计算机科学
COMPUTER SCIENCE

深度学习方法在二维人体姿态估计的研究进展

张国平, 马楠, 贯怀光, 吴祉璇

引用本文

张国平, 马楠, 贯怀光, 吴祉璇.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 Guo-ping, MA Nan, Guan Huai-guang, WU Zhi-xuan. Research Progress of Deep Learning Methods in Two-dimensional Human Pose Estimation [J]. Computer Science, 2022, 49(12): 219-228.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于TPH-YOLOv5和小样本学习的害虫识别方法

Pest Identification Method Based on TPH-YOLOv5 Algorithm and Small Sample Learning

计算机科学, 2022, 49(12): 257-263. https://doi.org/10.11896/jsjkx.221000203

基于改进Sigmoid卷积神经网络的手写体数字识别

Handwritten Numeral Recognition Based on Improved Sigmoid Convolutional Neural Network

计算机科学, 2022, 49(12): 244-249. https://doi.org/10.11896/jsjkx.211000179

基于反事实思考的视觉问答方法

Visual Question Answering Method Based on Counterfactual Thinking

计算机科学, 2022, 49(12): 229-235. https://doi.org/10.11896/jsjkx.220600038

面向深度卷积神经网络的小目标检测算法综述

Small Object Detection Based on Deep Convolutional Neural Networks:A Review

计算机科学, 2022, 49(12): 205-218. https://doi.org/10.11896/jsjkx.220500260

用于协同过滤的序列解耦变分自编码器

Disentangled Sequential Variational Autoencoder for Collaborative Filtering

计算机科学, 2022, 49(12): 163-169. https://doi.org/10.11896/jsjkx.211200080

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900041

深度学习方法在二维人体姿态估计的研究进展

张国平[1,3]　马　楠[2]　贯怀光[1]　吴祉璇[1]
1 北京联合大学北京市信息服务工程重点实验室　北京 100101
2 北京工业大学信息学部　北京 100124
3 北京联合大学机器人学院　北京 100101
(diffzhang@163.com)

摘　要　人体姿态估计的任务是对图像或视频中的人体关键点进行定位和检测,其一直是计算机视觉领域的热点研究方向之一,也是计算机理解人类行为动作的关键一步。近年来,图像和视频中的二维人体姿态关键点预测在许多领域有着广泛的应用,二维人体姿态估计利用深度学习强大的图像特征提取能力,提升了其鲁棒性、准确性并缩短了处理时间,而且表现效果远超传统方法。根据二维人体姿态研究对象数量的不同,可将其分为单人以及多人姿态估计方法。针对单人姿态估计,根据提取到的关键点表示的不同,可采用基于直接预测人体坐标点的坐标回归方法,以及预测人体关键点高斯分布的基于热图的检测方法;针对多人姿态估计,可采用的方法分为解决多人到单人过程的自顶向下方法,以及直接处理多人关键点的自底向上方法。根据现有的人体姿态估计方法对其进行总结,说明网络结构的内部机制及执行过程,并对常用的数据集、评价指标进行分析,最后阐述当前面临的问题及未来发展趋势。

关键词:二维人体姿态估计;深度学习;单人姿态估计;多人姿态估计;评价指标
中图法分类号　TP391

Research Progress of Deep Learning Methods in Two-dimensional Human Pose Estimation

ZHANG Guo-ping[1,3],MA Nan[2],Guan Huai-guang[1] and WU Zhi-xuan[1]
1 Beijing Key Laboratory of Information Service Engineering,Beijing Union University,Beijing 100101,China
2 Department of Information Science,Beijing University of Technology,Beijing 100124,China
3 College of Robotics,Beijing Union University,Beijing 100101,China

Abstract　The task of human pose estimation is to locate and detect the key points of human body in images or videos. It has always been one of the hot research directions in the field of computer vision,and it is also a key step for computers to understand human actions. In recent years,it has wide application for predicting the poses of two-dimensional human body key points in images and videos. Using the powerful image feature extraction capabilities of deep learning,two-dimensional human pose estimation has been improved in robustness,accuracy,and processing time,and the performance effect is far beyond traditional methods. According to the different number of objects in the two-dimensional human body pose,it can be divided into single-person and multi-person pose estimation methods. For single-person pose estimation,according to the different representations of the extracted key points,coordinate regression methods based on the direct prediction of human coordinate points and heat map detection methods based on predicting the Gaussian distribution of human key points can be used. In multi-person pose estimation,it is divided into the top-down method which solves the process from multiple people to a single person,and a bottom-up method that directly deals with the key points of multiple people. Based on the existing estimation methods of human body posture,this paper analyzes the internal mechanism of the network structure,analyzes the commonly used datasets and evaluation indicators,and elaborates the current problems and future development trends.

Keywords　Two-dimensional human pose estimation,Deep learning,Single-person pose estimation,Multi-person pose estimation,Evaluation metrics

---

到稿日期:2021-09-06　返修日期:2022-01-30
基金项目:国家自然科学基金(61871038,61931012)
This work was supported by the National Natural Science Foundation of China(61871038,61931012).
通信作者:马楠(manan123@bjut.edu.cn)

1 引言

人体姿态估计是计算机视觉中一个重要的研究课题，其任务是对图像或视频中人体关键点位置进行估计。近年来，人体姿态估计在无人驾驶、智能交互、行为动作分析、娱乐影视、视频监控等领域有着广泛的应用。现实中，人体有各种复杂的姿态，人体姿态之间的遮挡、复杂背景和视角变化等因素使得检测效果不理想，因此，人体姿态估计任务具有一定的挑战性。随着 GPU 算力的不断提升，深度学习技术得到迅速发展，尤其是在目标分类和目标检测领域[1-2]，已超过人类水平，基于深度学习的人体姿态估计方法表现效果逐年提升，极大地促进了人体姿态估计领域的发展。

传统的人体姿态估计方法是使用图结构模型，以满足结构化的预测任务。其通过可变形的部件集合来表示人体，各个部件采用模板匹配的方式来检测，然后计算出各个部件之间的空间连接关系，从而构成完整的人体姿态。但传统方法主要提取基于人工设计的特征，导致特征提取网络不能充分利用图像的特征，从而使得人体姿态估计方法的检测效果无法满足现实的需求。而基于深度学习的人体姿态估计方法则是利用深度卷积神经网络强大的特征提取能力，有效地提取图像的低级和高级的特征，进行准确的人体姿态估计，在现实中有着良好的效果。

本文主要对基于深度学习的二维人体姿态估计方法进行分析，在比较各种单人和多人姿态估计方法的基础上，系统地汇总了近年来的研究进展和应用；此外，还说明了视频中的人体姿态估计方法。

2 基于深度学习的二维人体姿态估计方法

随着 Toshev 等[3]提出的 DeepPose 方法识别效果远优于传统方法之后，很多人开始将人体姿态估计的研究从传统方法转向深度学习方法。当前的人体姿态估计方法普遍采用深度卷积神经网络进行提取特征，以取代人工特征提取。根据检测人数的不同，可将其分为单人姿态估计方法和多人姿态估计方法。

