the discriminative power of features belonging to the target individual, the model becomes adept at navigating scenarios where feature confusion is prevalent. Methods based on human body structure priors optimize occluded poses by leveraging prior knowledge of the human body structure. The use of human body structure priors is particularly effective in providing valuable information about the structural aspects of the human body. These priors serve as constraints that improve the robustness of the model during the inference process. By incorporating these priors, the model is further informed about the expected configuration of body parts, even in the presence of occlusion. This prior knowledge helps guide the model's predictions and ensures that the estimated poses adhere closely to anatomically plausible configurations. A comparative analysis is also conducted to highlight the strengths and limitations of each method in handling occlusion. This paper also discusses the challenges inherent to occluded pose estimation and offers some directions for future research in this area.

Key words: human pose estimation(HPE); occlusion; data augmentation; human structure a priori; insufficient occlusion labeling data

0 引 言

人体姿态估计(human pose estimation,HPE)一直是计算机视觉领域的热门研究方向,其主要目标是准确定位人体的关键点,如手腕和眼睛等部位。该基础任务是许多计算机视觉应用的基础,包括人类动作识别(Yan 等,2018;Crasto 等,2019;Shi 等,2019)、人机交互(Xu 等,2020)、行人重新识别(Li 等,2021;Xuan 和 Zhang,2021;Ye 等,2022)、人体解析(Liang 等,2015,2019;Gong 等,2018;Luo 等,2018;Liu 等,2022)、视频监控和动画生成等领域。

近年来,人体姿态估计取得了显著的进展,这主要归功于丰富的数据和深度卷积神经网络(deep convolutional neural network,DCNN)的不断进步。现有的多人人体姿态估计方法中,可以分为自顶向下和自底向上两种不同的类别。自底向上的方法,如 DeepCut(Pishchulin 等,2016)、OpenPose(Cao 等,2017)、Pifpaf(Kreiss 等,2019)、HigherHrnet(Cheng 等,2020)、SWAHR(scale-weight adaptive heatmap regression)(Luo 等,2021)以及 DEKR(disentangled keypoint regression)(Geng 等,2021),首先检测图像中的所有关键点,然后使用分组算法将检测到的关键点与目标人物进行匹配;而自顶向下的方法,如 Hourglass(Newell 等,2016)、PyraNet(Yang 等,2017)、CPN(cascaded pyramid network)(Chen 等,2018)、Corss-Stage(杨兴明 等,2019)则是首先检测图像中的人,将人体区域从图中裁剪出来,输入人体检测器进行每个人的关键点估计。

尽管近些年取得了不错的进展,但人体姿态估计仍然是一项具有挑战性的任务,特别是在面对复杂的姿态、关键点尺度的变化、遮挡和其他因素时。其中值得注意的是,当前基于热图的方法,如 HRNet(high-resolution net)(Sun 等,2019)、DARK(distribution-aware coordinate representation of keypoint)(Zhang 等,2020)、Litepose(Wang 等,2021)、PPNet(parallel pyramid network)(Zhao 等,2021)Poseur(Mao 等,2022)、Posetrans(Jiang 等,2022)等在遇到遮挡时都存在严重的性能下降,这仍然是该领域的一个关键挑战。

然而,当前虽然有一些方法尝试着解决遮挡所带来的问题,但由于遮挡方向在研究中相对较少且发展不够充分,因此遮挡问题依然存在,并且仍然只受到了相对较少的关注。

孔英会等人(2023)详细总结了近几年二维人体姿态估计相关发展内容,包括单人姿态估计、多人姿态估计以及轻量级人体姿态估计等。但其中未提及关于遮挡人体姿态估计的内容。因此,为了总结和讨论近年来遮挡人体姿态估计的发展,本文对 2018 年后涌现的代表性方法进行了分类、评述、总结和展望。同时,针对为何遮挡问题具有挑战性这一问题进行了探讨。首先,介绍了人体姿态估计的定义、遮挡问题的挑战以及用于评测的数据集。然后,对基于数据增广的预处理、基于特征区分的结构设计和基于人体先验的结果优化 3 类方法进行了概述。其中重点介绍了当前主流的基于热图的自顶向下人体姿态估计检测方法。接着,从性能评测的角度对主流方法进行了比较,并解释了为何遮挡问题具有挑战性。最后,对人体姿态估计领域未来的发展趋势进行了讨论。

1 问题定义及评测数据库

1.1 人体姿态估计的定义

人体姿态估计的目标是在输入图像中准确地估计人体的关节位置,如眼睛、鼻子、手腕等。按照自顶向下的方法,人体姿态估计的流程通常包括以下几个阶段:1)人体检测。首先需要判断输入图像中是否存在人体。通过应用人体检测算法,可以定位图像中的人体区域。一旦检测到人体,就可以对人体区域进行裁剪和仿射变换,进而得到规定尺寸的仅包含单个人体的输入图像。2)人体关键点检测。通过应用深度学习模型提取图像特征,并进行特征识别,可以推断出人体关键点的位置。一种常见的方法是使用热图(heatmaps)来表示每个关键点的概率分布。在热图中,每个像素对应于图像上的一个位置,并且具有一个与关键点相关的概率值。通常,通过选择具有最大概率值的位置来确定每个关键点

的最终位置。为了提高精度，还可以对最大概率值进行子像素级的微调。

1.2 遮挡人体姿态估计的主要挑战

本文在自顶向下的框架下，探讨了人体姿态估计面临遮挡情况时的挑战，主要讨论在给定人体检测框准确的前提下，人体关键点检测在遮挡情况下的困难：1）标注的遮挡数据不足。标注遮挡数据相对来说更具挑战性。当前的人体姿态估计数据集大部分关键点是可见的，只有很少一部分数据涵盖了遮挡情况并进行了标注。因此，在训练模型时，缺乏带有遮挡并进行标注的数据，这导致模型在面对遮挡时缺乏鲁棒性。2）特征混淆。自顶向下的人体姿态估计方法，通常是根据检测框从图像中裁剪出检测框区域中的人，从而进行关键点预测。如果存在遮挡情况，检测框中通常会包含除目标人物以外的其他人物，这些人物通常会对目标人物的关键点预测造成干扰。这是因为目标人和干扰人之间特征的高相似度所导致的，模型难以有效区分二者之间的特征，从而导致关键点的混淆。3）推理困难。大面积的遮挡会导致上下文信息和人体结构信息的丢失。这使得模型无法从相邻关键点中获取足够的上下文信息来推断出准确的位置，从而导致关键点的缺失或姿态异常。图1展示了遮挡情况下人体姿态估计的主要挑战。

