软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(11):4173–4191 [doi: 10.13328/j.cnki.jos.006390]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度学习的二维人体姿态估计综述*

张 宇[1], 温光照[1], 米思娅[2,3], 张敏灵[1,2], 耿 新[1]

[1](东南大学 计算机科学与工程学院, 江苏 南京 211189)

[2](东南大学 网络空间安全学院, 江苏 南京 211189)

[3](网络通信与安全紫金山实验室, 江苏 南京 211111)

通信作者: 米思娅, E-mail: SiyaMi@seu.edu.cn

摘 要: 人体姿态估计是计算机视觉领域的一个基础且具有挑战的任务, 人体姿态估计对于描述人体姿态、描述人体行为等至关重要, 是行为识别、行为检测等计算机视觉任务的基础. 近年来, 随着深度学习的发展, 基于深度学习的人体姿态估计算法展现出了极其优异的效果. 从单人人体姿态估计、自顶向下的多人人体姿态估计和自底向上的多人人体姿态估计这 3 种主流的人体姿态估计方式, 介绍近年来基于深度学习的二维人体姿态估计算法的发展, 并讨论目前二维人体姿态估计所面临的困难和挑战. 最后, 对人体姿态估计未来的发展做出展望.

关键词: 深度学习; 二维人体姿态估计; 关键点检测

中图法分类号: TP18

中文引用格式: 张宇, 温光照, 米思娅, 张敏灵, 耿新. 基于深度学习的二维人体姿态估计综述. 软件学报, 2022, 33(11): 4173–4191. http://www.jos.org.cn/1000-9825/6390.htm

英文引用格式: Zhang Y, Wen GZ, Mi SY, Zhang ML, Geng X. Overview on 2D Human Pose Estimation Based on Deep Learning. Ruan Jian Xue Bao/Journal of Software, 2022, 33(11): 4173–4191 (in Chinese). http://www.jos.org.cn/1000-9825/6390.htm

Overview on 2D Human Pose Estimation Based on Deep Learning

ZHANG Yu[1], WEN Guang-Zhao[1], MI Si-Ya[2,3], ZHANG Min-Ling[1,2], GENG Xin[1]

[1](School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

[2](School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China)

[3](Purple Mountain Laboratory, Nanjing 211111, China)

Abstract: Human pose estimation is a basic and challenging task in the field of computer vision. It is the basis for many of computer vision tasks, such as action recognition and action detection. With the development of deep learning methods, deep learning-based human pose estimation algorithms have shown excellent results. This study divides pose estimation methods into three categories, including single person pose estimation, top-down multi-person pose estimation, and bottom-up multi-person pose estimation. The development of 2D human pose estimation algorithms in recent years is introduced, and the current challenges of two-dimensional human pose estimation are discussed. Finally, the outlook for the future development of human pose estimation is given.

Key words: deep learning; 2D human pose estimation; keypoint detection

人体姿态估计是计算机视觉领域的一个基础且具有挑战的任务, 人体姿态估计对于描述人体姿态、人体行为等至关重要. 有许多计算机视觉任务都是以人体姿态估计任务作为基础的, 包括行为识别、行为检测等等[1−3]. 近些年, 随着深度学习技术的发展, 尤其是随着卷积神经网络算法的提出, 我们可以通过神经网络强大的拟合能力和特征提取能力[4,5]建立一种隐式的人体姿态估计模型, 大大降低了人体姿态估计的门槛, 同时也提高了人体姿态估计的准确率, 这也使得人体姿态估计得到快速的发展.

* 基金项目: 国家重点研发计划(2018AAA0100100); 国家自然科学基金(61702095); 江苏省自然科学基金(BK20190341)
收稿时间: 2020-01-18; 修改时间: 2021-01-06; 采用时间: 2021-06-03; jos 在线出版时间: 2021-08-03

基于深度的人体姿态估计模型发迹于 2014 年, Google 提出了 DeepPose[6], 首次利用深度神经网络进行了人体姿态估计; 同年也发布了目前最为常用的基准数据集: MPII 数据集[7]和 MS-COCO 数据集[8]. 之后, 基于深度学习的人体姿态估计方法就开始了快速的发展, 有关姿态估计的研究成果如雨后春笋般, 不断在各大国际会议和期刊上发表.

本文对近年来人体姿态估计的研究做一个归纳和总结, 为相关领域的研究者提供参考. 本文第 1 节概述二维人体姿态估计的研究现状. 第 2 节从单人人体姿态估计、自顶向下(top-down)的多人人体姿态估计和自底向上(bottom-up)的多人人体姿态估计这 3 种主流的人体姿态估计方式来介绍近年来主流的基于深度学习的二维人体姿态估计方法. 第 3 节给出目前主流方法的实验结果并进行对比分析. 第 4 节讨论目前二维人体姿态估计领域所面临的困难和挑战, 并对未来的研究方向给出建议.

1 二维人体姿态估计的研究现状

二维人体姿态估计的基本定义是: 从单张 RGB 图像中精确地识别多个人体的位置以及其骨架上的稀疏关键点的位置. 如图 1 所示, 是二维人体姿态估计的效果图. 从单张 RGB 图像中估计对应的关键点位置, 直观地想, 需要去思考每个关键点坐标周围的局部特征是什么样的、相邻的关键点坐标之间的约束关系是什么样的、如何将同一个人的各个关键点坐标关联起来等问题, 这些问题正是人体姿态估计要面对的基本问题.

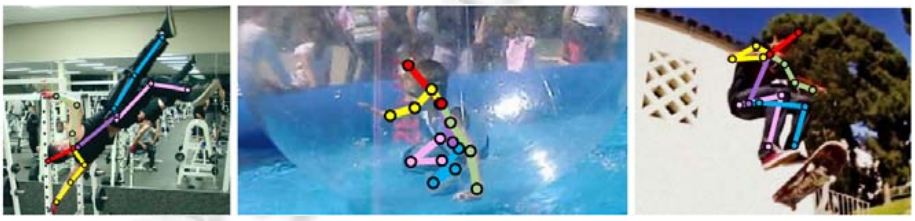


图 1 2D 人体姿态估计[14]

深度学习的快速发展, 尤其是卷积神经网络算法的提出, 使得我们可以通过利用神经网络的强大拟合和特征抽取能力, 建立一种隐式的人体姿态估计模型, 避开上述显式的人体姿态估计的基本问题, 大大降低了人体姿态估计研究的门槛, 使得人体姿态估计的研究和深度学习的发展紧密结合起来. 同时, 由于神经网络模型的引入, 也衍生了许多新的问题, 包括如何引导神经网络模型提取有效的人体姿态特征、如何有效地利用人体姿态特征、如何设计一个高效的人体姿态估计模型等等. 针对这些问题的研究, 使得近年来人体姿态估计的研究得到了快速的发展.

2 主流的二维人体姿态估计方法

目前主流的两种基于深度学习的二维人体姿态估计方法, 分为单人人体姿态估计、自顶向下的多人人体姿态估计和自底向上的多人人体姿态估计. 本节中将介绍单人人体姿态估计、自顶向下的多人人体姿态估计和自底向上的多人人体姿态估计这 3 种人体姿态估计方式, 以及近年来提出的基于以上 3 种方式的二维人体姿态估计算法, 并分析其发展历程.

2.1 单人二维人体姿态估计方法

单人二维人体姿态估计方法, 顾名思义, 指的是从单张 RGB 图片中, 精确地识别单个人体骨架上的稀疏关键点的位置. 当图片中包含多个人时, 只识别主体人的关键点. 由于受到算力的限制, 大部分早期的人体姿态估计方法都是单人人体姿态估计方法. 单人二维人体姿态估计方法的研究也是自顶向下和自底向上的多人人体姿态估计方法研究的重要的基础性研究, 至今仍有许多优异的单人二维人体姿态估计方法不断被提出.

早期的基于深度学习人体姿态估计方法如 DeepPose[6], 是一种基于数值坐标回归的单人人体姿态估计方法, 其通过训练类似 AlexNet[9]的深度神经网络, 无须使用任何人体模板, 就能直接从图像中回归人体的关键

点坐标, 并通过多阶段的提炼获得较为准确的人体关键点坐标. DeepPose 首次将深度学习引入到了人体姿态估计中, 对于人体姿态估计的研究具有开创性意义. 但是, 基于数值坐标回归的人体姿态估计方法是直接从图像到数值坐标的端到端的回归过程, 会丢失关键点的空间信息, 使得训练出来的模型缺乏空间泛化能力.

因而, 基于关键点热图回归的人体姿态估计方法应运而生. Tompson 等人[10]提出了一种基于深度卷积神经网络和图模型的人体姿态估计方法, 该方法采用热图回归人体姿态估计方法, 首次将热图引入到人体姿态估计中, 并利用人体关键点之间的结构关系, 结合马尔科夫随机场优化预测结果, 使得该方法成为当时最优越的人体姿态估计方法. 基于热图回归将人体姿态估计从原本的坐标回归问题转化成检测问题, 最大程度地保留了关键点坐标的空间信息, 大大提高了学习得到的姿态估计模型的空间泛化能力, 进而提高了姿态估计的准确率. 由于热图回归的优越性, 此后大多数基于深度学习的人体姿态估计方法都采用关键点的热图作为网络的回归目标, 使得姿态估计领域的研究迈出了重要的一步.

之后, 如图 2 所示, 另一个影响深远的单人人体姿态估计方法——卷积姿态机(convolution pose machine, CPM)[11], 于 2016 年被提出. 卷积姿态机通过一个多阶段序列卷积结构模型学习关键点的空间信息, 并引入了中间监督(intermediate supervision)解决了多阶段序列卷积结构模型学习过程中, 由于输出层的误差经过多层反向传播而导致梯度消失的问题. 同时也指出了网络模型的感受野对于精确的人体姿态估计的重要性, 并以较大的性能优势领先于当时的其他方法, 被公认为是当时最为优雅工整且性能较好的网络结构.