2.1 单人姿态估计方法

单人姿态估计方法是在输入的单人图像中定位出人体关键点的位置。根据关键点表示的不同，可分为基于坐标回归和基于热图的检测方法。基于坐标回归的方法是在图像中直接预测人体关键点的坐标；基于热图检测的方法旨在预测关键点的大概位置，并由一系列热图表示，关键点的位置由以关键点位置为中心的二维高斯分布表示。单人姿态估计识别流程如图 1 所示。

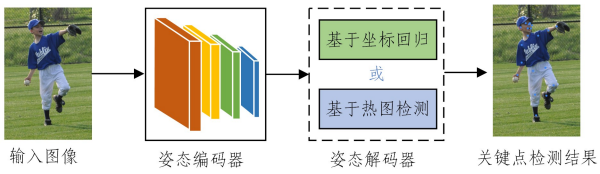


图 1 单人姿态估计识别流程

Fig. 1 Process of single-person pose estimation and recognition

2.1.1 基于坐标回归方法

早期的人体姿态估计方法大多是以局部特征为关键点建模，但算法表示能力具有局限性，只能为人体关键点之间所有关系的部分子集进行建模，忽略了关键点整体上下文的关系。为了解决这个问题，DeepPose 使用坐标回归的深度学习网络进行人体姿态估计，采用整体的方式来预测人体关键点位置并且使用级联回归优化关键点的精确位置，提高了对人体关键点的预测准确率。使用 DeepPose 等前馈神经网络通常能够学习到输入的空间特征，但是无法在输出空间对依赖关系进行显式建模，Carreira 等[4]提出一个通用的坐标回归框架，使用 GoogleNet 作为骨干网络，对输出特征与输入特征进行联合学习，同时对输入特征和输出特征进行建模，此外还引入自顶向下的反馈机制，这种机制不直接预测输出目标，而是在前馈过程中预测当前估计的偏差并使用迭代反馈修正预测结果。这种方法被称为迭代误差反馈机制（Iterative Error Feedback，IEF），它是一种自我修正的模型，可以获得更好的预期效果。该方法的运行流程如图 2 所示。

(1)利用统计的平均值初始化 y_0。

(2)前向模型 f 输入：t 时刻的 RGB 图像 x_t 和视觉表示映射 g 连接构成的增强输入空间，其中 g 以 y_t 作为输入。

(3)前向模型 f 输出：预测一个修正值 ε_t，目的是让预测值 y_t 与真实值 y 更加接近。

(4)更新估计结果：$y_{t+1}=y_t+\varepsilon_t$。

(5)与图像堆叠在一起，产生新的输入 $x_{t+1}=I\oplus g(y_t)$，经过迭代得出结果。

其中，I 代表输入图像；x_t 表示 t 时刻的 RGB 图像；y_0 代表初始化的关键点值；y_t 代表当前的关键点值；ε_t 代表输出的一个修正值；y_{t+1} 代表新的关键点值；f 代表卷积神经网络；g 代表高斯函数，用于进行高斯转换。

为了充分利用人体姿态内部的结构信息，Shuang 等[5]提出了一个结构化感知回归方法，这种方法采用重新参数化骨骼代替关键点来表达人体姿态，骨骼具有人体的直观性和稳定性，能更好地表达人体姿态结构。此外，Luvizon 等[6]提出一种使用 Soft-Argmax 函数的端到端回归方法，此方法将特征图转换为关键点坐标，可以直接将上下文信息用于人体姿态预测中，以提高识别精度。Mao 等[7]借助转换器中的注意力机制提出的方法能够自适应地关注与目标关键点最相关的特征，在很大程度上解决了以前基于回归的方法的特征错位问题，并显著提高了性能。

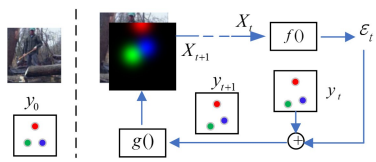


图 2 IEF 网络架构[4]

Fig. 2 IEF network architecture[4]

2.1.2 基于热图检测方法

目前已经有很多基于热图检测的人体姿态估计方法[8-9]。基于回归的检测方法通常对复杂的背景和身体遮挡比较敏感，为了获得更多的监督信息，国内外学者大多采用热图来

表示关键点的真实值，这种方法有助于深度卷积神经网络的训练。Lifshitz 等[10]通过关键点检测器和推理关键点关系联合生成最终的姿态估计结果，在关键点检测时，该方法使用空洞卷积和反卷积层来提高模型输出的特征图分辨率，使得在没有增加模型参数量的基础上能有效地扩大卷积感受野，提高关键点热图检测的准确率。

网络结构设计一直是重要的研究课题之一，高效的网络结构不仅参数量小，收敛速度快，而且易于预测关键点位置，对深度学习网络的实际应用具有重要意义。也有不少学者对应用于人体姿态估计的深度卷积神经网络的网络结构进行优化和改进。Wei 等[11]提出了一个卷积姿态机(Convolutional Pose Machines，CPM)网络，该方法使用卷积神经网络学习图像纹理信息和空间信息。在此之前，很多学者使用卷积神经网络来提取图像的纹理信息，使用图模型或者其他模型来表达身体各个部位在空间上的关系，没有同时利用这两种信息，而 Wei 等使用卷积神经网络同时学习这两种特征，使得学习效果更好，并且有助于端到端学习。经过不断的细化，最终将得到一个比较准确的关键点热图的预测值。随着残差网络的提出，Newell 等[12]设计了一个堆叠沙漏网络，这个网络也是多阶段结构，由多个堆叠起来的沙漏结构组成，每个沙漏结构都包含从高分辨率到低分辨率和从低分辨率到高分辨率的过程，以估计不同尺度的人体姿态关键点热图信息。在此基础上，Yang 等[13]加入金字塔残差模块，以增强深度卷积神经网络对尺度变化的鲁棒性。此外，Chu 等[14]改进残差单元，使分支滤波器具有更大感受视野，同时使用改进后的残差单元结构学习到多尺度特征，进一步提升了关键点热图预测的准确率。Wang 等[15]提出一种学习随机混合图像的数据增强方法，提高了姿态估计在各种损坏数据(如模糊和像素化)的情况下关键点检测的鲁棒性。Groos 等[16]提出了一种新的卷积神经网络架构 EfficientPose，此架构通过限制内存占用和计算成本，快速生成关键点热图，以支持边缘设备上的实际关键点检测应用程序。

2.1.3 小结

基于坐标回归方法是直接回归人体关键点坐标，而人体姿态估计任务是一个高度非线性的问题，因此该类方法具有明显的局限性，泛化能力较差。基于热图检测的方法采用改进的深度学习网络生成准确的热图，以更好地表示人体部位的关键点信息，具有更好的鲁棒性。因此，当前主要研究基于热图检测的方法。表 1 列出了这两种方法的优缺点。

表 1 单人姿态估计两类方法的比较

Table 1 Comparison of two methods of single-person pose estimation

方法	优点	缺点
基于坐标回归方法	检测速度快，可以实现端到端学习	回归学习映射困难，精确度不高
基于热图检测方法	识别精确度较高，适应于多种复杂环境，热图容易可视化分析	预测精确度依赖热图分辨率，并且计算量较大，检测速度慢