(a) 关键点交换　　(b) 关键点漏检

图1　遮挡人体姿态估计挑战可视化图

Fig. 1　Illustration of occlusion human pose estimation visualization images ((a) keypoint swap; (b) keypoint miss)

1.3 评测数据库

1）MSCOCO（Microsoft common objects in context）（Lin等，2014）数据集是目前最广泛使用的人体姿态估计数据集之一。该数据集包含超过20 000幅图像和25 000个人体实例，并对其中的17个关键点进行了标注。MSCOCO训练数据集（train2017）包含57 000幅图像和150 000个人体实例。MSCOCO验证数据集（val2017）和测试数据集（test-dev2017）分别包含5 000幅图像和20 000幅图像。对于MSCOCO数据集中的每个对象，关键点的ground truth标注具有如下形式：$[x_1, y_1, v_1, \cdots, x_k, y_k, v_k]$，其中，$x$，$y$表示关键点的位置，$v$表示可见性标志，$v=0$表示未标注，$v=1$表示标注且可见，$v=2$表示标注且不可见，$k$代表第$k$个关键点。

2）OCHuman（occluded human）（Zhang等，2019）数据集包含4 731幅图像和8 110个人体实例。每个人体实例都严重遮挡了一个或多个其他人体。OCHuman数据集仅用于评估。其注释格式与MSCOCO数据集类似。在使用MSCOCO数据集进行训练后，可以使用OCHuman数据集进行评估。通常认为OCHuman数据集是与人体实例遮挡相关的最具挑战性的数据集。它包含了许多遮挡的关键点，非常适合用于评估遮挡算法的有效性。

3）CrowdPose（efficient crowded scenes pose estimation）（Li等，2019）数据集包含20 000幅图像和80 000个人体标签。对于每一个人体实例，标注了14个关键点。训练数据集包含12 K幅图像，测试数据集包含8 K幅图像。CrowdPose数据集涵盖了许多复杂场景，包括拥挤场景、遮挡、多样的姿态以及背景复杂的情况。因此，CrowdPose数据集具有挑战性，有助于评估算法在真实世界场景中的鲁棒性和泛化能力。它十分适合对于遮挡算法的评价。

4）MSCOCO-RE。由于MSCOCO数据集标注的遮挡关键点较少，因此为了全面评估算法在解决遮挡问题方面的能力，本文对MSCOCO验证集进行了额外的标注，增加了一些更具挑战但可预测的遮挡关键点。为了准确地标注MSCOCO数据集中的遮挡关键点，招募了3名志愿者来手动标记被遮挡的关键点。为确保准确性，进行了交叉标注，即至少有两名志愿者对同一图像进行标注。如果两个注释之间存在差异，则重新对图像进行标注。最终，通过取两个注释的平均值来确定最终的关键点位置。标注完成后，相较于MSCOCO原标注，额外添加了5 961个遮挡关键点，详细信息参见表1。

新增遮挡关键点的可视化图如图2所示，其中左图为原MSCOCO数据集的标注，右图红色的点为新标注的遮挡关键点。MSCOCO-RE数据集地址：https://github.com/whffams2/COCO-RE。

表 1　MSCOCO-RE val2017 详细信息

Table 1　Detail information of MSCOCO-RE val2017

关键点名称	增加数量	关键点名称	增加数量
鼻子	118	左手腕	529
左眼	124	右手腕	499
右眼	123	左髋	422
左耳	435	右髋	408
右耳	409	左膝	381
左肩	154	右膝	393
右肩	138	左脚踝	343
左肘	561	右脚踝	335
右肘	589		

图 2　MSCOCO-RE 可视化图

Fig. 2　Illustration of MSCOCO-RE visualization images

2　遮挡人体姿态估计方法进展

根据神经网络包含的训练数据、模型结构以及输出结果，本文将解决遮挡方法细分为基于数据增广的预处理、基于特征区分的结构设计和基于人体先验的结果优化 3 类，如图 3 所示。

数据增广的方法旨在通过增加遮挡样本来提高模型的鲁棒性。传统的数据增广方式虽然可以应对图像形变、缩放等干扰，但在面对遮挡情况时，模型的性能仍会显著下降。因此，针对遮挡的特殊数据增广方法是缓解模型在面对遮挡时性能下降的一种途径。有一些方法尝试通过人为遮挡图像中的关键点来模拟遮挡情况。此外，一些方法通过生成对抗的方式生成具有挑战性的遮挡样本，以缓解遮挡造成的性能下降。

神经网络特征处理事关模型性能好坏，在遮挡情况下，由于目标人和干扰人之间具有高相似度的特征，常常会导致特征混淆，使得模型无法区分目标人和干扰人的特征，进而导致关键点的混淆。因此，一些方法尝试通过设计特征区分的网络结构来解决遮挡问题。常见的方法之一是采用注意力机制，该