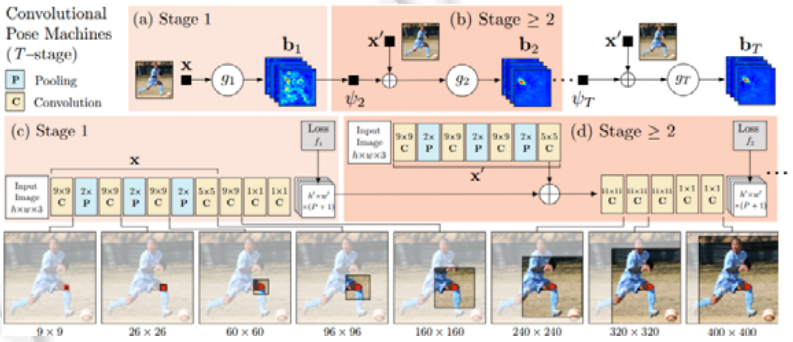


图 2 卷积姿态机的结构[11]

直到同年, Newell 等人[12]提出了二维人体姿态估计领域里程碑级别的方法——堆叠沙漏网络(stacked hourglass networks, SHN), 再次刷新了各大人体姿态估计比赛的榜单; 同时, 由于结构的简单灵活和优越的性能, 堆叠沙漏网络成为姿态估计领域的新宠儿. 至今, 许多后来提出的方法都是基于它做出的改进. 如图 3 是堆叠沙漏网络的基本结构.

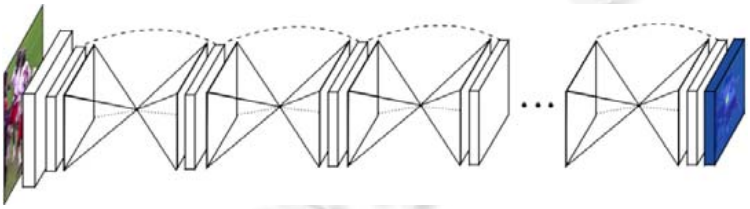


图 3 堆叠沙漏网络的结构[12]

堆叠沙漏网络, 顾名思义, 有许多类似沙漏结构的大大小小的特征图. 堆叠沙漏网络将图片中人体区域对应的子图作为输入, 经过一系列卷积和池化操作, 得到子图对应的基础特征图, 然后将特征图输入到对应的沙漏结构的网络中. 堆叠沙漏网络的操作过程.

(1) 通过卷积操作提取特征图的局部特征, 同时降低特征图的分辨率. 随着卷积层的不断增加, 对应的特征图也不断减小, 此时, 每一个特征值对应的感受野逐渐增大, 当到沙漏的中心时, 此时特征图

上的特征点的感受野基本上能够涵盖整个子图，相当于包含了图片的全局信息. 这一过程相当于对输入特征图的编码操作，各阶段的特征图对应不同尺度的特征信息.

(2) 沙漏的后半部分对特征图进行了最近邻上采样，同时和沙漏前半部分大小相同的特征图进行了跨层连接，将不同尺度局部特征和全局特征逐步混合，逐步逼近特征点的精确位置.

(3) 将多个沙漏结构进行堆叠，同时引入卷积姿态机中提出的中间监督，将每一个沙漏网络的输出都和目标热图进行损失的计算，从而提高人体姿态估计的准确率.

堆叠沙漏网络的核心是：通过反复地对特征图进行高低分辨率的编解码操作，更好地融合图片的全局和局部特征，进而提高人体姿态估计的准确率. 这一思想在人体姿态估计的研究中被不断借鉴，且得到了进一步的发展. Chu 等人[13]基于堆叠沙漏网络提出了一个用于人体姿态估计的多上下文注意力机制(multi-context attention)，首次在人体姿态估计中引入了注意力机制，改进了沙漏网络中的残差分支，提出了具有更大感受野的沙漏残差单元(hourglass residual units, HRU). 并利用沙漏网络中的多分辨率特征图，结合条件随机场生成多分辨率注意力，更好地利用了提取的多分辨特征图对应多尺度特征. 同时，对于不同层的沙漏网络，生成多语义注意力，并提出了分层的注意力生成机制：在底层的沙漏网络中生成整体注意力；而在高层的沙漏网络中，对每个关键点生成部分注意力，构建由粗糙到精细的注意力模型. 多上下文注意力机制大大提高了人体姿态估计的准确率，也使得该方法也成为 2016 年 COCO 人体姿态估计挑战赛的冠军方法.

之后，Sun 等人[14]发现：以堆叠沙漏网络为代表的一些网络结构，为了提取图片的多尺度特征，都存在利用卷积进行特征图的下采样操作，但是下采样操作会导致一定程度的信息损失，从而影响姿态估计的准确率. 简单基线网络通过反卷积操作，一定程度上缓解了这一点，从而提高了姿态估计的准确率，但是没有本质上解决这个问题. 因而，针对这个问题，他们提出了高分辨率特征表示学习网络(high-resolution representation networks, HRNet)，如图 4 所示. 高分辨率特征表示学习网络是以堆叠沙漏网络为代表的一系列多分辨率融合网络的进一步发展. 它始终保持高分辨率的特征图作为运行主线，在进行下采样时，生成一个新的低分辨率子网络，不断横向扩展网络，通过多个并行的子网络，从头到尾保持图像的高分辨率特征，从而避免了特征图在上下采样的过程造成信息损失；同时，每隔一定层数对多个并行子网络进行信息融合，从而实现高低分辨率的特征融合，进一步提高了姿态估计的准确率. 这一方法再次刷新了 2017 年 MS-COCO 数据集上人体姿态估计的准确率，成为二维人体姿态估计任务中更强的基线模型.

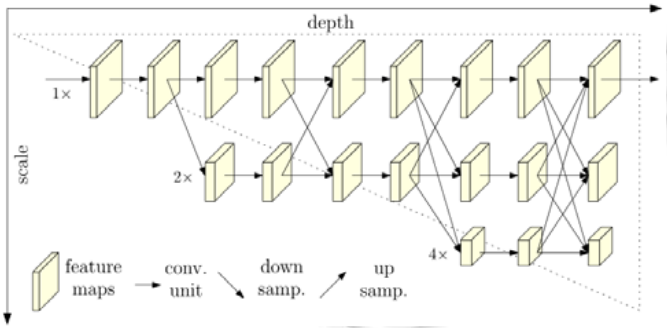


图 4 高分辨率特征表示学习网络的结构[14]

后来，Artacho 等人[15]认为：目前的人体姿态估计方法都是一个多阶段的人体姿态估计方法，存在一定的局限性，因而提出了一种统一的人体姿态估计方法——UniPose. 同时估计人体的边界框和人体关键点的位置，将两个阶段的人体姿态估计过程变成了单阶段的人体姿态估计过程，简化了人体姿态估计的步骤. 同时，Artacho 等人将其曾在语义分割领域中提出的瀑布空洞空间池化(water-fall atrous spatial pooling, WASP)模块[16]引入到了人体姿态估计中. 如图 5 所示，是用于姿态估计的瀑布空洞空间池化模块的结构图.

瀑布空洞空间池化模块也是一种多尺度特征融合模块，其利用空洞卷积在不降低特征图分辨率的情况下，渐近地提取不同尺度的感受野对应的不同尺度的特征，再通过瀑布流的形式引出并融合多尺度的特征，

同时估计人体边界框和人体关键点坐标. 该方法再次刷新了 MPII 数据集上单人人体姿态估计准确率, 成为在不引入额外数据的情况下, MPII 数据集单人人体姿态估计榜上最先进的人体姿态估计方法.

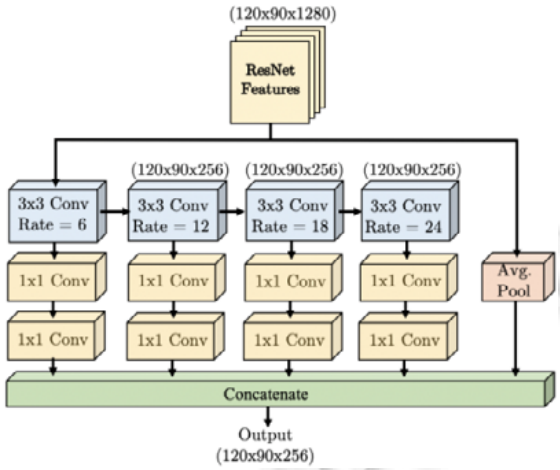


图 5 瀑布空洞空间池化模块[15]

除了针对提取和融合多尺度特征的改进以外, 有不少研究人员注意到: 人体本身具有一定的结构性特征, 可以根据人体结构特征, 利用较容易估计的关键点坐标去辅助较难、易被遮挡的关键点坐标的估计. 早期的时候, Chu 等人[17]人提出了一种用于人体姿态估计的结构特征学习方法, 引入了几何变换核去学习人体关键点之间的偏移. 同时, 为了避免两个较远关键点之间存在过大的偏移和较难学习关联的问题, 他们构建了一个双向树模型, 设计关键点之间的传播关系, 即只在相邻关键点之间传播热图特征去辅助关键点之间的估计, 利用结构特征提高人体姿态估计的准确率.