2.2 多人姿态估计方法

单人姿态估计是定位图像中单个人的关键点位置，而多人姿态估计由于事先不知道图像中存在的人数，因此需要处理检测和定位两个任务。多人姿态估计方法可分为由解决多人到解决每人的自顶向下方法，以及直接针对多人关键点分组的自底向上方法两大类。自顶向下方法识别流程如图 3(a)所示，先检测出图像中所有目标对象，然后将目标对象从原图中裁剪出来重置大小后，输入到网络中进行姿态估计，相当于把多人姿态估计处理成多个单人姿态估计；而自底向上方法识别流程如图 3(b)所示，首先检测出图像中所有目标对象的关键点信息，再对提取的关键点信息进行筛选分组，最终得到每个人的姿态。

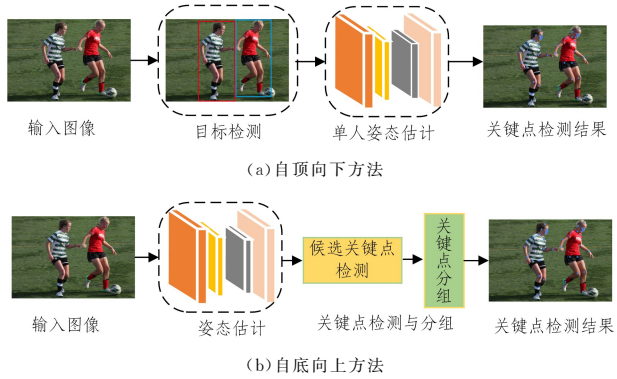


图 3 多人姿态估计识别流程

Fig. 3 Process of multi-person pose estimation and recognition

2.2.1 自顶向下方法

自顶向下的人体姿态估计方法[17-20]有两个最重要的组成部分，分别是人体候选区域检测器和单人姿态估计器，研究者大多基于现有的人体候选区域检测器进行人体姿态估计的研究。Iqbal 等[21]提出了一种基于 Faster R-CNN 检测器的多人的姿态估计方法，将一组在图像中检测到的候选关键点构造成一个完全连接的图，然后使用整数线性规划解决关键点与人的关联关系，该方法能很好地解决人体姿态间的遮挡问题。Fang 等[22]采用 Faster R-CNN 作为检测器，并使用堆叠沙漏网络作为人体姿态估计器，该网络模型进行人体姿态估计时，通过对图像进行微小的变换和修剪来提高下一阶段人体姿态估计的效果，可以减少许多对人体关键点的定位错误，以消除人体定位产生的冗余姿态，提高了性能；并通过适当地增强数据来提高人体姿态估计方法适应复杂场景的人体候选区域的定位效果。Papandreou 等[23]提出了一种高效的多人姿态估计方法，在第一阶段采用 Faster R-CNN 作为人体检测方法，检测出图像中的多个人，并对其进行裁剪；第二阶段采用全卷积 ResNet 网络对上一阶段检测出的人体进行关键点和偏移量的预测；最后通过融合预测到的关键点和偏移量来得到人体关键点的精确位置。Huang 等[24]提出了一种以 Inception 网络为骨干的由粗到精的细化网络结构，该网络结构在多个级别上进行监督学习，目的是得到粗略和精细化的预测，进而精确估计人体姿态。Kumar 等[25]提出了一种快速有效的联合人员检测和姿态估计方法，该方法引入了一种两阶段的评估策略，该策略更适合自动驾驶，并且与目前最先进的评估方法相比有着显著的性能改善。

为了解决多人姿态估计中面临的关键点遮挡、关键点不可见和复杂姿态等难题，Chen 等[26]提出了一种级联金字塔网络(Cascaded Pyramid Network，CPN)，如图 4 所示，该网络

包括 GlobalNet 和 RefineNet 两个子网络。GlobalNet 使用特征金字塔网络提取含有简单关键点的不同尺度特征，但无法准确识别被遮挡或者视觉不可见的关键点；RefineNet 将 GlobalNet 网络得到的不同分辨率下的特征表示融合到一起，使得被遮挡的关键点通过融合后的上下文信息被准确定位。Su 等[27]设计了两个新颖的模块来增强人体姿态估计的信息，分别是通道混合模块(Channel Shuffle Module，CSM)和空间注意力残差模块(Spatial Channel-wise Attention Residual Bottleneck，SCARB)。CSM 通过对不同层次的特征图进行混合操作，促进了金字塔特征图之间跨信道的信息通信；SCARB 利用注意力机制增强残差单元，在空间和通道上下文中增强显示特征图的信息，实现通道方向和空间信息增强，以在遮挡场景下更好地进行多人姿态估计。Qiu 等[28]提出了一个新的网络 OPEC-Net(Occluded Pose Estimation and Correction Networks)，同时提出了一个渐进式图像引导的 GCN(Graph Convolutional Networks)模块，该模块可以对图像上下文和姿态结构进行全面理解，很好地解决了姿态遮挡问题。在人体姿态估计中。想要更加准确地定位人体关键点，需要同时利用高分辨率的特征表示和低分辨率的语义信息。Sun 等[29]提出了一种新的网络结构 HRNet(High-Resolution Network)，如图 5 所示，该方法始终保持主干网络为高分辨率进行特征提取，使得预测关键点更准确。Wang 等[30]提出了一种注意力改进网络 HR-ARNet 来增强人体姿态估计的多尺度特征融合，采用通道和空间注意机制来强化重要特征并抑制不必要的特征，解决关键点之间不一致的问题。

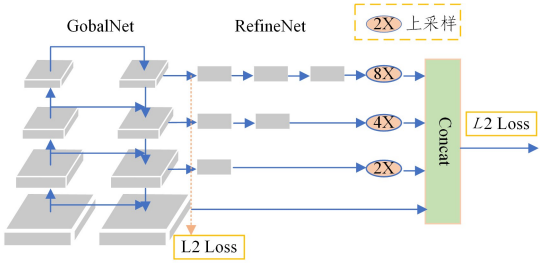


图 4 CPN 网络结构[26]

Fig. 4 CPN network architecture[26]

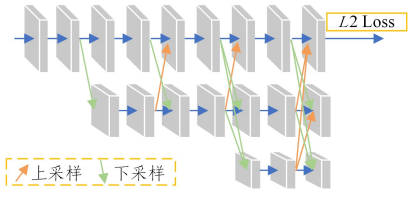


图 5 HRNet 网络结构[29]

Fig. 5 HRNet network architecture[29]