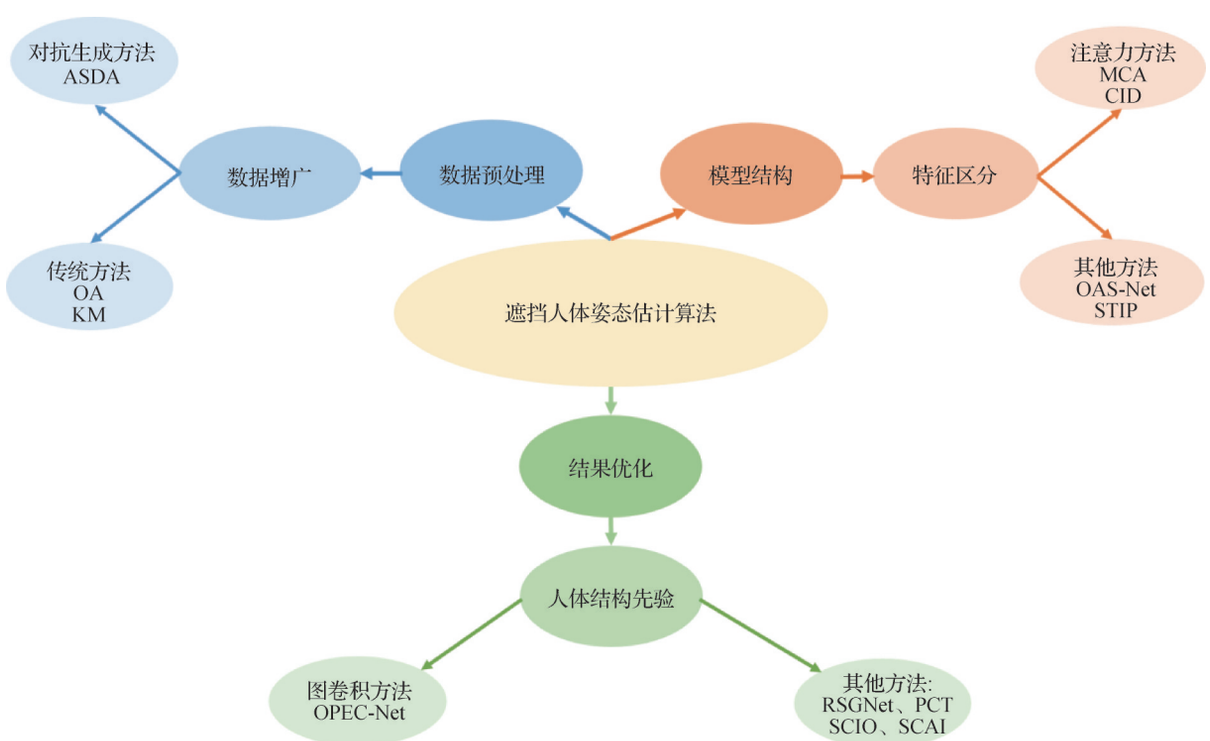

图 3　遮挡算法分类

Fig. 3　Illustration of occlusion methods classification

机制可以有效增强目标特征并抑制非目标特征，从而使模型能够更好地学习目标任务，以缓解遮挡对性能影响。

人体结构的先验知识可以有效弥补因遮挡而导致的人体结构信息丢失。同时，人体结构的先验知识还能够有效抑制异常姿态的生成。基于人体先验的结果优化方法尝试利用人体结构的先验知识，以限制或推断遮挡关键点的位置，对遮挡的关键点进行细化。

2.1 基于数据增广的预处理方法

数据增广（data augmentation）是一种常见且通用的训练预处理策略。它可以缓解神经网络训练中的过拟合问题，增强神经网络的泛化能力，并提高神经网络的性能。一般而言，数据增广指的是在保持原始数据的语义标签不变的情况下，对原始数据进行修改，以增加遮挡训练样本的数量。

2.1.1 传统数据增广预处理

在人体姿态估计的任务中，常用的数据增广方法包括随机翻转（概率为50%）、随机旋转（旋转角度范围为[-45°, 45°]）、随机缩放（范围为[0.65, 1.35]），但这些方法不能够减轻遮挡的影响。因此，Huang等人（2020）受信息丢弃法的启发，提出了OA（occlusion augmentation）方法。OA方法通过随机丢弃图像中人体关键点的外观信息，并保留响应图的监督，从而使得神经网络更加关注被丢弃的关键点。这样可以增强神经网络的鲁棒性，从而可以预测被遮挡的关键点。OA方法可视化图如图4所示。

具体而言，在给定的N个标注的关键点$\{k_1, k_2, \cdots, k_N\}$中，随机选择一个关键点$k$，将$k$周围区域的像素设置为零，并将该区域形状设置为圆。圆的半径r在$[0.1w, 0.2w]$的范围内随机选择，其中w代表网络输入的宽度。由于遮挡形状过于简单，为了防止神经网络过度拟合这种线索，遮挡区域的中心至少包含一个关键点。因此，遮挡区域的中心位置会根据随机向量$\boldsymbol{\delta}$进行移动。

Ke等人（2018）提出KM（keypoint mask training）方法。该方法的数据增广方式包含两种：1）随机遮挡（random occlusion，RO）。在背景上随机选择一个补丁，并确保该补丁不包含任何关键点。然后，使用该背景补丁覆盖关键点，以模拟关键点被遮挡的情况。这样的样本有助于学习如何从遮挡中恢复关键点。2）随机粘贴（random paste，RP）。随机选择一个关键点，并将其粘贴到附近的背景上。这种操作旨在模拟密集人群中存在多个关键点的情况。图5展示了这两种方法的可视化图。总的来说，KM方法有助于解决遮挡、比例变化和复杂背景等挑战，使得神经网络能够更好地处理被遮挡或部分可见的关键点。

图4 OA方法可视化图

Fig. 4 Illustration of the method of occlusion augmentation

(a) 随机遮挡图 (b) 随机粘贴图

图5 KM方法可视化图

Fig. 5 Illustration of the method of keypoint mask training ((a) RO visualization image; (b) RP visualization image)

2.1.2 基于生成对抗网络数据增广预处理

传统的数据增广预处理方法（OA、KM等）可以帮助神经网络缓解由于轻微遮挡所带来的性能下降，但无法提高其在面对挑战性案例时的鲁棒性。因此，引入生成对抗网络（generative adversarial network，GAN）来生成更具迷惑性的样本，即更真实的

样本。GAN是深度学习领域中重要的生成模型，其通过同时训练生成器和鉴别器两个网络，并在极小极大算法中进行竞争。这种对抗性的方法避免了一些传统生成模型在实际应用中遇到的困难，例如随机生成的样本不符合实际情况。

Bin等人(2020)提出了ASDA(adversarial semantic data augmentation)方法。由于传统的数据增广预处理方法(OA、KM等)具有随机性，不能保证生成最真实的遮挡样本，因此，ASDA尝试利用生成对抗网络来动态生成粘贴参数，从而模拟最真实的遮挡样本。具体来说，ASDA首先利用人体解析(human parsing)的方法构建语义池(semantic part pool)，然后将一些部分(如右鞋和右腿)组合起来，形成具有更高语义颗粒度的新部分，并将其放入语义池。最后，过滤掉尺寸低于35×35像素且语义较低的部件。然后随机选择人体部件，将这些部件粘贴到原始图像中。由于随机粘贴不具备生成最真实遮挡样本的能力，因此ASDA利用空间变换器网络(spatial Transformer network，STN)以对抗性方式来优化语义部分的参数。STN充当生成器，生成最混乱的变换，以增加姿态估计网络的损失。而目标网络则作为鉴别器，从粘贴的语义增广中学习，其过程如图6所示。

图6　ASDA过程图

Fig. 6　Illustration of ASDA method process diagram

通过对抗性训练，生成器网络学习如何生成更具挑战性的合成图像，以迷惑判别器网络。这种对抗性训练促使生成器网络产生更真实和多样的合成图像，从而提供了更广泛的姿态估计情况，使姿态估计算法具有更好的鲁棒性和泛化能力。使用增广后的数据进行训练可以改善姿态估计算法的准确性和鲁棒性。