后来, 如图 6 所示, Ke 等人[18]结合多尺度和人体结构性特征的研究, 提出了一种用于人体姿态估计的多尺度结构感知网络(multi-scale structure-aware network, MSSAN). 在堆叠沙漏网络的基础上, 引入多尺度监督和多尺度回归, 利用沙漏网络中间生成的多分辨率特征图, 生成多分辨率热图实现多尺度的监督. 最后, 融合多分辨的特征图实现多尺度回归. 并且提出了一种结构感知损失, 绑定相邻关键点计算损失, 从而使得网络能够捕捉到关键点之间的连接信息. 考虑到训练数据中困难场景的样本量太少, 他们还提出了一种关键点掩蔽(keypoint masking)的数据增强方法, 通过遮挡和粘贴部分关键点区域, 生成带遮挡和类似多人重叠的样本, 进一步提高模型对遮挡和多人重叠样本等困难样本人体姿态估计的准确率.

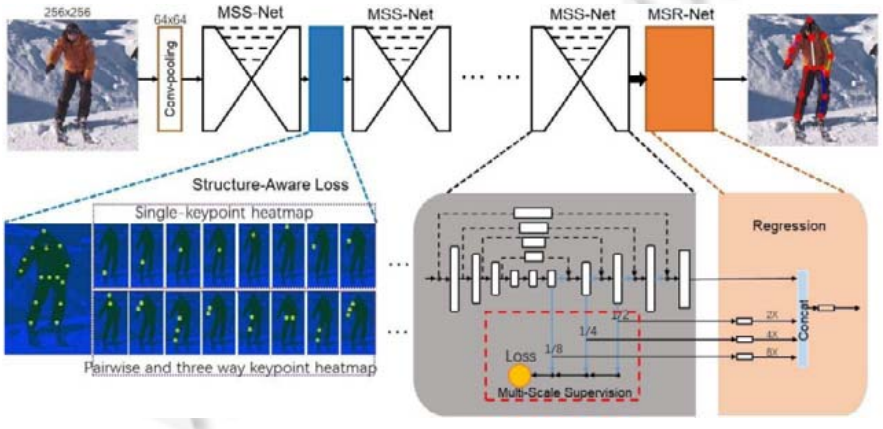


图 6 多尺度结构感知网络[18]

近些年，由于图神经网络对于学习图结构信息传播模型的有效性，受到了很大一部分研究人员的青睐，这一特性也和人体姿态估计中人体结构特征的学习相契合. 如图 7 所示, Zhang 等人[19]提出了一种基于空间上下文信息(spatial contextual information, SCI)的人体姿态估计方法. 该方法以级联预测融合(cascade prediction fusion, CPF)的堆叠沙漏网络作为基本框架，学习获得基础的关键点热图，并引入了用于姿态估计的姿态图神经网络(pose graph neural network, PGNN)，基于人体结构的图模型去学习关键点热图之间的特征关联，促进了模型对人体结构特征的利用，优化了关键点估计，获得了更为精确的人体姿态估计结果.

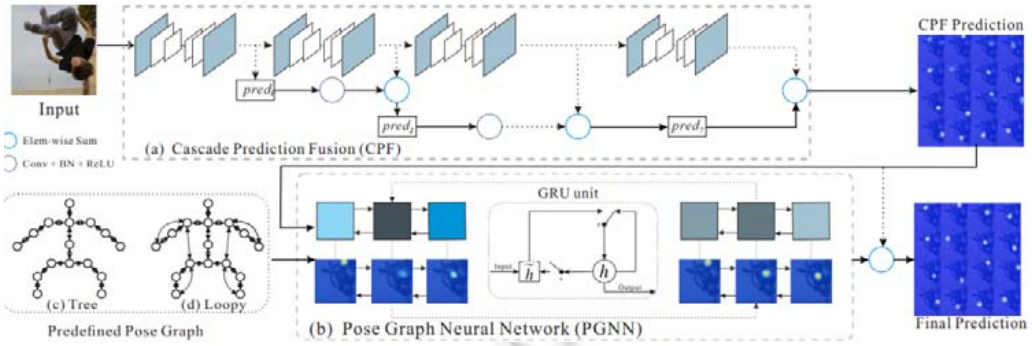


图 7　基于空间上下的信息的人体姿态估计[19]

用于人体姿态估计的图模型控制特征的传播路径，但是没有注意对传播的特征中存在的部分冗余信息的过滤. Bulat 等人[20]通过研究特征分布，发现类似堆叠沙漏网络等结构中的残差连接传播的特征中，只有少部分通道的特征对于下一个阶段是有价值的，因而直接的连接可能会引入一些没有价值甚至阻碍整体性能的特征. 因而提出了一个软门控跳跃连接(soft-gated skip connection, SGSC)，通过通道加权，控制跳跃连接特征之间的信息传递，进而提高了人体姿态估计的准确率.

此外, Nie 等人[21]发现：人体姿态估计的研究和人体解析的研究都是针对人体部分的，两者之间存在一定的相关性，人体解析提取的特征一定程度上也能辅助人体姿态的研究. 因而，他们提出了一种用于人体姿态估计的解析诱导学习器(parsing induced learner, PIL). 通过训练一个自适应卷积核，将人体解析标签训练学习到的人体特征融合到人体姿态估计的网络中，提高了人体姿态估计的准确率. 之后, Nie 等人[22]还提出了一个人体解析和人体姿态估计的相互学习适应(mutual learning to adapt, MuLA)的多任务框架，多次交叉融合人体解析和人体姿态估计提取的特征，从而同时提高人体解析和人体姿态估计的准确率.

深度学习是受数据驱动的，而用于训练的数据集往往是有限的，因而，如何利用好已有的数据对于深度学习的研究而言非常重要. 受 Ke 等人[18]提出的关键点掩蔽的数据增强方法的启发，如图 8 所示，Bin 等人[23]提出了用于人体姿态估计的对抗语义数据增强(adversarial semantic data augmentation, ASDA)，利用语义分割，将人体分成多个有语义的部件，构建一个部件池，再将这些人体部件和训练样本进行重组，并通过训练生成对抗网络(generative adversarial network, GAN)[24]去控制重组时每个人体部件的空间变换参数，进而生成更有挑战的人体姿态估计样本，从而使得模型能够更好地学会估计遮挡、重叠等复杂条件下的人体姿态，进而提高人体姿态估计的准确率.

随着用于单人人体姿态估计的深度学习模型研究日益成熟，单人人体姿态估计准确率不断地提高，单人人体姿态估计中的一些基本问题都能得到较好的解决，人们开始关注姿态估计中存在的细节性的问题. 从数值坐标到热图估计，虽然大大提高了人体姿态估计的准确率，但是标准的热图表示也引入了一些固有的缺陷. 因为原始图像到热图是存在降采样操作的，也就意味着真实关键点坐标在热图上应该是一个浮点数，但是由于热图上的像素点坐标都是整数，所以在生成热图标签时不免需要做一些近似，进而引入了统计误差. 针对这个问题，Zhang 等人[25]提出了一种感知分布的坐标表示方法，即基于统计策略更加精确的坐标编解码方式. 编码时，利用原始图像中的坐标，即未量化引入统计误差的坐标，去生成热图中对应点的高斯分布. 解码时，

调整输出热图分布, 使输出的热图符合高斯分布, 并计算高响应位置的一阶和二阶导数, 从而推算出对应高斯分布精确的峰值坐标点, 作为人体姿态估计的输出. 他将提出的感知分布的坐标表示方法和高分辨率特征表示学习网络相结合, 刷新了 COCO2017 数据集测试集上的人体姿态估计准确率.

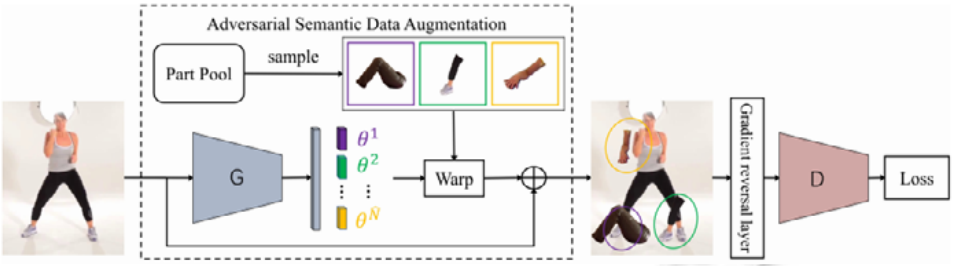


图 8 对抗语义数据增强[23]

紧随其后, Huang 等人[26]也针对标准的热图回归中存在的统计误差的问题, 提出了一种无偏数据处理(unbiased data processing, UDP)的姿态估计方法, 利用一个基于圆形区域标注的热图和两个偏置图作为标签, 提出了基于偏移量的编码方式替代原本的单一热图的编码方式. 解码时, 先利用热图确定关键点的位置, 然后利用两个偏置图去修正关键点的位置, 使得姿态估计统计误差的期望为 0, 实现无偏的姿态估计. 此外, 他们还发现了在姿态估计过程中, 数据增强以及测试步骤中常需要使用的基于像素点的翻转操作, 会使得翻转图像对应的关键点位置和原图像对应的关键点位置存在像素点偏移, 因而提出了一种无偏的数据处理方法, 利用单位长度而不是像素点作为图像尺寸测量的标准进行图像翻转. 他们将提出的无偏的姿态估计方法和数据处理方式, 与高分辨率特征表示学习网络相结合, 再次刷新了在不利用额外数据的情况下, COCO2017 人体姿态估计数据集测试集上的人体姿态估计准确率.

在对准确率的不断刷新之外, 模型的轻量化[27,28]对于应用落地而言也是不可或缺的. 因而, Zhang 等人[29]提出了更加适用于实际应用的快速人体姿态估计. 其根据堆叠沙漏网络的对称性结构将堆叠沙漏网络进行了压缩, 同时利用多阶段的堆叠沙漏网络模型进行知识蒸馏, 引导压缩后的堆叠沙漏网络模型的训练. 进而在保证一定模型准确率的基础上, 大大降低了模型的大小和计算复杂度.