随着基于图像的二维人体姿态估计方法的不断发展，视频中的二维人体姿态估计方法识别效果也在逐步提升。与基于图像的人体姿态估计方法不同，视频中的人体姿态估计方法利用视频帧之间的时序信息来处理视频帧间的运动模糊、遮挡和外观变化等问题，同时也会与追踪任务相结合。因此，直接使用基于图像的人体姿态估计方法来处理视频中的人体姿态往往不能达到最佳的效果。基于视频的人体姿态估计方法首先检测每帧中的人和关键点，然后跨帧传播边界框和关键点。Girdhar 等[31]提出了一个 3D Mask R-CNN 网络，该网络全部采用 3D 卷积来检测视频片段中每个人的二维关键点热图表示，然后使用关键点追踪器通过比较检测到的边界距离来连接和预测姿态。此外，Wang 等[32]改进基于视频片段的追踪器，通过设计的 3D 卷积层以扩展 HRNet 网络学习二维关键点之间的对应关系，并且设计时空融合算法来估计最佳关键点输出，该方法可以在具有严重遮挡的复杂场景下进行人体姿态估计和追踪。与基于视频片段的检测不同，Xiao 等[33]提出的方法是建立在单帧检测的基础上，并利用基于光流的时间姿态相似性来关联不同帧的关键点，但该方法会使得视频帧中的检测缺失。为了处理单帧中的缺失检测问题，Bao 等[34]设计了一个网络，将基于图像的检测器与人物位置检测器相结合，以补偿缺失的检测框，同时该方法还引入了分层姿态引导的图卷积网络，该网络利用人体结构关系来增强人体表示和数据关联。为了更加有效地进行数据关联，Ruan 等[35]设计了一个端到端的姿态引导网络，用于多人姿态追踪中的数据关联，它联合了学习特征提取、相似性估计和人体分组。Umer 等[36]提出了一个自我监督的关键点对应关系网络，该方法不仅可以恢复丢失的姿态检测，还可以跨帧关联检测到恢复的姿态。与使用高维图像表示来追踪人体不同，Snower 等[37]提出了一个基于转换器的追踪器，它只依赖于 15 个关键点，并且该网络利用二元分类来预测姿态空间是否在时间上跟随另一个姿态，而不需要使用任何光流信息来实时追踪人体关键点，提高了人体姿态识别和追踪的效率。Yang 等[38]采用不同的 ResNet 模块，并将模块中的普通卷积改进为深度可分离卷积，降低了参数量；并且在反转积模块中丰富了特征图的信息，使得网络检测更加精确。

2.2.2 自底向上方法

自底向上的人体姿态估计方法[39-40]，首先使用高效的人体关键点检测方法检测出所有人的关键点，然后使用分组算法对候选关键点进行分组，其中候选关键点分组的准确性是关键。Cao 等[41]提出一种基于 CPM 网络的方法来预测具有部分亲和力场(Part Affinity Fields，PAF)的所有候选关键点信息，可以有效地检测图像中多个人的关键点位置，PAF 结构用于表示关键点的位置和方向，并且可以计算出两个关键点的关联程度。该算法的整个网络结构如图 6 所示，分为关键点生成分支和 PAF 分支两个分支结构，关键点生成分支用来获得置信图，PAF 分支用来获得部分亲和力关系 PAF。网络分为多个阶段，每一个阶段结束时都有中继监督，每一个阶段结束之后，特征图 S，L 都和阶段一中的特征图 F 进行融合。此方法最大的贡献是提出了部分亲和力关系 PAF，从而在保持高精度的同时实现实时的效果。Yu 等[42]提出了一种基于部分亲和力关系 PAF 的改进方法，将离散的关键点分组成个体，并检测关键点之间的向量或连接，但是这种机制存在一个致命的缺陷，即当 PAF 将 n 个关键点与$(n-1)$条线连接时，表示关联所有连接的最小连接数，这表明为了获得整个人体姿态，需要每个连接预测和关键点预测完全正确，这种特性削弱了 PAF 的稳健性和鲁棒性，同时由于关键点在树形结构中逐个连接，如果父连接断开，那么即使子连接和关键点被正确检测，也不会被连接。为了解决这个问题，Yu 等设计一种

冗余的部分亲和力关系，如图 7(b)所示，通过增加关键点间的连接数，使连接变得多余；通过添加这些冗余连接，使得关键点之间的结构从树型转换为图型，提高了连接的容错率，从而使分组准确率更高。为了提高算法在低分辨率和遮挡情况下的识别效果，Kreiss 等[43]设计了一种多人姿态估计方法 PifPaf，该方法采用部分强度场关系(Part Intensity Field，PIF)来定位身体各个部位，同时使用部分亲和力关系 PAF 关联身体部位来分组成完整的人体姿态，复合的 PAF 可以对细粒度的信息进行编码，因此该方法在低分辨率以及遮挡等复杂场景下效果优于其他方法。

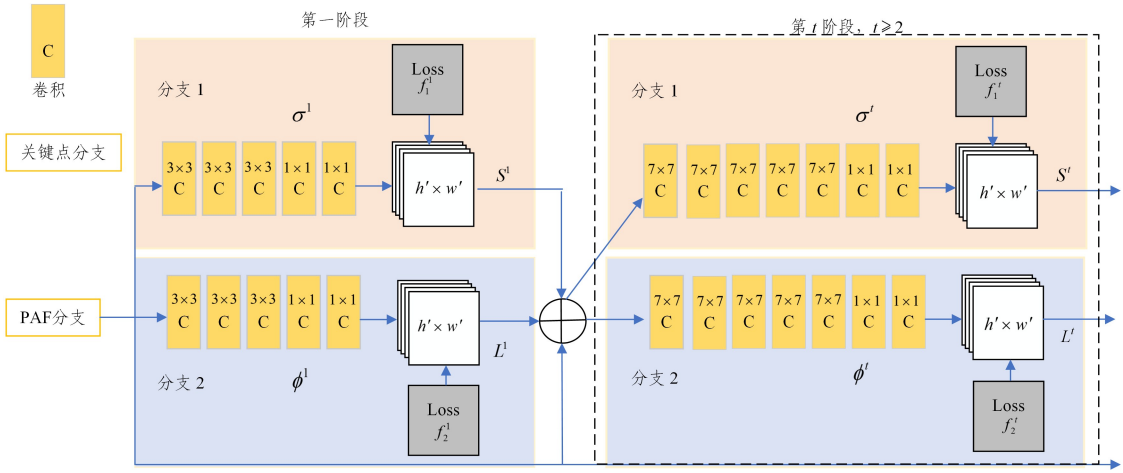
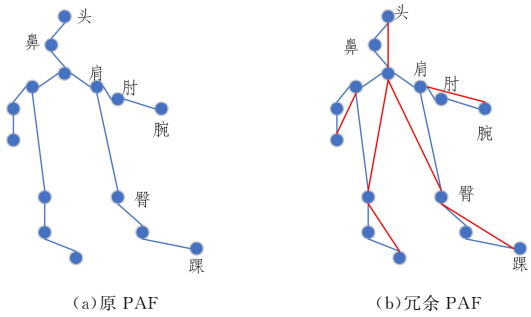


图 6 OpenPose 网络结构[41]

Fig. 6 OpenPose network architecture[41]



(a)原 PAF (b)冗余 PAF

图 7 原 PAF 和冗余 PAF[42]

Fig. 7 Original PAF and redundant PAF[42]

为了提高多人姿态估计的检测效率，国内外学者研究了单阶段多人姿态估计方法，Newell 等[44]提出了一种单阶段网络，同时输出关键点检测和分组分配。该方法提出了关联嵌入(Associative Embedding)网络，该网络为一种表示联合检测分组的新方法，主要思想是把标签和同一组中的其他部分关联到一起，提高了运行的速度，但是由于算法对尺度信息不敏感，使得检测的准确性不高。为了解决多尺度问题，Cheng 等[40]在 HRNet 的基础之上设计网络，通过在 HRNet 网络的最后加入反卷积层以得到更高分辨率的特征图，这有助于提高对小目标的检测，在一定程度上解决了人体尺度多样性的难题。