2.2　基于特征区分的结构设计方法

注意力机制已经证明在计算机视觉任务中非常有效，如姿态估计、图像分类和动作识别等。它通过选择性地关注相关信息区域和抑制不相关的全局信息，能够有效增强目标信息、抑制非目标信息。因此，人体姿态估计中，有一些方法尝试设计特征区分的结构，利用注意力机制来增强遮挡情况中目标人的特征、抑制非目标人的特征，从而缓解遮挡带来的混淆。

Chu等人(2017)提出了MCA(multi-context attention)方法，这是首个尝试设计特征区分结构，利用注意力机制来解决人体姿态估计任务的方法。该模型能够在不同环境下对人体姿态进行建模。其利用多环境注意力机制，使模型能够自适应地关注不同环境中与人体姿态相关的特征。多环境注意力机制由两个关键组件组成：环境感知模块和注意力融合模块。环境感知模块用于提取不同环境下的特征表示，以捕捉环境的上下文信息；注意力融合模块则将不同环境的特征进行加权融合，以产生最终的姿态估计结果。该方法利用注意力机制来使模型更加关注到人体，增强人体特征，削弱环境特征的影响。使用多环境注意力可以改善姿态估计算法在不同环境下的准确性和鲁棒性。模型能够更好地适应不同环境的变化，并更准确地预测人体的姿态。

Zhou等人(2020)设计了遮挡感知的孪生网络，Siamese框架OAS-Net(occlusion-aware siamese network)。此外，还提出了两种机制，一种是特征清除，另一种是特征重建。OAS-Net利用数据集中的遮挡注释生成遮挡感知的注意力图，来作为特征擦除和特征重建的基础。特征擦除是指去除那些被污染的特征图，获得一个更清晰的表示。在遮挡感知注意图的指导下，由遮挡引起的模糊特征被去除，并获得相对更干净的特征。然而，特征擦除在擦除被遮挡污染的特征的同时也会擦除其他有用的信息，这就导致了特征擦除后的特征图中缺少语义信息或人体结构信息。因此，引入了特征重建，它不仅可以恢复被删除的一些有用信息，还可以获得更新的信息来替换被遮挡的特征。为了给特征重建提供足够的指导，使被遮挡的分支模仿未被遮挡的分支的行为，OAS-Net采用了孪生(Siamese)框架。Siamese框架有两个分支，模型权重在两个分支之间共享。第1个分支将人工遮挡作为输入，并对不太可信的信息进行编码；第2个分支将没有人工遮挡的原始图像作为输入，提供更可信的信息。为了更紧密地拉动

两个分支，使用了最优传输与附加掩码作为正则化代替对应基方法。使用遮挡感知的孪生网络可以显著改善姿态估计算法在遮挡情况下的准确性和鲁棒性。该方法能够有效地处理部分遮挡的情况，并生成更准确的姿态估计结果。

Wang等人(2021)提出了一种基于实例自适应的语义感知迁移(semantic-aware transfer with instance-adaptive parsing, STIP)方法。STIP首先通过语义感知机制增强了像素级表示的判别能力，智能地决定要增强哪些像素和要添加哪些语义嵌入，从而缓解了缺失关键点检测的问题。其次，提出了一种新的实例自适应的句法分析方法，该方法不采用具有固定参数的标准回归器，而是动态生成特定于实例的参数，以减少歧义标记带来的不利影响。从而来解决由于遮挡所带来的特征混淆问题。语义感知的迁移学习策略结合实例自适应解析模块使模型更好地理解复杂场景背景，并能够更准确地估计遮挡环境中个体的姿态。

Wang等人(2022)则是提出了CID (contextual instance decoupling)方法，利用通道注意力，基于人的实例相似性来区分不同个体之间的特征。主要过程为，首先从目标人中心提取目标人实例特征，其次利用通道注意力加强提取的目标人实例特征，并以此抑制干扰人的特征。

2.3 基于人体先验的结果优化方法

人体结构具有丰富的知识，作为人体姿态估计的先验知识，能够有效地帮助模型学习人体结构的信息，以及各关键点之间的联系。这种先验知识使得模型能够更好地推理出被遮挡的关键点，优化结果。同时也有助于解决人体姿态估计中的长距离问题。

Qiu等人(2020)提出了OPEC-Net (occluded pose estimation and correction)。OPEC-Net尝试利用图卷积神经网络(graph convolution neural network, GCNN)根据人体结构进行建模，利用人体结构信息来解决优化结果。OPEC-Net利用一个图像引导的渐进式GCN模块IGP-GCN(image guided progressive)，以全面理解图像背景和姿态结构，并从推断的角度估计不可见关节的位置。该方法的核心思想是，利用上下文信息和空间约束来推测被遮挡的关键点位置。具体来说，人体结构提供了关节之间的基本约束信息。最初的姿态和级联特征被送入IGP-GCN。IGP-GCN作为修正模块，提供了一种明确的身体结构信息建模方式，有利于修正关节。除此之外，遮挡关节的另一个重要线索是其相关的图像背景。OPEC-Net同时结合了人体结构信息和图像背景信息，这两种信息相互促进，共同预测被遮挡的关节。OPEC-Net能够有效地预测遮挡关键点的位置，并提高整体姿态估计的准确性，在处理拥挤场景中的姿态估计问题方面具有潜力。

Dai等人(2021)提出了RSGNet，旨在通过利用骨架关系来解决遮挡问题。RSGNet主要包含两个部分，TRP(target-aware relation parser)和SGM (skeleton graph machine)。TRP用于对所有预测关节的关系进行建模，从而产生目标编码。这个新的管道在很大程度上缓解模型在看到具有完全不同标签的相同关节时的混乱(例如，同一只手存在于两个边界框中)。SGM是一种基于骨骼常识知识的建模方法，利用基于骨架的人体姿态表示，其中人体被表示为相互连接的关节图。然后，构建了一个基于关系的骨架图网络。目的是在人体结构约束下估计目标姿态。从推理的角度来看，这种基于骨架的约束可以帮助处理拥挤场景中的挑战。RSGNet能够更好地处理遮挡和复杂的空间关系，即使在具有挑战性的拥挤场景中也能获得更准确的姿态估计。