2.2 自顶向下的多人二维人体姿态估计方法

自顶向下的方法, 即先定位人体, 再针对定位的人体逐个进行关键点坐标的估计. 一般的流程是: 利用人体检测器, 从图像中检测人体位置, 获得单个人体所对应的边界框(bounding box)或人体对应区域, 裁剪图像或特征图, 将裁剪获得的子图作为网络的输入进行多次单人人体姿态估计, 将多个子图的输出作为原图人体姿态估计的结果. 自顶向下的多人人体姿态估计方法利用检测获得的人体边界框或人体区域, 将原本多人人体姿态估计问题转化成为多个单人人体姿态估计问题, 简化了关键点估计的复杂度, 一定程度上提高了人体姿态估计的准确率.

最一般的自顶向下的多人人体姿态估计方法, 是利用互相独立人体检测器和单人人体姿态估计方法的组合进行自顶向下的多人人体姿态估计, 这也就意味着: 只需要为单人人体姿态估计方法提供一个人体检测器, 就可以无缝地将单人人体姿态方法转化为自顶向下的多人人体姿态估计方法, 所以这类方法的研究和单人人体姿态估计的研究是高度一致的.

这类方法中, 比较具有代表性的是 2017 年 COCO 挑战赛的冠军方法——旷世[30]提出的级联金字塔网络(cascaded pyramid networks, CPN). 该方法利用基于特征金字塔的目标检测器作为人体边界框检测器, 然后利用提出的级联金字塔网络进行单人人体姿态估计. 如图 9 所示, 级联金字塔网络分成两个部分: GlobalNet 和 RefineNet. GlobalNet 进行基本的人体关键点的检测, 采用了特征金字塔(feature pyramid networks, FPN)的结构, 该结构和单个沙漏网络的结构十分相似, 都是先下采样提取多尺度特征, 再进行上采样融合多尺度特征的两个过程. 不同的是, 特征金字塔引入了多尺度监督的部分, 对所有尺度的特征图都生成关键点估计损失,

使得模型对不同尺度的检测具有更强的包容性；而 RefineNet 通过卷积和上采样融合多分辨率的特征图，综合全局和局部特征，对 GlobalNet 无法精准估计的关键点进行了修正，使得姿态估计模型对于复杂背景或遮挡的关键点的估计准确率进一步提高.

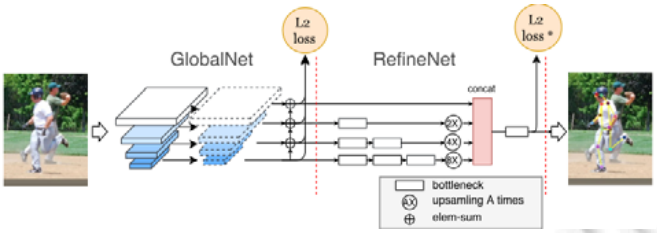


图 9　级联金字塔网络的结构[30]

相似地，微软亚洲研究院[31]提出了一个用于二维人体姿态估计的简单基线(simple baseline)，采用一种简单直接的网络结构，实现了高准确率的人体姿态估计. 简单基线第 1 阶段利用 Faster R-CNN[32]作为人体边界框检测器. 第 2 阶段以 ResNet[33]作为主干网络，也采用了类似堆叠沙漏网络的下采样和上采样操作. 但不同的是，简单基线取消了其中跨层连接的部分，并利用反卷积结构替代堆叠沙漏网络中的最近邻上采样操作,进行特征图的上采样操作. 简单基线引入反卷积操作，替代传统的上采样操作，以一种简单而有效的方法提高了人体姿态估计的准确率，进一步刷新了 COCO2017 数据集上人体姿态估计的准确率.

还有 Su 等人[34]提出了一种通道和空间信息增强的人体姿态估计方法，该方法的核心主要包含两点：首先是受到堆叠沙漏网络、级联金字塔网络等多分辨率融合的人体姿态估计方法以及 ShuffleNet[35]的通道洗牌(channel shuffle)方法的启发，提出了一种基于通道洗牌的多分辨融合机制，增强了多分辨特征之间的信息流转和融合；另一个受到 SENet[36]和 SCA-CNN[37]的启发，继 Chu 等人[13]提出的用于人体姿态估计的多上下文注意力机制之后，再次将注意力机制引入到了人体姿态估计中，在普通的残差单元中引入空间注意力和通道注意力模块，提出了一个注意力残差瓶颈(attention residual bottleneck)模块，增强了多分辨率特征的空间和通道信息，提高了人体姿态估计的准确率.

除了单人人体姿态估计部分以外，对于自顶向下的多人人体姿态估计而言，人体边界框或区域定位的精度也会大大影响人体姿态估计的准确率. 因而存在不少自顶向下的多人人体姿态估计方法，通过改进人体边界框或区域定位的方法，进而提高人体姿态估计的准确率. 著名的 Mask-RCNN[38]是一种典型的自顶向下的多人人体姿态估计方法. Mask R-CNN 扩展了 Faster R-CNN，用 ROIAlign 替代了 ROIPooling 优化了特征图上人体区域的提取，然后采用类似语义分割的方式，每一个人体关键点对应一个独热码的掩膜，实现自顶向下的人体姿态估计.

自顶向下的人体姿态估计中，优化提取人体边界框的方法还有上海交通大大学卢策吾组[39]提出的区域多人姿态估计(regional multi-person pose estimation，RMPE)框架，即著名的 AlphaPose. 如图 10 所示，是 AlphaPose 的基本框架.

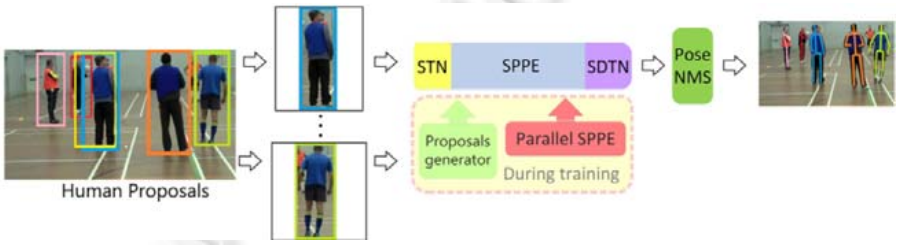


图 10　区域多人人体姿态估计[39]

AlphaPose 设计了一个新型的对称空间转换网络(sysmmetric spatial transformer network, SSTN)，该网络分

为主要3个部分: 首先, 第1个部分是一个空间转换网络(spatial transformer network, STN), 从一个不准确的人体边界框中提取一个高质量、更有利于识别的单人人体区域, 然后进行单人人体姿态估计; 第 2 个部分是根据提取出来的高质量的单人人体区域进行单人人体姿态估计(single person pose estimation, SPPE); 最后一个部分是一个空间逆转换网络(spatial de-transformer network, SDTN), 将估计出来的人体姿态重映射到原图上, 获得原图上的人体姿态坐标. 此外, AlphaPose 中也采用了一个参数化的姿态非极大值抑制算法, 利用估计出来的关键点坐标消除冗余的姿态估计.

Google[40]也提出了一个对边界框提取进行优化的自顶向下的人体姿态估计基线, G-RMI. 类似一般的自顶向下的人体姿态估计方法, G-RMI 也将人体姿态估计看成两个阶段的过程, 即先提取人体边界框, 再进行人体姿态估计的过程. 但是除了优化这两个阶段过程以外, G-RMI 还提出了基于关键点置信度的边界框置信度重定义算法和基于关键点相似度的非极大抑制算法, 利用第 2 阶段估计获得关键点坐标优化第一个阶段人体边界框提取, 从而提高整个自顶向下的多人人体姿态估计框架的可靠性. 此外, 如图 11 所示, G-RMI 还提出了一个精确的姿态估计方法, 将基于圆形区域标注的热图和包含两个通道的偏置图作为标签, 利用双线性插值核融合热图和偏置图获得更加精确的融合激活图, 将融合激活图中的最大值区域作为关键点估计的结果, 一定程度上弥补了人体姿态估计的输出和真实标签相比存在下采样、降低量化精度的问题. 后来, 单人人体姿态估计中, Huang 等人[26]提出的无偏姿态估计和该方法非常相似, 不同的是, G-RMI 最终还是基于一个融合激活图最大的响应位置去获得关键点坐标位置, 是一个依赖于输出融合激活图的分辨率大小的整型坐标. 而 Huang 等人[26]提出的无偏姿态估计的关键点坐标是热图最大值对应坐标加上偏移量的值, 对应于一个浮点型坐标.

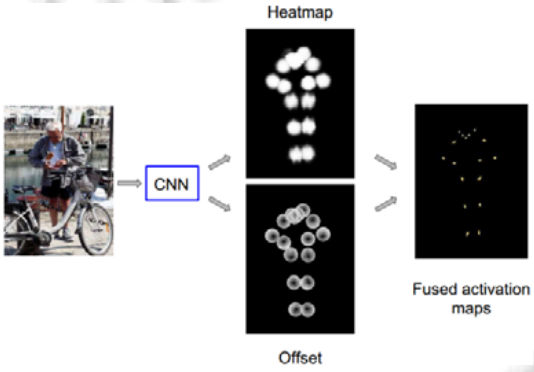


图 11 G-RMI 中精确的姿态估计方式[40]

此外, 还有 Wang 等人[41]提出了一种在视频中结合检测和跟踪的人体姿态估计方法. 该方法定义了一个切片跟踪网络(clip tracking network), 先在关键帧中估计出基本的人体边界框, 然后从视频切片的全部帧中裁剪出该区域的子图, 组成一个区域视频切片, 再将该切片输入到一个 3D 高分辨率表示网络中, 从而输出区域视频切片所有帧中属于关键帧中所定位人的人体姿态. 最后, 再通过一个时间和空间的平滑, 合并估计出来的人体姿态, 从而获得一个准确的人体姿态估计. 该方法通过传播人体区域, 利用前后帧的信息去定位关键帧中没有准确定位的关键点坐标, 从而降低自顶向下的人体姿态估计对于人体边界框的依赖, 再利用跟踪的方法修正关键点坐标, 进一步提高人体姿态估计的准确率.