人体姿态估计用于多任务学习时可以和相关的任务联合学习，这比单个任务学习的识别效果更好。也有相关学者把多任务学习用于人体姿态估计。Papandreou 等[45]提出了一种同时进行姿态估计和实例分割的多任务网络，该网络可以同时预测人体所有关键点的联合热图及其相对位移。Kocabas 等[46]将多任务模型与姿态残差网络(Pose Residual Network，PRN)结合在一起，可以同时处理人体关键点检测和人体分割任务，聚类检测到的关键点信息，最终将关键点分配给人体实例以生成准确的人体姿态。

视频中的人体姿态估计方法是使用单帧姿态估计方法来预测每帧中的所有关键点，然后以时空优化的方式跨帧分配关键点。Insafutdinov 等[47]基于图分割方法扩展了基于图像的自底向上的多人姿态估计方法，然后又构建了一个时空图，在空间和时间上连接候选关键点，将其公式化为线性规划问题。然而，空间和时间图划分通常会导致大量计算。为了解决这个问题，Xiu 等[48]利用姿态流测量不同帧中的姿态距离来追踪同一个人。Zhang 等[49]提出了一种基于姿态估计的多摄像机系统语义同步框架，利用从视频中获得的语义人体姿态估计来对多个摄像机进行时间同步，证明相机帧同步可以提高精度。由于受到 OpenPose 方法的 PAF 结构的启发，一些方法[50-52]利用时间流场来表示不同帧中关键点的传播方向。此外，Jin 等[53]提出一个时空嵌入的关键点关联策略方法，该网络将关键点与嵌入特征相关联以实现时间一致性。Helmstetter 等[54]设计了一种专门针对工艺应用中测量用户需求的移动人体捕获系统，该系统基于立体相机系统和开源的人体姿态估计算法 OpenPose，但相机的定位和标定以及图像的三角化仍然存在挑战，导致手腕关节角度的偏移。Stenum 等[55]使用基于开源的人体姿势估计方法 OpenPose 对测量的时空和矢状运动步态参数与同时记录的健康成人地上行走的 3D 运动捕捉进行了比较，证明参与者相对于相机的位置会影响空间测量。

自顶向下方法中，随着图像中人数的增加，计算成本也会随之显著增加。而自底向上方法与图像中人数无关，因此检测速度较快，但是如果人体之间有较大重叠，对检测到的关键点进行分组时就变得比较困难。因此，想要获得更高的识别精度，就需要基于自顶向下的姿态估计方法进行研究；想要获得更快的识别速度，自底向上的姿态估计方法将是最佳选择，表 2 列出了这两种方法的优缺点。

表 2 多人姿态估计两类方法的比较

Table 2 Comparison of two methods of multi-person pose estimation

方法	优点	缺点
自顶向下方法	检测精确度高，分两步易于训练	随图像中人数增加，计算量也增加，速度较慢
自底向上方法	检测速度较快，与图像中人数无关	检测精确度较低，受环境因素影响较大，在复杂环境中，容易导致分组错误

目前视频中的人体姿态估计方法大多是基于自顶向下方法，这种方法可以获得更高的识别精度，并且可以有效解决遮挡和运动模糊等问题，但是识别速度较慢。多人自底向上的方法检测速度较快，但是识别效果不好，精度较低，而且容易出现分组错误。但是随着计算机算力和内存大小的不断提升，计算速度从硬件上得到一定的缓解，更多的研究利用自顶向下的人体姿态估计方法来提高识别精度。

2.3 人体姿态估计方法应用于其他任务

2.3.1 智能交互

智能交互是一个重要的研究课题，它在图像理解和机器人学习中有大量应用，人体姿态信息对检测到的人和物体之间的交互非常重要。Fang 等[56]提出了一个基于身体部位的注意力模型，专注于关键部位及其对智能交互识别的相关性，在该方法中使用姿态估计器检测人体关键点，然后检测身体部位。Li 等[57]进一步结合视觉外观、空间位置和人体姿态信息进行交互，并且采用姿态估计来预测 17 个关键点以构建空间姿态信息。Wan 等[58]训练 CPN 网络作为姿态估计器，并利用识别的人体姿态全局空间信息作为引导，在语义部分提取局部特征，使得该方法能够对细粒度人体对象交互进行可靠的预测。

2.3.2 行为动作识别

人体行为动作识别是理解动态场景的重要任务，人体姿态的正确检测对于识别人体行为至关重要，如 Luvizon 等[59]提出了一个多任务方法，可以联合处理视频序列的人体姿态估计和动作识别方法。与其不同的是，Du 等[60]提出了一种姿态注意力机制，可以在循环神经网络每个时间步的动作预测中自适应地学习姿态相关特征。Ludwig 等[61]提出使用伪标签作为自监督训练技术以及伪标签的过滤方法，提高模型一致性。随着二维姿态估计的发展，最近的研究[62-63]倾向于研究基于骨架的动作识别。

2.3.3 人体解析

人体解析旨在将人体图像分割成不同的细粒度语义，这是许多高级应用的基础，如人体行为分析、人体重识别和视频监控等。人体姿态可以为身体部位分割提供结构化信息，如 Dong 等[64]提出了一个同时进行人体解析和姿态估计的方法，该方法验证了这两个任务的互补性质可以提高彼此的性能；此外，Liang 等[65]提出了一个联合人体解析和姿态估计的网络来进行有效的上下文建模，该网络可以同时高质量地预测人体解析和人体姿态。

当前很多与人体相关的任务都结合了人体姿态估计进行研究，从而进一步提高识别效果，尤其在智能交互、行为动作识别和人体解析任务中的应用更多，而且实践证明人体姿态估计方法与其他任务联合训练可以互相提高各自的识别效果。

3 常用数据集和评价指标

3.1 常用数据集

数据集是算法研究中重要的一部分，随着深度学习基础网络的不断加深，数据集的大小也在不断增加，只有数据集的数据量足够大、多样性足够丰富，才能训练出准确率高和泛化效果好的算法。近年来，人体姿态估计领域发布了很多数据量比较大的公开数据集，这些数据集的出现促进了人体姿态估计算法的研究与发展。表 3 列出了常用的二维人体姿态估计数据集。

表 3 常用数据集

Table 3 Common data set

数据集	类型	关键点数量	数据大小
LSP[66]	单人	14	≈2 K
MPII[67]	单人/多人	16	≈25 K
COCO[68]	多人	17	>300 K
CrowdPose[69]	多人	14	≈20 K
PennAction[70]	单人	13	≈2 K
J-HMDB[71]	单人	15	≈1 K
PoseTrack[72]	多人	15	≈0.5 K

LSP 是比较早期的数据集，而且只能用于单人姿态估计，标注的关键点也较少，因此现在基本不再使用此数据集。而 MPII 和 COCO 数据集规模较大，因此很多研究者常用这两个数据集训练算法，并且它们都可以用于多人姿态估计算法。CrowdPose 是拥堵场景下的数据集，可以使得在其上训练的算法对拥堵情况有很强的鲁棒性。PennAction，J-HMDB 和 PoseTrack 是视频数据集，主要用于基于视频的时序人体姿态估计研究。这些数据集的提出对人体姿态估计的研究作出了巨大贡献，同时我们也希望有更多新的数据集被提出，以进一步推动人体姿态估计的研究。