Kan等人(2022)认为人体姿态在不同身体部位之间由于生物学约束而展现出强烈的组内结构相关性和关键点之间的空间耦合。这种组内结构相关性可以用来提高人体姿态估计的准确性和鲁棒性。因此，提出了SCIO(self-constrained inference optimization)方法。具体来说，SCIO根据人体生物结构将关键点分为不同的组。在每个组内，关键点进一步分为两个子集，即高置信度的基本关键点和低置信度的终端关键点(如人的眼睛、肩膀等关键点比较容易预测，则被认为是高置信度关键点；而人的耳朵、手腕等难以预测，则被认为是低置信度关键点)。同时，还提出了一个自我约束的预测—验证网络，该网络在关键点子集之间进行前向和后向预测。在姿态估计中，当真值不可用时，无法验证预测的结果是否准确。而预测网络如果能够成功预测，则验证网络可以用做前向姿态预测的准确性验证模块。在推断阶段，它可以用来指导对低置信度关键点的姿势估计结果进行局部优化，以高置信度关键点上的自我约束损失作为目标函数，从而显著提高了人体姿态

估计的性能。

Geng等人(2023)认为虽然通过身体关节的坐标向量或它们的热图嵌入来表示人体姿势很容易，但由于缺乏身体关节之间的依赖关系建模，导致了不现实的姿态估计结果。因此，Geng等人(2023)提出一种结构化表示方法，称为PCT(pose as compositional tokens)，旨在探索关节之间的依赖关系。PCT通过M个离散标记来表示一个姿势，每个标记都代表一个具有若干相互依赖关节的子结构。这种组合设计使其能够以较低的成本实现小的重构误差，并且将姿态估计视为一个分类任务。具体来说，PCT学习了一个分类器，用于从图像中预测M个标记的类别。然后，使用预先训练好的解码网络来从这些标记中恢复姿势，无需进一步的后处理。结果表明，PCT在一般情况下，与现有方法相比，能够实现更好或相媲美的姿势估计结果，同时在遮挡出现的情况下仍能良好地工作，而遮挡在实际应用中是普遍存在的。

Kan等人(2023)提出的SCAI(self-correctable and adaptable inference)是SCIO的延伸，出自同一作者。Kan等人(2023)认为已学习的网络缺乏表征预测误差、从测试样本生成反馈信息并为每个独立的测试样本实时纠正预测误差的能力，这导致了泛化性能的降低。因此，提出了SCAI来解决网络预测的泛化性挑战。简单来说，SCAI在SCIO中预测—验证网络的基础上增加了一个反馈网络，预测结果映射到原始输入域并与原始输入进行比较。反馈网络用来回馈验证网络，以便验证网络能够通过反馈网络的反馈信息来纠正预测的错误结果。该方法具有较强的泛化能力，在人体姿态估计任务中具有较大的性能提升，面对遮挡情景具有较强的鲁棒性。

3 代表性遮挡人体姿态估计算法性能对比

表2列举了2018年以来的代表性遮挡人体姿态估计方法，以及这些方法在MSCOCO、OCHuman、CrowdPose和MSCOCO-RE数据集上的平均精度均值(mean average precision，mAP)。mAP能够可靠地衡量算法的性能，广泛用于人体姿态估计算法的性能评价。

1)MSCOCO数据集以及MSCOCO-RE数据集结果分析。对于MSCOCO数据集来说，数据中的人分散较为稀疏，存在较少的遮挡关键点。从表2的结果可知，基于数据增广预处理的方法对于性能的提升有限，最大提升由方法RC所得，mAP提升了0.5%。结果表明基于数据增广预处理的方法只能稍微缓解遮挡所带来的困难。基于特征区分的结构设计方法相较于数据增广预处理的方法更具有性能优势，提升更加明显。最大提升由STIP方法所得，mAP提升了1.2%。这表明基于特征区分的结构设计方法能够缓解遮挡所带来的性能下降。基于人体先验的结果优化方法在所有方法中取得了最高的性能，有较大的提升。最大提升由方法SCAI所得，mAP提升了5.1%。结果表明，基于人体先验的结果优化方法能够有效缓解人体姿态估计遮挡问题带来的性能下降。

MSCOCO-RE数据集相较于MSCOCO数据集增加了更具有挑战性的遮挡关键点，大多数为物体遮挡人的关键点。

各类方法在MSCOCO-RE数据集中的性能相较于MSCOCO数据集均有下降，说明MSCOCO-RE中标注的遮挡的关键点更加难以预测，也证明了MSCOCO-RE数据集中所标注的遮挡关键点的有效性。在3类方法中，基于数据增广预处理的方法在MSCOCO-RE数据集中提升最少，最大提升由方法RC(RP、ASDA)所得，mAP提升了0.2%。说明数据增广预处理的方法在遮挡较少的数据集中，提升非常有限。基于特征区分的结构设计方法在MSCOCO-RE数据集中提升明显。最大提升由方法STIP所得，mAP提升了1%。说明基于特征区分的方法能够有效解决遮挡问题。基于人体先验的结果优化方法在MSCOCO-RE数据集提升最为明显，最大提升由PCT所得，mAP提升了5.1%。说明基于人体先验的结果优化方法不仅适合人体姿态估计遮挡类问题，同时也适合一般性问题，具有较强的鲁棒性。

2)OCHuman数据集结果分析。对于OCHuman数据集来说，数据中人与人之间存在大量重叠情形，但样本中人数较少，通常为1～3个。从表2的结果可知，基于数据增广的预处理方法在OCHuman数据集中提升明显，最大提升由OA方法所得，mAP提升1.6%。结果表明，基于数据增广的预处理方法，在人与人重叠情况的遮挡数据集中，提升更加明显。

表2　代表性遮挡人体姿态估计算法总结

Table 2　Summary of representative occluded human pose estimation algorithms

	方法	Backbone	图像尺寸/像素	mAP/%			
				MSCOCO	OCHuman	CrowdPose	MSCOCO-RE
Baseline	HRNet(Sun等,2019)	HRNet-W32	256×192	74.4	60.7	71.7	71.3
		HRNet-W48	384×288	76.3 75.5[t]	64.8	73.9	73.0
数据增广	OA(Huang等,2020)	HRNet-W32	256×192	74.7	62.3	73.5	71.4
	KM(Ke等,2018)	HRNet-W32	256×192	74.6	61.2	73.4	71.2
	RC(Ke等,2018)	HRNet-W32	256×192	74.9	61.5	73.3	71.5
	RP(Ke等,2018)	HRNet-W32	256×192	74.8	61.5	73.3	71.5
	ASDA(Bin等,2020)	HRNet-W32	256×192	74.8	61.8	73.5	71.5
特征区分	OAS-Net(Zhou等,2020)	HRNet-W32	256×192	75.0	*	*	*
	STIP(Wang等,2021)	HRNet-W32	256×192	75.6	64.2	**74.1**	72.3
人体结构先验	OPEC-Net(Qiu等,2020)	ResNet-101	256×192	73.9[t]	*	70.6	*
	RSGNet(Dai等,2021)	HRNet-W32	256×192	75.7	64.1	73.6	72.3
	SCIO(Kan等,2022)	HRNet-W48	384×288	79.5	*	*	*
	PCT(Geng等,2023)	Swin-Huge	256×256	78.3	**70.2**	*	**76.4**
	SCAI(Kan等,2023)	HRNet-W48	384×288	**80.6**[t]	*	*	*