随着多人人体姿态估计的发展, 研究者们开始聚焦于多人人体姿态估计的主要难点——人体的遮挡和重叠等. 上海交通大学卢策吾组的 Li 等人[42]发现: 目前, 主流的人体姿态估计方法在面对拥挤的人群这种大面积遮挡和重叠的环境时, 人体姿态估计的准确率会大幅度下降, 意味着目前的人体姿态估计模型对于解决拥挤人群的人体姿态估计的能力有限. 若要解决这个问题, 一个针对拥挤人群的人体姿态估计数据集是必不可少的, 因而他们发布了 CrowdPose 数据集, 专门用于拥挤人群的人体姿态估计的研究, 并提出了一种针对拥挤人群的人体姿态估计算法——AlphaPose+. 传统的自顶向下的人体姿态估计方法无法基于拥挤人群实现准

确的人体姿态估计的主要原因是: 传统的自顶向下方法往往需要基于一个边界框或人体区域进行单人人体姿态估计, 这也就限制了人体姿态估计的特征区域. 而拥挤情况下, 人体检测器常常会将一个人的部分人体关键点划分到其他人的边界框或人体区域中, 这也就导致了姿态估计的结果出现了较大的偏差. 针对这个问题, 如图 12 所示, 他们提出一个关键点候选的单人人体姿态估计方法, 引入了新的热图损失去抑制却不完全消除人体边界框中属于其他人的干扰关键点坐标, 并提出了一种全局关联算法, 在关键点候选的单人人体姿态估计的基础上, 跨边界框全局关联属于同一个人的人体关键点坐标. AlphaPose+借此也成为当时针对拥挤人群最好的人体姿态估计方法.

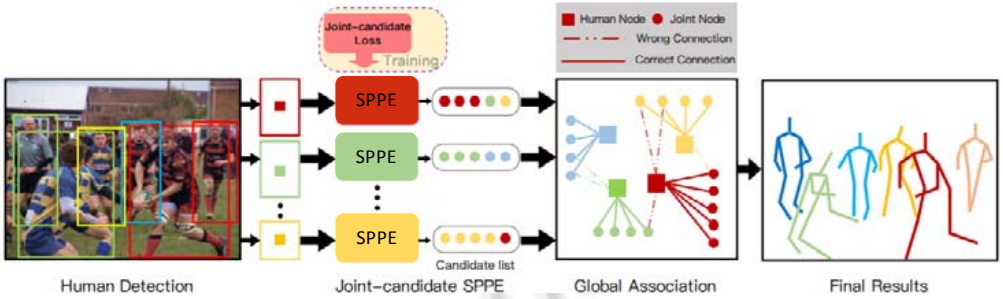


图 12 AlphaPose+的基本结构[42]

拥挤人群也就意味着存在许多的遮挡, 遮挡关键点的估计, 也是二维人体姿态估计的主要难点之一. 而遮挡的人体关键点坐标往往无法通过简单的视觉信息去获取, 需要依赖于人体的结构信息和已有的关键点坐标去推导. 而图神经网络就是发掘人体结构信息去推导被遮挡的关键点坐标的一种有效的手段, 因而, Qiu 等人[43]提出一个 OPEC-Net 的姿态估计框架, 以 AlphaPose+作为初始化的姿态估计网络, 再利用图神经网络结合图像对初步估计得到的关键点进行修正, 从而推断获得更精确的关键点坐标, 进而刷新了 CrowdPose 数据集上人体姿态估计的准确率.

虽然目前自定向下的人体姿态估计方法已经取得了较好的效果, 但其仍存在许多局限性: 首先, 自顶向下的人体姿态估计的准确率严重依赖于第 1 步人体检测获得的人体边界框的精确度; 其次, 由于自顶向下的人体姿态估计是以人体边界框或人体区域对应的子图作为输入的, 那么有几个人体边界框或人体区域, 人体姿态估计的网络就要运行几次, 也就导致了自顶向下的方法应对密集人群时效率大大降低.

2.3 自底向上的二维人体姿态估计方法

另一种主流的多人人体姿态估计方式就是自底向上的人体姿态估计方式, 自底向上的人体姿态估计方式不同于自顶向下的人体姿态估计方式, 它是将原图像作为输入, 首先估计出图中所有的关键点坐标, 然后再对关键点坐标按人进行划分, 从而生成各个人对应的二维人体姿态估计. 自底向上的方法不依赖于获得人体的边界框, 其输入只有原始图片, 无论图中有多少个人, 自底向上方法的关键点检测过程都无须依赖于人体边界框或区域, 且只需执行一次姿态估计估计就能获得图中所有人的人体姿态, 所以与自顶向下的人体姿态估计方法相比, 其效率往往要高得多, 往往能够更好地达到实时性的要求, 因而近些年, 越来越多关于自底向上的人体姿态估计方法涌现而出.

早期基于深度学习的自底向上的多人人体姿态估计方法有德国的马克普朗克研究所的 Pishchulin 等人[44]提出的 DeepCut, DeepCut 先使用卷积神经网络检测图像中存在的所有人体关键点坐标, 再将每个关键点坐标作为一个图节点, 节点之间的关联性作为图节点之间的权重, 形成的一个密集连接图, 将关键点的划分看成一个整数线性规划(integer linear program, ILP)问题, 使用数学方法将属于同一个人的关键点关联起来, 以获得最后的多人人体姿态估计结果.

随着深度学习的发展, 尤其是 ResNet 的提出, 在一定程度上解决了神经网络过深带来的梯度消失问题, 更深的神经网络也带来了更强的性能提升, 因而在 DeepCut 之后, 马克普朗克研究所的 Insafutdinov 等人[45]

又提出了精度更高、速度更快的 DeeperCut. DeeperCut 引入了更深的 ResNet, 提高了人体部分的检测精度. 并针对 DeepCut 中利用关键点的密集连接图, 将关键点的划分看成一个 ILP 问题而导致计算复杂度过高的情况, 提出了一种基于图像的成对关系匹配方法, 引入深度神经网络去预测成对的部分到部分的关联, 去计算关键点匹配的可能性, 同时也采用了一种增量优化策略去探索关键点匹配的搜索空间, 提高了人体姿态估计的效率.

虽然之前的 DeeperCut 在 DeepCut 的基础上实现了超 20 倍的速度提升, 但是距离实时的多人人体姿态估计仍有一定的距离. 2017 年, 卡耐基梅隆大学[46]提出的一个包含实时的多人人体姿态估计的系统, 即著名的 OpenPose 系统, 如图 13 所示, 是 OpenPose 中多人人体姿态估计基本流程. 为了达到实时性要求, OpenPose 使用了多阶段反复迭代的卷积神经网络结构, 该网络有两个分支, 分别用于计算部分置信图(part confidence maps)和部分关系场(part affinity field). 部分置信图的回归目标是人体关键点对应的热图, 其根据生成热图中超过一定置信度的所有波峰的位置, 去估计图中存在的关键点坐标; 而部分关系场用于关联属于同一个人的人体关键点, 其实现的方式是为每一个位置生成一个单位向量, 该向量的回归目标是指向与该关键点关联的关键点方向的单位向量. OpenPose 估计时, 利用部分置信图获得最高且超过一定阈值的关键点坐标作为初始关键点坐标, 通过部分关系场输出单位向量和该关键点指向其他关键点方向的单位向量点乘, 并沿着枝干方向求均值,获得的值最大且超过一定的阈值的点就是该关键点关联的关键点. 重复这个操作, 直到无法再找到新的点, 就找到了图中属于同一个人的所有关键点坐标. OpenPose 通过部分置信图和部分关系场相结合, 通过一次姿态估计过程估计出一张图中所有人的姿态, 大大提高了关键点的估计效率.

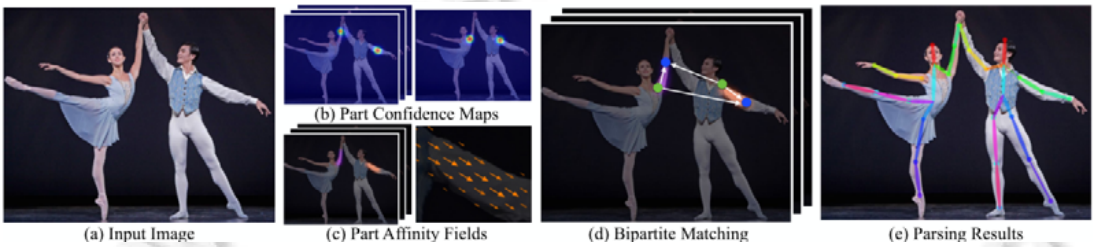


图 13 OpenPose 中多人人体姿态估计基本流程[46]

此外, 虽然不像单人人体姿态估计和自顶向下的多人人体姿态估计之间存在直接的联系, 单人人体姿态估计的发展也同样推动了自底向上的多人人体姿态估计的发展. 堆叠沙漏网络的提出者 Newell 等人[47]在堆叠沙漏网络的基础上提出了依赖于联系嵌入向量(associative embedding)的一种端到端的自底向上人体姿态估计方法, 该方法将原本分为检测关键点和分组两步的自底向上的多人人体姿态估计变成端到端的过程, 在估计关键点的同时生成联系嵌入向量, 指示关键点属于哪个人体实例, 既提高了人体姿态估计的效率, 又避免了相互依赖的两个步骤会造成误差的叠加, 从而提高了姿态估计的准确率.