3.2 评价指标

评价指标用于衡量人体姿态估计模型的性能，不同的任务类型和数据集具有不同的评价指标。表 4 所列为常用的评价指标。

表 4 评价指标

Table 4 Evaluation metrics

评价指标	释义	典型数据集
PCP	Percentage of Correct Parts	LSP
PCK	Percentage of Correct Keypoints	MPII
AP	Average Precision	MPII，PoseTrack，COCO
OKS	Object Keypoint Similarity	COCO

PCP 是早期广泛使用的评价指标，其值越高，表示模型性能越好。如果预测的关键点和真实的肢体关键点之间的距离小于肢体尺寸的一半，则认为肢体被检测到，但是该评价指标有个缺点，即当肢体具有较小的阈值时，对较短肢体不友好。其计算式如(1)所示：

$$PCP=\frac{\sum_i \delta(d_i<kL_{\text{norm}})\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{1}$$

其中，d_i 代表预测的关键点和真实的关键点之间的欧氏距离，L_{norm} 代表标准肢体长度，k 是比例，一般为 0.5。

PCK 多用于 MPII 数据集，如果预测的关键点和真实

关键点之间的距离小于设定的阈值，则可以认为检测到的关键点是正确的。由于较短的肢体具有较小的躯干，因此 PCK 可以解决 PCP 无法很好地检测出较短肢体的问题。在 MPII 数据集上，将阀值设置为头部长度 L_{head}，并将其称为 PCKh，其计算式如(2)所示：

$$PCKh=\frac{\sum_i \delta(d_i<kL_{\text{head}})\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{2}$$

AP 即平均精度，是衡量关键点准确率的指标，如式(3)所示，而平均精度均值 mAP(Mean Average Precision)是计算所有关键点 AP 的均值。OKS 即目标关键点相似度，主要目的是计算预测的人体关键点和真实的人体关键点之间的相似度，其计算式如(4)所示。

$$AP=\frac{\sum_p \delta(OKS>s)}{\sum_p 1} \tag{3}$$

$$OKS=\frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{4}$$

其中，s 代表体检测框面积的平方根，k_i 代表归一化参数，v_i 代表当前关键点的可见性。

4 人体姿态估计方法性能比较

表 5 列出了基于深度学习的二维单人姿态估计方法在 MPII 数据集上的性能比较，测量标准使用 PCKh@0.5，其中阈值设置为 0.5，从表中可以看出，基于热图检测的方法精确度高于基于坐标回归的方法。

表 5 单人姿态估计方法性能比较

Table 5 Performance comparison of single-person pose estimation methods

方法	类型	头	肩	肘	腕	臀	膝	踝	平均
文献[4]	基于坐标回归方法	95.7	91.7	81.7	74.2	82.8	73.2	66.4	81.3
文献[5]	基于坐标回归方法	97.5	94.3	87.0	81.2	86.5	78.5	75.4	63.4
文献[6]	基于坐标回归方法	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
文献[11]	基于热图检测方法	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
文献[12]	基于热图检测方法	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
文献[13]	基于热图检测方法	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
文献[14]	基于热图检测方法	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5

表 6 列出了基于深度学习的二维多人姿态估计方法在 COCO 数据集上的性能比较，测量标准使用 AP，其中 AP 0.5 表示 $OKS=0.50$，AP 0.75 表示 $OKS=0.75$，AP(M)表示中等大小的对象，AP(L)表示大对象。从表中可以看出，自顶向下的多人姿态估计方法精确度更高，鲁棒性更强。由此可以得出，基于热图检测的多人自顶向下的方法是研究的重点，虽然目前此方法识别速度较慢，但是随着轻量化网络和蒸馏模型的不断发展，识别速度将会有较大改善。而对于自底向上的方法，提出更准确的分组方法将会大幅度提高此类方法的准确性。

表 6 多人姿态估计方法的性能比较

Table 6 Performance comparison of multi-person pose estimation methods

方法	类型	AP	AP 0.5	AP 0.75	AP(M)	AP(L)
文献[19]	自顶向下方法	78.6	94.3	86.6	75.5	83.3
文献[20]	自顶向下方法	76.2	92.5	83.6	72.5	82.4
文献[23]	自顶向下方法	64.9	85.5	71.3	62.3	70.0
文献[26]	自顶向下方法	72.1	91.4	80.0	68.7	77.2
文献[29]	自顶向下方法	74.9	92.5	82.8	71.3	80.9
文献[39]	自底向上方法	67.6	85.1	73.7	62.7	74.6
文献[40]	自底向上方法	70.5	89.3	77.2	66.6	75.8
文献[41]	自底向上方法	61.8	84.9	67.5	57.1	68.2
文献[44]	自底向上方法	65.5	86.8	72.3	60.6	72.6
文献[45]	自底向上方法	68.7	89.0	75.4	64.1	75.5

5 当前面临的问题

近年来，国内外很多学者对人体姿态估计方法进行研究，并且已经取得了巨大的进步，提出了许多解决特定问题的算法，但是仍然存在许多复杂问题有待突破。

(1)算法对环境敏感

光照、遮挡和雨天等现实环境复杂多样，很容易引起多人姿态估计中关键点无法被正确识别、关键点分组错误等问题。

(2)人体姿态复杂多样

人体的姿态多种多样，简单的姿态容易检测，但是复杂的姿态就很难检测，如瑜伽动作和跌倒等。在多人姿态估计中，一张图像会有多个人，不同的人可能有不同的姿态，对于复杂的姿态，检测的准确率会降低，甚至会出现漏检或者错误识别。

(3)算法实时性不高

随着深度学习网络层数的不断加深，人体姿态的识别精度有了很大的提升，但是网络的执行速度较慢，对于自顶向下的算法而言，图像中人数越多，检测速度会越慢，无法满足实际的需求。

(4)多视角数据不易检测

多视角数据可以有效地解决信息欠缺问题。研究算法通常都是在单视角情况下，但实际数据视角并不单一，对于正视角下的数据可以很好地检测出姿态，但对于左视角和右视角下的数据就不容易检测。同时，多视角数据集采集困难，现有数据集都在实验室环境中采集，导致真实野外场景或真实应用场景数据集缺失。

6 未来的研究进展

基于二维的人体姿态估计方法不断发展[73-79]，极大地提升了算法的检测效果。为了解决当前面临的问题，研究者可以尝试解决一些更有挑战性的任务，下面分析了人体姿态估计未来的发展趋势。

(1)优化多人姿态估计网络模型

在人体姿态估计方法中，网络模型优化是重要的主题，而现阶段的网络模型层数都较深，参数量都较大，对网络模型进行压缩以减少参数量可以提高网络的运行速度，加速人体姿态估计的实际落地应用。

(2)加强多任务学习

多任务学习是把人体姿态估计任务跟人体相关或相似的

其他任务放在一起联合学习，如人体分割和人体解析。人体多任务联合学习对于充分理解行人意图具有重要意义，同时也可以互相提升各自的效果。

(3)充分利用时序信息

基于单帧图像的人体姿态估计方法逐渐成熟，可准确高效地估计出人体的关键点，但仍然面临挑战，如遮挡、拥挤、复杂姿态等问题。对于视频格式的输入数据，基于单帧图像的人体姿态估计方法是逐帧对视频数据进行特征提取，虽然能取得一定的效果，但这类方法没有利用视频帧之间的时序特征，无法从时间序列中挖掘到更加丰富的特征信息。采用基于视频的时序人体姿态估计方法能高效地提取视频的时序特征，充分利用视频帧之间的时序信息，例如，在遮挡、拥挤、复杂姿态等复杂环境中，能充分利用相邻帧之间的完整性来克服这些难题。