注:加粗字体表示各列最优结果,"-"表示无数据,"*"表示未开源代码或模型无法公平对比。MSCOCO一栏中,无标记的数值为MSCOCO vol2017上的评测结果;带上标t的数值为MSCOCO test-dev 2017上的评测结果,MSCOCO test-dev 2017与MSCOCO val2017相比,预测更加困难,一般同一模型在MSCOCO val2017上的性能要高于MSCOCO test-dev2017上的性能。

基于特征区分的结构设计方法在OCHuman数据集中提升较大,最大提升由方法STIP所得,mAP提升了3.5%。结果表明,基于特征区分的结构设计方法,不仅在遮挡数据较少的MSCOCO数据集上有明显提升,在人与人重叠情况严重的遮挡数据集OCHuman也有明显提升,说明此类方法能够有效解决人体姿态估计中的遮挡问题。基于人体先验的结果优化方法,在3类方法中提升最大,最大提升为PCT方法所得,mAP提升了9.5%。结果表明,基于人体先验的结果优化方法,在面对遮挡难题时,具有鲁棒性。

3)CrowdPose数据集结果分析。对于CrowdPose数据集来说,数据中人与人之间存在较多遮挡,且单个样本包含较多的人,通常在3个以上。从表2的结果可知,基于数据增广的预处理方法在CrowdPose数据集上的性能普遍提升,最大提升由OA和ASDA所得,mAP提升了1.8%。基于特征区分的结构设计方法在CrwodPose数据集中性能明显提升。最大提升由STIP方法所得,mAP提升了2.4%。基于人体先验的结果优化方法,在CrowdPose数据集上具有明显提升,最大提升由RSGNet所得,mAP提升了2.0%。

综上所述,基于数据增广的预处理方法在具有较多遮挡样本的数据集,例如OCHuman和CrowdPose等数据集具有较大的提升,能够有效提升模型面对遮挡时的性能,而在MSCOCO以及MSCOCO-RE这类遮挡关键点较少的数据集中,则没有较大的提升。基于特征区分的结构设计方法在MSCOCO、MSCOCO-RE、OCHuman和CrowdPose数据集性能均有不错的提升,说明此类方法不仅能够解决遮挡问题,还能够有效解决人体姿态估计一般类问题,例如复杂背景、复杂人体姿态等。基于人体先验的结果优化方法,在3类方法中表现最为优秀,也是近年来探索最多的方法,此类方法具有较大的潜力,无论是面对遮挡问题还是一般性问题都能够具有较大的提升。

表2的代表性方法在各数据集中使用的人体检测框结果依次为：1）MSCOCO：使用在MSCOCO val2017数据集上具有56.4%mAP的人物检测器检测所得。2）OCHuman：Ground Truth Test检测框。3）在CrowdPose test数据集上具有71.0%mAP人体检测器检测所得。4）MSCOCO-RE：使用在MSCOCO val2017数据集上具有56.4%mAP的人物检测器检测所得。

此外，表2的代表性方法是在基于不同开源库上进行复现。主要分为两类：1）基于开放库MMpose（MMPose Contributors，2020）复现。主要包括数据增广类别的方法。2）基于原作者代码库复现。主要包括特征区分以及人体结构先验的方法。不同框架对实验结果精准度差距较小，影响可忽略不计。

训练策略如下：

1）MSCOCO。所有实验均在train2017上进行训练。并遵循HRNet（Sun等，2019）将人体检测框的高度和宽度扩展为固定的长宽比，即高宽比为4比3。并根据检测框对原始图像进行裁剪，将其调整为固定尺寸为256×192像素（除OPEC-Net为320×256像素）。除了数据增广方法外，其余方法均采用常规数据增广策略，如随机旋转（[−45°，45°]）、随机缩放（[0.5，1.5]），以及翻转和半身数据增广等。所有实验均使用Adam（adaptive moment estimation optimizer）优化器。基础学习率设定为5×10^{-4}，在第170次和第200次历时中分别降为5×10^{-5}和5×10^{-6}。本文将总的迭代次数设置为210次。

2）OCHuman。该数据仅为测试数据集，于MSCOCO数据集训练结束，直接测试。

3）CrowdPose。实验均在CrowdPose训练集上进行训练，其余设置均与MSCOCO一致。

4）MSCOCO-RE。该数据仅为测试数据集，于MSCOCO数据集训练结束，直接测试。

测试策略如下：

采取自顶向下的方式，先检测再预测。

1）MSCOCO。在MSCOCO val2017数据集进行测试，并使用在MSCOCO val2017数据集上具有56.4%mAP的人物检测器检测所得检测框进行测试，本文通过对从原始图像和翻转的图像中获得的热图进行平均，计算出最终的目标热图。然后根据热图中最高响应与次高响应之间的方向上偏移1/4来确定关节的坐标。

2）OCHuman。采取OCHuman Ground Truth Test检测框进行测试，其余设置与MSCOCO一致。

3）CrowdPose。在CrowdPose test数据集上进行测试，并使用在CrowdPose test数据集上具有71.0%mAP的人物检测器检测所得检测框进行测试，其余设置与MSCOCO一致。

4 结　语

作为一种人体姿态分析技术，人体姿态估计取得了相当的进展，但仍无法完全满足实际应用需求。在遮挡情况下，人体姿态估计的性能不可避免地会下降，而相关方法仍存在探索空间。主要有以下几个关键问题值得关注：

1）人体姿态估计数据库遮挡标注数据不足的问题。遮挡标注数据的缺乏，仍然是遮挡人体姿态估计领域的主要问题。尽管可以利用数据增广或是利用对抗生成网络来增加遮挡的样本，但实际情况下，这些遮挡的样本并不能够较好地模拟真实的情况，从实验结果也能看出，基于数据增广的预处理方法效果，并不如预期。因此，从提升数据增广真实性出发，是一个值得考虑的问题。目前的遮挡数据增广，绝大多数参数都是人为设定，其随机性往往会使得其他的一些关键点被遮挡亦或不符合现实遮挡情况，这使得生成的遮挡样本不仅无法帮助网络学习更好的表示，反而给网络带来了更大的干扰。而当前热门基于技术Diffusion的AI绘画则可以作为考虑的方法之一，其强大的能力，能够生成更加符合真实世界的样本。亦或可以优化对抗生成网络，使得网络能够学习到更加合理的信息，从而生成更符合真实情况的样本。