同时估计关键点和隐向量能提升姿态估计效果, 是因为关键点估计和关键点分组之间的学习本身包含一定的相互促进的关系, 因而, 中东技术大学(Middle East Technical University)的 Kocabas 等人[48]提出了多姿态网络(MultiPoseNet), 利用一个多任务模型实现自底向上的人体姿态估计. 该模型先用骨干网络提取图片特征, 然后基于该图片特征, 分别用姿态估计网络将图片中的所有关键点检测出来, 人体检测网络检测每个人体的边界框; 并提出了一个姿态残差网络(pose residual network, PRN), 学习姿态的结构特征, 进而消除人体边界框内属于其他人的关键点响应值, 获得每一个人对应人体姿态坐标.

Google[49]调整了之前用于自顶向下的多人人体姿态估计的 G-RMI 算法, 提出了著名的 PersonLab, 同时引入了短偏移和中偏移, 将偏移从关键点的周边升级到了躯干之间, 类似于 OpenPose 的部分关系场, 确定了下一个关键点的位置, 从而减少人体关键点误匹配的问题.

受到 PersonLab 等算法的启发, Kreiss 等人[50]扩张了其中场的概念, 提出了级联场(composite fields, PifPaf)用于自底向上的多人人体姿态估计. 级联场包括两个部分.

- 第 1 个是部分密集场(part intensity fields, Pif), 该场相当于 PersonLab 中提出的短偏移的进一步发展. 部分密集场除了偏移向量, 其还引入了关键点所对应人的尺度信息, 根据人体的偏移向量和估计的人体尺度信息, 将关键点估计的结果映射到更高分辨率的热图上, 实现对热图的精修. 在小分辨率的特征图上估计高分辨的关键点坐标位置, 提高人体姿态估计的精度.
- 第 2 个是部分联系场(part association fields, Paf), 如图 14 所示, 部分联系场将 Personlab 中的中偏移中关键点到关键点的关联变成了中间像素点到两个关键点之间的关联. 中偏移中关键点到关键点之间的关联只有一个单向向量, 也就是需要默认出发点在原来关键点所在特征图点的中心位置, 然后去寻找下一个关键点坐标, 这个地方实际上对出发关键点的位置做了一个近似. 而部分联系场中, 使用关键点之间的点作为参照点, 利用两个向量使得相关的关键点位置都是理论上精确的, 避免了这一近似过程, 精确定位了相联系的两个关键点的相对位置.

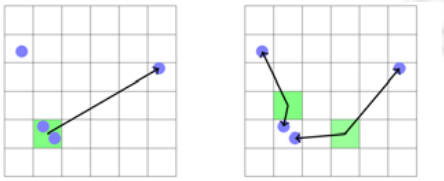


图 14 中偏移和部分联系场的对比[50]

类似的堆叠沙漏网络和 G-RMI, 高分率特征表示学习网络也被扩展用于实现自底向上的多人人体姿态估计. Cheng 等人[51]优化了单人人体姿态估计的高分率特征表示学习网络, 利用反卷积实现上采样, 引入多尺度监督, 并结合联系嵌入向量[47], 提出了一个尺度感知的更高的高分辨率特征表示学习网络, HigherHRNet. HigherHRNet 也成为当时最优秀的自底向上的多人人体姿态估计方法. 但是 HigherHRNet 作为一个自底向上的方法, 比起最新的自顶向下多人人体姿态估计方法, 其性能表现仍有所差距, 但是在 CrowdPose 数据集上, 其表现超越了所有自顶向下和自底向上的多人人体姿态方法, 成为 CrowdPose 数据集上最优秀的多人人体估计方法. 这也在一定程度上说明: 对于拥挤的场景, 自底向上的方法与自顶向下的方法相比可能具有更强的适应性.

自底向上的人体姿态估计方法中, 除了一系列基于场的分组方法之外, Nie 等人[52]提出一种基于中心点估计的多人人体姿态估计方法, 姿态分区网络(pose partition networks, PPN). 虽然原文中提出该方法是独立于自顶向下和自底向上的一种新的多人人体姿态估计方法, 但是从估计方式上看, 该方法实际上是一个优化了的关键点分组方式的自底向上的人体姿态估计方法. 类似于之前的方法, 姿态分区网络先估计所有候选关键点, 不同的是, 该方法不是为每个关键点去选择去生成属于同一个人的相邻关键点之间的联系, 而是为每一个关键点生成指向对应人体中心的质心嵌入向量, 然后根据该向量获得人体中心, 并基于人体中心进行局部贪心推导实现关键点的划分, 大大降低了关键点分组的复杂度.

还有 Varamesh 等人[53]考虑到姿态数据实际上是存在多种不同的状态, 如不同的视角和动作, 而使用单个模型去进行人体姿态估计往往会导致姿态的特征表示过于复杂而难以学习, 设计网络根据不同模式学习不同的组件, 将降低学习特征表示的复杂度, 简化学习过程, 提高姿态估计的效率和准确率. 因而他们提出了一个混合密集网络(mixture dense network, MDN)和一个用于密集空间回归的混合公式, 将输入划分成具有意义的多个模式, 根据输入调整合适的输出头部进行人体姿态估计, 并实验验证了该混合密集网络模型的有效性. 通过分析学习得到的混合密集网络的参数信息, 他们发现人体姿态的视角是混合密集网络划分模式的主要因素, 说明了现实世界的姿态数据确实是多模式的, 并提出: 设计网络根据不同模式学习不同的组件, 将会是未来研究的一个重要方向.

虽然对比自顶向下的方法, 自底向上的方法只需进行一次人体关键点的估计就能实现多人人体姿态估计, 因而速度较快, 但同时也意味着自底向上的方法需要一次性去估计未知数量的关键点坐标, 与自顶向下

一个特征图只估计一个关键点的方式相比具有更高的复杂度, 因而大部分情况下的自顶向下的方法与自底向上的方法相比都有较高的人体姿态估计准确率. 但是在拥挤条件下, 自顶向下的方法中的人体检测器常常会将一个人的部分人体关键点划分到其他人的边界框或人体区域中, 这也就导致了基于边界框或人体区域对应的局部特征进行的姿态估计会出现较大的偏差. 而对于自底向上的方法而言, 无论关键点的估计还是分组的过程都是基于全局特征的, 所以与自顶向下的方法相比, 其对于拥挤人群的人体姿态估计具有更强的适应性.

3 实验结果对比

本节主要介绍现在二维人体姿态估计领域的 3 个主流数据集: MPII 人体姿态估计数据集、COCO2017 数据集以及最新提出来的 CrowdPose 数据集. 并对比主流的人体姿态估计方法在这 3 个数据集上的表现, 进而分析近年来二维人体姿态估计领域的发展.

MPII 人体姿态估计数据集是记录真实世界中人类活动的图片数据集, 图片主要来源于 YouTube 的视频, 包括了大约 25 000 张图片, 包含超过 40 000 个带有人体关键点注释的人. 这些图片按照既定的人类日常活动去收集, 每张图片都有一个对应活动标签, 包含了人们日常生活的方方面面. 如图 15 左侧是 MPII 数据集中定义的人体姿态. MPII 人体姿态估计数据集的人体姿态注释包含 16 个关键点, 分别为 0–右踝, 1–右膝, 2–右臀, 3–左臀, 4–左膝, 5–左踝, 6–骨盆, 7–胸膛, 8–脖子上部, 9–头顶, 10–右腕, 11–右肘, 12–右肩, 13–左肩, 14–左肘, 15–左腕.

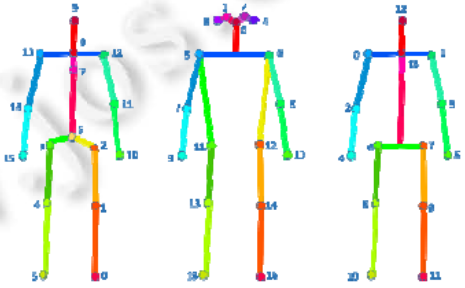


图 15 人体姿态. 从左到右依次对应为 MPII, COCO2017, CrowdPose 数据集对于人体姿态的定义

MPII 人体姿态数据集上主要评估两个任务, 分别是单人人体姿态估计和多人人体姿态估计. 对于单人人体姿态估计, MPII 提供真实的人体边界框, 对应的评价指标为 *PCKh*@0.5. *PCK*, 即关键点正确估计的比例(percentage of correct keypoints), *h* 对应于头部, @0.5 代表 0.5 倍的头部长度, *PCKh*@0.5 的意义就是误差少于头部长度一半的关键点所占的比例. *PCKh*@α的表达式如公式(1)所示.

$$PCKh@\alpha=\frac{\sum_{p\in P}\Lambda(\|p-\bar{p}\|_2<\alpha\times l_{head})}{n} \tag{1}$$

其中, P 代表了估计得到的所有的关键点坐标, p 代表了其中一个关键点坐标, $\bar{p}$ 代表对应关键点坐标的真实坐标; α代表长度比例系数; l_{head} 代表对应人的头部长度; Λ代表一元函数, 参数为真则返回 1, 反之返回 0; n 代表了测试的关键点总数. 表 1 中为目前本文中提到的主流人体姿态估计方法在 MPII 测试集上的实验结果, * 代表该模型在训练时引入了额外的数据.