(4)使用无监督学习模型

目前，人体姿态估计方法大多集中在已标注好的数据集上，人工标注图像数据或者视频数据非常耗时耗力，而且人工标注数据集质量参差不齐，具有一定的局限性，标注质量差的数据集会降低模型的泛化能力，使用无监督学习模型在未标注数据上可自我学习。

结束语 基于深度学习的二维人体姿态估计方法应用于自动驾驶、安全监控、肢体语言理解等领域发挥着重要作用。本文从单人和多人进行分析，对方法模型、准确率及各方法的优缺点进行对比分析。未来亟待通过提高网络的运行速度，利用多任务学习，充分利用相邻帧之间的互补性和加强无监督学习模型来提高方法的自我学习性，满足长视频检测要求、适配应用环境、可靠估计姿态等将成为姿态估计研究的热点。

参 考 文 献

[1] CHEN L, MA N, PANG G L, et al. Research on multi-view data fusion and balanced YOLOv3 for pedestrian detection[J]. CAAI Transactions on Intelligent Systems, 2021, 16(1): 57-65.

[2] TAN M, LE Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]// Proceedings of the 36th International Conference on Machine Learning. PMLR 97, 2019: 6105-6114.

[3] TOSHEV A, SZEGEDY C. DeepPose: Human Pose Estimation via Deep Neural Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1653-1660.

[4] CARREIRA J, AGRAWAL P, FRAGKIADAKI K, et al. Human Pose Estimation with Iterative Error Feedback[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4733-4742.

[5] SUN X, SHANG J X, LIANG S, et al. Compositional Human Pose Regression[J]. arXiv: 1704.00159, 2017.

[6] LUVIZON D C, TABIA H, PICARD D. Human Pose Regression by Combining Indirect Part Detection and Contextual Information [J]. Computers & Graphics, 2019, 85: 15-22.

[7] MAO W, GE Y, SHEN C, et al. TFPose: Direct Human Pose Estimation with Transformers[J]. arXiv: 2103.15320, 2021.

[8] ZHANG H, OUYANG H, LIU S, et al. Human Pose Estimation with Spatial Contextual Information[J]. arXiv: 1901.01760, 2019.

[9] ARTACHO B, SAVAKIS A. UniPose: Unified Human Pose Estimation in Single Images and Videos[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 7035-7044.

[10] LIFSHITZ I, FETAYA E, ULLMAN S. Human Pose Estimation using Deep Consensus Voting[C]// European Conference on Computer Vision. Cham: Springer, 2016: 246-260.

[11] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional Pose Machines[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.

[12] NEWELL A, YANG K, JIA D. Stacked Hourglass Networks for Human Pose Estimation[C]// European Conference on Computer Vision. Cham: Springer International Publishing, 2016: 483-499.

[13] YANG W, LI S, OUYANG W, et al. Learning Feature Pyramids for Human Pose Estimation[C]// IEEE Computer Society. IEEE Computer Society, 2017: 1281-1290.

[14] CHU X, YANG W, OUYANG W, et al. Multi-Context Attention for Human Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1831-1840.

[15] WANG J, JIN S, LIU W, et al. When Human Pose Estimation Meets Robustness: Adversarial Algorithms and Benchmarks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11855-11864.

[16] GROOS D, RAMAMPIARO H, IHLEN E. EfficientPose: Scalable single-person pose estimation[J]. Applied Intelligence, 2021, 51(4): 2518-2533.

[17] 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. Cham: Springer, 2020: 492-508.

[18] HUANG J, ZHU Z, GUO F, et al. The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[19] CAI Y, WANG Z, LUO Z, et al. Learning Delicate Local Representations for Multi-Person Pose Estimation [C]// European Conference on Computer Vision. Cham: Springer, 2020: 455-472.

[20] ZHANG F, ZHU X, DAI H, et al. Distribution-Aware Coordinate Representation for Human Pose Estimation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7093-7102.

[21] IQBAL U, GALL J. Multi-Person Pose Estimation with Local Joint-to-Person Associations [C] // European Conference on Computer Vision (ECCV) Workshops, Crowd Understanding, 2016. Cham: Springer International Publishing, 2016: 627-642.

[22] FANG H S, XIE S, TAI Y W, et al. RMPE: Regional Multi-person Pose Estimation[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 2334-2343.

[23] PAPANDREOU G, ZHU T, KANAZAWA N, et al. Towards Accurate Multi-person Pose Estimation in the Wild [C]// Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：4903-4911.

[24] HUANG S，GONG M，TAO D. A Coarse-Fine Network for Keypoint Localization [C] // 2017 IEEE International Conference on Computer Vision(ICCV). IEEE，2017：3028-3037.

[25] KUMAR C，RAMESH J，CHAKRABORTY B，et al. VRU Pose-SSD：Multiperson Pose Estimation For Automated Driving [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2021：15331-15338.

[26] CHEN Y，WANG Z，PENG Y，et al. Cascaded Pyramid Network for Multi-person Pose Estimation [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2018：7103-7112.

[27] SU K，YU D，XU Z，et al. Multi-Person Pose Estimation with Enhanced Channel-wise and Spatial Information [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2019：5674-5682.

[28] QIU L，ZHANG X，LI Y，et al. Peeking into occluded joints：A novel framework for crowd pose estimation [C] // European Conference on Computer Vision. Cham：Springer，2020：488-504.

[29] SUN K，XIAO B，LIU D，et al. Deep High-Resolution Representation Learning for Human Pose Estimation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：5693-5703.

[30] WANG X，TONG J，WANG R. Attention Refined Network for Human Pose Estimation [J]. Neural Processing Letters，2021(4)：1-20.

[31] IRDHAR R，GKIOXARI G，TORRESANI L，et al. Detect-and-Track：Efficient Pose Estimation in Videos [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE，2018：350-359.

[32] WANG M，TIGHE J，MODOLO D. Combining detection and tracking for human pose estimation in videos [C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2020：11088-11096.

[33] XIAO B，WU H，WEI Y. Simple Baselines for Human Pose Estimation and Tracking [C] // Proceedings of the European Conference on Computer Vision(ECCV). 2018：466-481.

[34] BAO Q，LIU W，CHENG Y，et al. Pose-Guided Tracking-by-Detection：Robust Multi-Person Pose Tracking [J]. IEEE Transactions on Multimedia，2020，23：161-175.

[35] RUAN W，LIU W，BAO Q，et al. POINet：Pose-Guided Ovonic Insight Network for Multi-Person Pose Tracking [C] // Proceedings of the 27th ACM International Conference on Multimedia. 2019：284-292.

[36] UMER R，DOERING A，LEIBE B，et al. Self-supervised Keypoint Correspondences for Multi-Person Pose Estimation and Tracking in Videos [J]. arXiv：2004. 12652，2020.

[37] SNOWER M，KADAV A，LAI F，et al. 15 Keypoints Is All You Need [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：6738-6748.