2）遮挡情况下特征混淆问题。遮挡在不同情况下，其遮挡程度亦不相同。尽管现有的方法在解决特征混淆问题上取得了一定的进展，但是当遮挡情况严重时，算法的表现则可能会出现较大的下降。因此，如何优化特征区分方法在遮挡严重情景下的性能依然值得探究。后续，可以考虑通过联合人体解析方法来辅助特征区分。人体解析方法可以将目标人体大部分部位定位出来，通过该方法可以进一步利用注意力机制或其他方式来增强目标人的特征。

3）遮挡情况下人体结构信息缺失问题。在现实

情况下,人体可能通常会被物体所遮挡。由于物体的遮挡,会使得人体结构信息不完整。缺乏完整的结构信息,往往会使得模型推理困难。因此,常用的解决方法是引入人体结构先验知识来弥补所缺失的人体结构信息。当前绝大部分方法都是基于人体结构利用GCN进行建模。大部分的GCN通常只能捕获相邻关键点的信息,而在现实情况中,人类的姿态各式各样、千变万化,因此还有可能会出现长距离问题,需要通过合理设计GCN来捕获更长距离关键点信息来解决。

尽管目前有一些方法尝试来解决遮挡人体姿态估计这个难题,但是相较于常规的人体姿态估计方法,解决遮挡的方法相对有限。而解决人体姿态估计中的遮挡问题具有重要的价值,不仅能够使得人体姿态估计技术贴近于现实情景,同时还能够增加人体姿态估计的实际应用价值。因此,希望遮挡人体姿态估计能够吸引更多人的关注,从而为该领域的发展迈出有意义的一步。

参考文献(References)

Bin Y R, Cao X, Chen X Y, Ge Y H, Tai Y, Wang C J, Li J L, Huang F Y, Gao C X and Sang N. 2020. Adversarial semantic data augmentation for human pose estimation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 606-622 [DOI: 10.1007/978-3-030-58529-7_36]

Cao Z, Simon T, Wei S E and Sheikh Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1302-1310 [DOI: 10.1109/cvpr.2017.143]

Chen Y L, Wang Z C, Peng Y X, Zhang Z Q, Yu G and Sun J. 2018. Cascaded pyramid network for multi-person pose estimation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7103-7112 [DOI: 10.1109/cvpr.2018.00742]

Cheng B W, Xiao B, Wang J D, Shi H H, Huang T S and Zhang L. 2020. HigherHRNet: scale-aware representation learning for bottom-up human pose estimation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5385-5394 [DOI: 10.1109/cvpr42600.2020.00543]

Chu X, Yang W, Ouyang W L, Ma C, Yuille A L and Wang X G. 2017. Multi-context attention for human pose estimation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5669-5678 [DOI: 10.1109/cvpr.2017.601]

Crasto N, Weinzaepfel P, Alahari K and Schmid C. 2019. MARS: motion-augmented RGB stream for action recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7874-7883 [DOI: 10.1109/cvpr.2019.00807]

Dai Y, Wang X H, Gao L L, Song J K and Shen H T. 2021. RSGNet: relation based skeleton graph network for crowded scenes pose estimation//Proceedings of the 35th AAAI Conference on Artificial Intelligence.Virtual Event, AAAI: 1193-1200 [DOI: 10.1609/aaai.v35i2.16206]

Geng Z G, Sun K, Xiao B, Zhang Z X and Wang J D. 2021. Bottom-up human pose estimation via disentangled keypoint regression//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2021: 14671-14681: [DOI: 10.1109/cvpr46437.2021.01444]

Geng Z G, Wang C Y, Wei Y X, Liu Z, Li H Q and Hu H. 2023. Human pose as compositional tokens//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 660-671 [DOI: 10.1109/cvpr52729.2023.00071]

Gong K, Liang X D, Li Y C, Chen Y M, Yang M and Lin L. 2018. Instance-level human parsing via part grouping network//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 805-822 [DOI: 10.1007/978-3-030-01225-0_47]

Huang J J, Shan Z G, Cai Y H, Guo F, Ye Y, Chen X Z, Zhu Z, Huang G, Lu J W and Du D L. 2020. Joint COCO and LVIS workshop at ECCV 2020: COCO keypoint challenge track technical report: UDP++ [EB/OL]. [2023-06-10]. http://presentations. cocodataset. org. s3. amazonaws. com/ECCV20/keypoints/UDP.pdf

Jiang W T, Jin S, Liu W T, Qian C, Luo P and Liu S. 2022. PoseTrans: a simple yet effective pose transformation augmentation for human pose estimation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 643-659 [DOI: 10.1007/978-3-031-20065-6_37]

Kan Z H, Chen S S, Li Z and He Z H. 2022. Self-constrained inference optimization on structural groups for human pose estimation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 729-745 [DOI: 10.1007/978-3-031-20065-6_42]

Kan Z H, Chen S S, Zhang C, Tang Y S and He Z H. 2023. Self-correctable and adaptable inference for generalizable human pose estimation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 5537-5546 [DOI: 10.1109/cvpr52729.2023.00536]

Ke L P, Chang M C, Qi H G and Lyu S W. 2018. Multi-scale structure-aware network for human pose estimation//Proceedings of the 15th

European Conference on Computer Vision. Munich, Germany: Springer: 731-746 [DOI: 10.1007/978-3-030-01216-8_44]

Kong Y H, Qin Y F and Zhang K. 2023. Deep learning based two-dimension human pose estimation: a critical analysis. Journal of Image and Graphics, 28(7): 1965-1989 (孔英会, 秦胤峰, 张珂. 2023. 深度学习二维人体姿态估计方法综述. 中国图象图形学报, 28(7): 1965-1989) [DOI: 10.11834/jig.220436]

Kreiss S, Bertoni L and Alahi A. 2019. PifPaf: composite fields for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11969-11978 [DOI: 10.1109/cvpr.2019.01225]

Li J F, Wang C, Zhu H, Mao Y H, Fang H S and Lu C W. 2019. CrowdPose: efficient crowded scenes pose estimation and a new benchmark//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10855-10864 [DOI: 10.1109/cvpr.2019.01112]