MPII 多人人体姿态估计的评价指标是 *mAP*@0.5, 即关键点估计的平均精确率(average precision, AP)的均值, 这里的 *AP* 的计算方式类似于 *PCKh*, 这里的@0.5 也是代表 0.5 倍的头部长度, 即和真实关键点坐标误差在 0.5 倍头部长度的关键点坐标被认为是正确的关键点估计. 不同的是, 在计算 *AP* 前需要根据 *PCKh* 将估计出来的姿态和地表真实的姿态匹配, 只需计算与地表真实关键点成功匹配的关键点, 并未匹配的关键点坐标被视为假阳性估计. *mAP*@α的表达式如公式(2)所示.

$$mAP@\alpha=\frac{1}{K}\sum_{i=1}^{K}\frac{\sum_{p\in P_i}\Lambda(\|p-\bar{p}\|_2<\alpha\times l_{head})}{m_i} \tag{2}$$

其中, K 代表了一个完整的人体姿态包含的关键点的个数, 即关键点的种类数; P_i 代表了估计得到且匹配成功的第 i 类关键点的集合, p 代表了其中的一个关键点坐标, $\overline{p}$ 代表对应关键点坐标的真实坐标; α代表长度比例系数; l_{head} 代表对应人的头部长度; Λ代表一元函数, 参数为真则返回 1, 反之返回 0; m_i 代表了估计得到的所有第 i 类关键点. 表 2 中为目前本文中提到的主流人体姿态估计方法在 MPII 多人数据集测试集上的表现, Subset 代表是采用部分数据集测试(288 张测试图片)的结果.

表 1 MPII 测试集上的效果(*PCKh*@0.5)

Method	*Hea*	*Sho*	*Elb*	*Wri*	*Hip*	*Kne*	*Ank*	Total
Joint CNN[10]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
DeepCut[44]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
DeeperCut[45]	96.6	94.6	88.5	84.4	87.6	83.9	79.4	88.3
CPM[30]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
SHN[12]	98.2	96.3	92.2	87.8	90.6	88.0	82.7	90.9
FPD[29]	98.3	96.4	91.5	87.4	90.9	87.1	83.7	91.1
SimpleBaseline[31]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
Multi-Context Attention[13]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
MSSAN[18]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
HRNet[14]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
PIL[21]	98.6	96.9	93.0	89.1	91.7	89.0	86.2	92.4
SGSC[20]	98.6	97.0	93.0	89.2	91.7	88.9	86.0	92.4
SCI[19]	98.6	97.0	92.8	88.8	91.7	89.9	86.8	92.5
UniPose[15]	–	–	–	–	–	–	–	92.7
SGSC*[20]	98.8	97.5	94.4	**91.2**	**93.2**	92.2	**89.3**	**94.1**
ASDA*[23]	**98.9**	**97.6**	**94.6**	**91.2**	93.1	**92.7**	89.1	**94.1**

表 2 MPII 多人数据集测试集上的效果(*mAP*@0.5)

Method	*Hea*	*Sho*	*Elb*	*Wri*	*Hip*	*Kne*	*Ank*	*mAP*
Top-down								
AlphaPose[39]	**91.3**	**90.5**	**84.0**	**76.4**	**80.3**	**79.9**	**72.4**	**82.1**
Bottom-up								
Deepcut[44](Subset)	78.4	72.5	57.9	39.9	56.7	44.0	32.0	54.1
DeeperCut[45]	79.1	72.2	59.7	50.0	56.0	51.0	44.6	59.4
OpenPose[46]	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Embedding[47]	92.1	89.3	78.9	69.8	76.2	71.6	64.7	77.5
PPN[52]	**92.2**	**89.7**	**82.1**	**74.4**	**78.6**	**76.4**	**69.3**	**80.4**

COCO2017 数据集源于微软举办的姿态估计挑战赛, 数据集中包含了超过 200 000 张图片和 250 000 个标注了人体关键点坐标的人体实例, 其中, 公开的训练和验证集包含超过 150 000 个人和 1 700 000 个关键点坐标的注释. 图 15 中间是 COCO2017 数据集中定义的人体姿态. 不同于 MPII 对人体姿态的注释, COCO2017 数据集中的人体姿态包含 17 个关键点, 分别为 0–鼻子, 1–左眼, 2–右眼, 3–左耳, 4–右耳, 5–左肩, 6–右肩, 7–左肘, 8–右肘, 9–左腕, 10–右腕, 11–左臀, 12–右臀, 13–左膝, 14–右膝, 15–左踝, 16–右踝. COCO2017 评价指标是根据对象关键点的相似度(object keypoint similarity, OKS)计算获得的. OKS 对应的表达式如公式(3)所示.

$$OKS=\frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)U(v_i>0)}{\sum_i U(v_i>0)} \tag{3}$$

其中, v_i 代表了第 i 个关键点是否标注, 0 代表未标注点, 1 代表标注但不可见点, 2 代表了标注且可见点; $U(v_i>0)$ 代表 $v_i>0$ 即点标注时取 1, 其他时候取 0; d_i 代表了第 i 个关键点坐标和地表真实关键点坐标的差值; s 对应人体大小的目标尺度; k_i 代表的是第 i 个关键点标注时产生的标准差. 评价时, *OKS* 大于一定阈值的关键点作为正例, 反之为反例. 其中, *AP* 代表的是 *OKS* 阈值从 0.05 到 0.95、间隔 0.05 获得的关键点估计准确率的平均值, $AP^{0.5}$ 代表 *OKS* 阈值为 0.5 时的关键点估计准确率, $AP^{0.75}$ 代表 *OKS* 阈值为 0.75 时关键点估计的平均准确率, AP^M 代表中等大小人体(人体面积大于 322 小于 962)关键点估计的 *AP* 值, AP^L 代表大的人体(人体面积大于 962)关键点估计的 *AP* 值, *AR* 代表的是 *OKS* 阈值从 0.05 到 0.95、间隔 0.05 获得的关键点召回率的平均值. 表 3 是主流方法在 COCO2017 测试集上的表现, *代表了引入了额外的数据集.

CrowdPose 数据集是专门用于拥挤人群的人体姿态估计数据集, 该数据集包含 20 000 张图片和 80 000 个人体, 其中很大一部分的人体样本存在极高的重叠和遮挡. 如图 15 右侧是 CrowdPose 数据集定义的人体姿态. 不同于 MPII 和 COCO, CrowdPose 数据集的人体样本包含 14 个关键点坐标, 分别为 0–左肩, 1–右肩, 2–左肘, 3–右肘, 4–左腕, 5–右腕, 6–左臀, 7–右臀, 8–左膝, 9–右膝, 10–左踝, 11–右踝, 12–头, 13–脖子. CrowdPose 数据集采用了和 COCO2017 一样的评价指标, 其中, $AR^{0.5}$ 代表 *OKS* 阈值为 0.5 时的关键点平均召回率, $AR^{0.75}$ 代表 *OKS* 阈值为 0.75 时的关键点平均召回率. 表 4 是主流方法在 CrowdPose 数据集上的表现.

表 3 在 COCO2017 测试集上的效果

Method	*AP*	AP^{50}	AP^{75}	AP^{M}	AP^{L}	*AR*
Top-down						
Mask R-CNN[38]	63.1	87.3	68.7	57.8	71.4	–
G-RMI[40]	64.9	85.5	71.3	62.3	70.0	69.7
AlphaPose+[42]	72.2	90.1	79.3	–	–	–
AlphaPose[39]	72.3	89.2	79.1	68.0	78.6	–
CPN[30]	73.0	91.7	80.9	69.5	78.1	79.0
Simple Baseline[31]	73.7	91.9	81.1	70.3	80.0	79.0
OPEC-Net[43]	73.9	91.9	82.2	–	–	–
CS Attention[34]	74.6	91.8	82.1	70.9	80.6	80.7
HRNet[14]	75.5	92.5	83.3	71.9	81.5	80.5
DarkPose[25]	76.2	92.5	83.6	72.5	82.4	81.1
UDP[26]	76.5	**92.7**	84.0	73.0	82.4	81.6
HRNet*[14]	77.0	**92.7**	84.5	73.4	83.1	82.0
DarkPose*[25]	**77.4**	92.6	**84.6**	**73.6**	**83.7**	**82.3**
Bottom-up						
OpenPose[46]	61.8	84.9	67.5	57.1	68.2	66.5
MDN[53]	62.9	85.1	69.4	58.8	71.4	–
Embedding[47]	65.5	86.8	72.3	60.6	72.6	70.2
PifPaf[50]	66.7	–	–	62.4	72.9	–
PersonLab[49]	68.7	89.0	75.4	64.1	75.5	75.4
MultiPoseNet[48]	69.6	86.3	76.6	65.0	**76.3**	**75.5**
HigherHRNet[51]	**70.5**	**89.3**	**77.2**	**66.6**	75.8	74.9

表 4 在 CrowdPose 测试集上的效果

Method	*AP*	AP^{50}	AP^{75}	*AR*	AR^{50}	AR^{75}
Top-down						
Mask R-CNN[38]	57.2	83.5	60.3	65.9	89.5	69.4
Simple Baseline[31]	60.8	81.4	65.7	67.3	86.3	71.8
AlphaPose[39]	61.0	81.3	66.0	67.6	86.7	71.8
AlphaPose+[42]	66.0	84.2	71.5	**72.7**	**89.5**	**77.5**
OPEC-Net[43]	**70.6**	**86.8**	**75.6**	–	–	–
Bottom-up						
HigherHRNet[51]	**67.6**	**87.4**	**72.6**	–	–	–

根据实验结果对比可以看出: 无论是单人人体姿态估计、自顶向下的多人人体姿态估计还是自底向上的多人人体姿态估计, 近年来都有了长足的进步. 从 MPII 数据集上单人人体姿态估计的实验结果分析可以发现: 一些目前主流的人体姿态估计方法仅在 MPII 数据集训练得到的人体姿态估计模型在 MPII 测试集上已经能实现 92%+的人体姿态估计准确, 且新提出的方法在 MPII 测试集准确率的上升开始逐渐放缓, 说明了基于 MPII 数据集的单人人体姿态估计的研究逐渐成熟. 深度学习是受数据驱动的, 一些方法通过引入额外的人体姿态数据, 大大提高了 MPII 测试集上的准确率, 使得 MPII 测试集上的准确率到达了 94%+. 这说明仅仅依赖于 MPII 数据集, 很难实现一些 MPII 测试集中遮挡、重叠等困难关键点的估计. 引入额外的数据可以大大提高 MPII 测试集上人体姿态估计的准确率, 这也意味着整合多源数据扩充样本, 对于人体姿态估计的研究将非常重要.