[38] YANG L P，SUN Y B，ZHANG H L，et al. Human Keypoint Matching Network Based on Encoding and Decoding Residuals [J]. Computer Science，2020，47(6)：114-120.

[39] JIN S，LIU W，XIE E，et al. Differentiable Hierarchical Graph Grouping for Multi-Person Pose Estimation [C] // European Conference on Computer Vision. Cham：Springer，2020：718-734.

[40] CHENG B，XIAO B，WANG J，et al. HigherHRNet：Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：5386-5395.

[41] CAO Z，SIMON T，WEI S E，et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：7291-7299.

[42] YU D，SU K，SUN J，et al. Multi-person Pose Estimation for Pose Tracking with Enhanced Cascaded Pyramid Network [C] // European Conference on Computer Vision. Cham：Springer，2018：221-226.

[43] KREISS S，BERTONI L，ALAHI A. PifPaf：Composite Fields for Human Pose Estimation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：11977-11986.

[44] NEWELL A，HUANG Z，DENG J. Associative Embedding：End-to-End Learning for Joint Detection and Grouping [J]. arXiv：1611. 05424，2016.

[45] 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] // Proceedings of the European Conference on Computer Vision(ECCV). 2018：269-286.

[46] KOCABAS M，KARAGOZ S，AKBAS E. MultiPoseNet：Fast Multi-Person Pose Estimation using Pose Residual Network [C] // Proceedings of the European Conference on Computer Vision(ECCV). 2018：417-433.

[47] INSAFUTDINOV E，ANDRILUKA M，PISHCHULIN L，et al. ArtTrack：Articulated Multi-Person Tracking in the Wild [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：6457-6465.

[48] XIU Y，LI J，WANG H，et al. Pose Flow：Efficient Online Pose Tracking [J]. arXiv：1802. 00977，2018.

[49] ZHANG Z，WANG C，QIN W. Semantically Synchronizing Multiple-Camera Systems with Human Pose Estimation [J]. Sensors，2021，21(7)：2464.

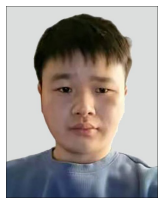
[50] FABBRI M，LANZI F，CALDERARA S，et al. Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World [C] // Proceedings of the European Conference on Computer Vision(ECCV). 2018：430-446.

[51] HWANG J，LEE J，PARK S，et al. Pose estimator and tracker using temporal flow maps for limbs [C] // 2019 International Joint Conference on Neural Networks(IJCNN). IEEE，2019：1-8.

[52] RAAJ Y，IDREES H，HIDALGO G，et al. Efficient Online Multi-Person 2D Pose Tracking With Recurrent Spatio-Temporal Affinity Fields [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2019：4620-4628.

[53] JIN S，LIU W，OUYANG W，et al. Multi-Person Articulated Tracking With Spatial and Temporal Embeddings [C] // Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:5664-5673.
[54] HELMSTETTER S,SNGER J,GERMANN R,et al. How to use human pose estimation to measure the hand-arm motion in craft application with no influence on the natural user behavior [J]. Procedia CIRP,2021,100:631-636.
[55] STENUM J,ROSSI C,ROEMMICH R T. Two-dimensional video-based analysis of human gait using pose estimation[J]. PLoS Computational Biology,2021,17(4):e1008935.
[56] FANG H S,CAO J,TAI Y W,et al. Pairwise Body-Part Attention for Recognizing Human-Object Interactions [C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:51-67.
[57] LI Y L,LIU X,WU X,et al. Transferable Interactiveness Knowledge for Human-Object Interaction Detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:3585-3594.
[58] WAN B,ZHOU D,LIU Y,et al. Pose-aware Multi-level Feature Network for Human Object Interaction Detection [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:9469-9478.
[59] LUVIZON D C,PICARD D,TABIA H. 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:5137-5146.
[60] DU W,WANG Y,YU Q. RPAN:An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:3725-3734.
[61] LUDWIG K,SCHERER S,EINFALT M,et al. Self-Supervised Learning for Human Pose Estimation in Sports [C]// 2021 IEEE International Conference on Multimedia & Expo Workshops(ICMEW). IEEE,2021:1-6.
[62] LI M,CHEN S,CHEN X,et al. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:3595-3603.
[63] SHI L,ZHANG Y,CHENG J,et al. Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks [J]. IEEE Transactions on Image Processing,2020,29:9532-9545.
[64] DONG J,CHEN Q,SHEN X,et al. Towards Unified Human Parsing and Pose Estimation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014:843-850.
[65] LIANG X D,GONG K,SHEN X H,et al. Look into Person:Joint Body Parsing Pose Estimation Network and a New Benchmark [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,41(4):871-885.
[66] JOHNSON S,EVERINGHAM M. Clustered pose and nonlinear appearance models for human pose estimation[C]// Proceedings of the British Machine Vision Conference. Wales,2010:1-11.
[67] ANDRILUKA M,PISHCHULIN L,GEHLER P,et al. 2d human pose estimation:New benchmark and state of the art analysis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014:3686-3693.
[68] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft COCO:Common Objects in Context[C]// European Conference on Computer Vision. Cham:Springer,2014:740-755.
[69] LI J,WANG C,ZHU H,et al. CrowdPose:Efficient Crowded Scenes Pose Estimation and A New Benchmark [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:10863-10872.
[70] ZHANG W,ZHU M,DERPANIS K G. From Actemes to Action:A Strongly-Supervised Representation for Detailed Action Understanding[C]// Proceedings of the IEEE International Conference on Computer Vision. 2013:2248-2255.
[71] JHUANG H,GALL J,ZUFFI S,et al. Towards understanding action recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. 2013:3192-3199.
[72] ANDRILUKA M,IQBAL U,MILAN A,et al. PoseTrack:A Benchmark for Human Pose Estimation and Tracking [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:5167-5176.
[73] KITAMURA T,TESHIMA H,THOMAS D,et al. Refining OpenPose with a new sports dataset for robust 2D pose estimation[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022:672-681.
[74] WANG Y,LI M,CAI H,et al. Lite pose:Efficient architecture design for 2d human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:13126-13136.
[75] COTTON R J. Posepipe:Open-source human pose estimation pipeline for clinical research[J]. arXiv:2203.08792,2022.
[76] GUPTA D,ARTACHO B,SAVAKIS A. HandyPose:Multi-level framework for hand pose estimation[J]. Pattern Recognition,2022,128:108674.
[77] AN S,ZHANG X,WEI D,et al. FastHand:Fast monocular hand pose estimation on embedded systems[J]. Journal of Systems Architecture,2022,122:102361.
[78] ZHANG M,ZHOU Z,DENG M. Cascaded hierarchical CNN for 2D hand pose estimation from a single color image[J]. Multimedia Tools and Applications,2022,81(18):25745-25763.
[79] LIANG S,CHU G,XIE C,et al. Joint relation based human pose estimation[J]. The Visual Computer,2022,38(4):1369-1381.



ZHANG Guo-ping, born in 1995, master. His main research interests include human pose estimation, interactive cognition and action recognition.

MA Nan, born in 1978, Ph.D, professor. Her main research interests include interactive cognition, intelligent driving, knowledge discovery and intelligent system.

(责任编辑:何杨)