Li Y L, He J F, Zhang T Z, Liu X, Zhang Y D and Wu F. 2021. Diverse part discovery: occluded person re-identification with part-aware Transformer//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2897-2906 [DOI: 10.1109/cvpr46437.2021.00292]

Liang X D, Gong K, Shen X H and Lin L. 2019. Look into person: joint body parsing and pose estimation network and a new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(4): 871-885 [DOI: 10.1109/tpami.2018.2820063]

Liang X D, Xu C Y, Shen X H, Yang J C, Liu S, Tang J H, Lin L and Yan S C. 2015. Human parsing with contextualized convolutional neural network//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1386-1394 [DOI: 10.1109/iccv.2015.163]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu K L, Choi O, Wang J M and Hwang W. 2022. CDGNet: class distribution guided network for human parsing//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4463-4472 [DOI: 10.1109/cvpr52688.2022.00443]

Luo Y W, Zheng Z D, Zheng L, Guan T, Yu J Q and Yang Y. 2018. Macro-micro adversarial network for human parsing//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 424-440 [DOI: 10.1007/978-3-030-01240-3_26]

Luo Z X, Wang Z C, Huang Y, Wang L, Tan T N and Zhou E J. 2021. Rethinking the heatmap regression for bottom-up human pose estimation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13259-13268 [DOI: 10.1109/cvpr46437.2021.01306]

Mao W A, Ge Y T, Shen C H, Tian Z, Wang X L, Wang Z B and Van Den Hengel A. 2022. Poseur: direct human pose regression with Transformers//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 72-88 [DOI: 10.1007/978-3-031-20068-7_5]

MMPose Contributors. 2020. OpenMMLab pose estimation toolbox and benchmark [EB/OL]. [2023-06-10]. https://github.com/open-mmlab/mmpose

Newell A, Yang K Y and Deng J. 2016. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 483-499 [DOI: 10.1007/978-3-319-46484-8_29]

Pishchulin L, Insafutdinov E, Tang S Y, Andres B, Andriluka M, Gehler P and Schiele B. 2016. DeepCut: joint subset partition and labeling for multi person pose estimation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4929-4937. [DOI: 10.1109/cvpr.2016.533]

Qiu L T, Zhang X Y, Li Y R, Li G B, Wu X J, Xiong Z X, Han X G and Cui S G. 2020. Peeking into occluded joints: a novel framework for crowd pose estimation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 488-504 [DOI: 10.1007/978-3-030-58529-7_29]

Shi L, Zhang Y F, Cheng J and Lu H Q. 2019. Skeleton-based action recognition with directed graph neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7904-7913 [DOI: 10.1109/cvpr.2019.00810]

Sun K, Xiao B, Liu D and Wang J D. 2019. Deep high-resolution representation learning for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5686-5696 [DOI: 10.1109/cvpr.2019.00584]

Wang D K and Zhang S L. 2022. Contextual instance decoupling for robust multi-person pose estimation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11050-11058 [DOI: 10.1109/cvpr52688.2022.01078]

Wang X H, Gao L L, Dai Y, Zhou Y X and Song J K. 2021. Semantic-aware transfer with instance-adaptive parsing for crowded scenes pose estimation//Proceedings of the 29th ACM International Conference on Multimedia. [s. l.]: ACM: 686-694 [DOI: 10.1145/3474085.3475233]

Wang Y H, Li M Y, Cai H, Chen W M and Han S. 2022. Lite pose: efficient architecture design for 2D human pose estimation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13116-13126 [DOI: 10.1109/cvpr52688.2022.01278]

Xu C J, Yu X Y and Wang Z G. 2020. Multi-view human pose estimation in human-robot interaction//Proceedings of the 46th Annual

Conference of the IEEE Industrial Electronics Society. Singapore, Singapore: IEEE: 4769-4775 [DOI: 10.1109/iecon43393.2020.9255211]

Xuan S Y and Zhang S L. 2021. Intra-inter camera similarity for unsupervised person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11921-11930 [DOI: 10.1109/CVPR46437.2021.01175]

Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: #12328 [DOI: 10.1609/aaai.v32i1.12328]

Yang W, Li S, Ouyang W L, Li H S and Wang X G. 2017. Learning feature pyramids for human pose estimation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1290-1299 [DOI: 10.1109/iccv.2017.144]

Yang X M, Zhou Y H, Zhang S R, Wu K W and Sun Y X. 2019. Human pose estimation based on cross-stage structure. Journal of Image and Graphics, 24(10): 1692-1702 (杨兴明, 周亚辉, 张顺然, 吴克伟, 孙永宣. 2019. 跨阶段结构下的人体姿态估计. 中国图象图形学报, 24(10): 1692-1702) [DOI: 10.11834/jig.190028]

Ye M, Shen J B, Lin G J, Xiang T, Shao L and Hoi S C H. 2022. Deep learning for person Re-identification: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 2872-2893 [DOI: 10.1109/TPAMI.2021.3054775]

Zhang F, Zhu X T, Dai H B, Ye M and Zhu C. 2020. Distribution-aware coordinate representation for human pose estimation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7091-7100 [DOI: 10.1109/cvpr42600.2020.00712]

Zhang S H, Li R L, Dong X, Rosin P, Cai Z X, Han X, Yang D C, Huang H Z and Hu S M. 2019. Pose2seg: detection free human instance segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 889-898 [DOI: 10.1109/cvpr.2019.00098]

Zhao L, Wang N N, Gong C, Yang J and Gao X B. 2021. Estimating human pose efficiently by parallel pyramid networks. IEEE Transactions on Image Processing, 30: 6785-6800 [DOI: 10.1109/TIP.2021.3097836]

Zhou L, Chen Y Y, Gao Y Z, Wang J Q and Lu H Q. 2020. Occlusion-aware Siamese network for human pose estimation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 396-412 [DOI: 10.1007/978-3-030-58565-5_24]

作者简介

徐琳皓,男,硕士研究生,主要研究方向为计算机视觉、人体姿态估计。E-mail:linh@njust.edu.cn

赵林,通信作者,男,副教授,主要研究方向为计算机视觉、人体姿态估计和跟踪、三维人体形状重建和异常检测。E-mail:linzhao@njust.edu.cn

孙辛欣,女,副教授,主要研究方向为智能交互设计、产品创新设计。E-mail:sunxinxinde@126.com

颜克冬,男,副教授,主要研究方向为数据挖掘、神经网络形式化验证、优化理论及应用。E-mail:yan@njust.edu.cn

李广宇,男,讲师,主要研究方向为智能汽车。E-mail:guangyu.li2017@njust.edu.cn