对比 MPII 数据集上多人人体姿态估计的结果和 COCO2017 测试集上的结果可以发现, 同时期, 自顶向下的基于深度学习人体姿态估计方法与自底向上的基于深度学习人体姿态估计方法相比准确度更高. 因为自底向上方法要在一张图中同时估计多个人的关键点坐标, 其复杂性高于自顶向下的多人人体姿态估计, 自然也

就导致了其准确率相对较低. 但是自底向上的人体姿态估计方法由于可以一次性获得图中所有关键点坐标, 所以在进行多人人体姿态估计时, 自底向上的方法的效率比自顶向下的方法要高得多.

此外, 在 CrowdPose 数据集和 COCO2017 数据集上对比 HigherHRNet[51]和 Simple Baseline[31]可以发现: 虽然 Simple Baseline 在 COCO2017 测试集上的表现更好(*AP* 73.7 vs. *AP* 70.5), 但是在包含大量拥挤样本的 CrowdPose 数据集上, 却远远地被 HigherHRNet 甩开(*AP* 60.8 vs. *AP* 67.6), 在一定程度上说明了对于拥挤人群, 自底向上的多人人体姿态估计方法比自顶向下的多人人体姿态估计方法具有更强的适应性. 主要原因就像第 2.3 节中提到的那样: 自顶向下的人体姿态估计方法通过人体检测缩小了特征空间, 虽然降低了关键点估计的复杂度, 但是在拥挤的场景中, 人体检测框可能无法将目标人的人体部分包含在一个人体边界框或人体区域内, 导致一些人体部分的信息丢失; 再加之其他人的人体部分的混入, 导致了一般的自顶向下的人体姿态估计方法往往无法适应拥挤条件下的多人人体姿态. 而自底向上的多人人体姿态估计方法大多数本身就是基于全局去估计关键点和分组的, 因而在面对拥挤人群时, 自底向上的方法具有更强的适应性.

4 结论与未来的研究方向

近年来, 基于深度学习的二维人体姿态估计处于高速发展时期, 很多优秀的人体姿态估计方法被提出, 不同数据集的榜单也不断地被刷新. 多尺度融合和监督的改进、注意力机制的引入、人体结构特征的探究、人体解析的联合估计、用于姿态估计的数据增强以及基于知识蒸馏的模型压缩等, 大大促进了单人人体姿态估计的发展. 单人人体姿态估计发展是多人人体姿态估计的基础, 如堆叠沙漏网络[12]、高分辨率特征表示网络[14]等单人人体姿态估计的方法都被推广到多人人体姿态估计中, 直接推动了多人人体姿态估计的发展.

目前, 多人二维人体姿态估计方法的研究包括自顶向下的多人人体姿态估计方式和自底向上的多人人体姿态估计方式, 两者互有利弊, 适用于不同的场景. 大多数场景下, 自顶向下的人体姿态估计比自底向上的人体姿态估计方法要更加精确, 但是自底向上的人体姿态估计方法由于可以一次性估计图中所有关键点坐标, 在进行多人人体姿态估计时, 速度上更具优势. 而且由于自底向上的多人人体姿态估计方法是基于全局特征去估计关键点和分组的, 而自顶向下的多人人体姿态估计方法依赖于边界框分组, 基于局部特征进行关键点的估计, 因而自底向上的多人人体姿态估计方法在面对拥挤人群的多人人体姿态估计时性能下降更小,具有更强的适应性. 但是自底向上的人体姿态估计方法由于本身的姿态估计精度较低, 这也一定程度上限制了其对拥挤人群多人人体姿态估计的上界.

拥挤人群等带来的遮挡重叠等问题, 是未来二维人体姿态估计研究的一个重要方向. 针对拥挤人群提出的 AlphaPose+[42]实际上是一种引入部分自底向上方法的特性的自顶向下的方法. 它为了克服自顶向下方法中局部特征导致信息缺失而无法准确估计拥挤状态下人体关键点的位置的问题, 利用自顶向下的方式估计边界框内的主体人的关键点坐标, 再利用类似自底向上的方式中的分组方法, 基于全局特征关联关键点坐标, 从而优化原本自顶向下的方法中被误分到其他区域的关键点坐标, 实现针对拥挤人群的多人人体姿态估计方法. 这样的人体姿态估计方式一定程度上融合了自顶向下和自底向上的优势, 为了解决拥挤人群下的多人人体姿态估计提供了一个基本方案. 随后, OPEC-Net[43]在 AlphaPose+基础上引入图神经网络学习人体的结构特征, 进一步刷新了 CrowdPose 数据集上人体姿态估计的准确率. 我们认为, 未来如果想要进一步解决拥挤人群下的多人人体姿态估计问题, 可以考虑融合现有的自顶向下和自底向上的人体姿态估计方法, 以自顶向下的人体姿态估计方法作为主框架, 缩小特征空间, 精确估计关键点坐标, 再引入自底向上方法中的分组方式, 利用全局特征优化估计关键点坐标.

还有多模式学习的人体姿态估计算法, 也将是未来研究的一个重要方向. Varamesh 等人[53]提出并验证了混合密集网络的有效性, 说明了现实世界的数据是多模式的, 使用单个模型去进行人体姿态估计往往会导致姿态的特征表示过于复杂而难以学习. 因而, 如何设计网络根据不同的姿态模式学习不同的组件、降低学习特征表示的复杂度、简化学习过程, 将是未来基于多模式学习的人体姿态估计算法研究的重点.

此外, 所有深度学习的算法都是受数据驱动的. 如何有效地利用现有的数据是非常重要的. Nie 等人[21]提

出了解析诱导学习器, 用人体解析的特征学习一个自适应卷积核来辅助人体姿态估计, 实际上就是引入人体解析学习到的信息来指导人体姿态估计. 人体解析和人体姿态估计接近, 所以可以利用人体解析的学习辅助人体姿态估计的学习. 此外, 不同数据集上人体姿态的标注更加接近, 因而未来也可以探究以类似自适应卷积核的形式级联多个人体姿态估计数据集上人体姿态估计模型的学习, 从而融合多个数据集上相似却不同的姿态信息, 以多任务学习的形式, 提高训练获得的人体姿态估计模型的准确率.

人体的结构特征仍是未来研究的一个重要方向. 虽然现在已有不少的基于人体结构特征的二维人体姿态估计方法, 但是这些方法主要是构建人体关键点之间的数据传输关系, 试图直接通过神经网络去学习关键点之间的隐式联系. 但是在现实中, 实际上除了关键点之间的连接以外, 我们有更多关于人体姿态的先验知识, 如人体在三维空间内, 左、右骨骼长度是相等的. 因而, 如何在构建一个隐式的三维人体模型时引入更多人体结构的先验知识去辅助二维人体姿态估计, 仍有待进一步探寻.

References:

[1] Zhao X, Liu Y, Fu Y. Exploring discriminative pose sub-patterns for effective action classification. In: Proc. of the ACM Multimedia. Barcelona: ACM, 2013. 273−282. [doi: 10.1145/2502081.2502094]

[2] Wang L, Wu J, Zhou ZM, Zhao X, Liu YC. Human action recognition through part-configured human detection response feature maps. Ruan Jian Xue Bao/Journal of Software, 2015, 26: 128−136 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15023.htm

[3] Desai C, Ramanan D. Detecting actions, poses, and objects with relational phraselets. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, eds. Proc. of the European Conf. on Computer Vision. Florence: Springer, 2012. 158−172.

[4] Zhang QJ, Zhang L. Convolutional adaptive denoising autoencoders for hierarchical feature extraction. Frontiers of Computer Science, 2018, 12(6): 1140−1148.

[5] Huang LL, Peng JF, Zhang RM, *et al*. Learning deep representations for semantic image parsing: A comprehensive overview. Frontiers of Computer Science, 2018, 12(5): 840−857.

[6] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014.

[7] Andriluka M, Pishchulin L, Gehler PV, Schiele B. 2D human pose estimation: New benchmark and state of the art analysis. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 3686−3693.

[8] Lin TY, Maire M, Belongie S, *et al*. Microsoft coco: Common objects in context. In: Proc. of the European Conf. on Computer Vision. Cham: Springer, 2014. 740−755.

[9] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84−90.

[10] Tompson JJ, Jain A, LeCun Y, *et al*. Joint training of a convolutional network and a graphical model for human pose estimation. Advances in Neural Information Processing Systems, 2014, 27: 1799−1807.

[11] Wei SE, Ramakrishna V, Kanade T, *et al*. Convolutional pose machines. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4724−4732.

[12] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: Proc. of the European Conf. on Computer Vision. Cham: Springer, 2016. 483−499.

[13] Chu X, Yang W, Ouyang W, *et al*. Multi-context attention for human pose estimation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1831−1840.

[14] Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 5693−5703.

[15] Artacho B, Savakis A. UniPose: Unified human pose estimation in single images and videos. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 7035−7044.

[16] Artacho B, Savakis A. Waterfall atrous spatial pooling architecture for efficient semantic segmentation. Sensors, 2019, 19(24): 5361.

[17] Chu X, Ouyang W, Li H, *et al*. Structured feature learning for pose estimation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4715−4723.

[18] Ke L, Chang MC, Qi H, *et al*. Multi-scale structure-aware network for human pose estimation. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 713−728.

[19] Zhang H, Ouyang H, Liu S, *et al*. Human pose estimation with spatial contextual information. arXiv:1901.01760, 2019.

[20] Bulat A, Kossaifi J, Tzimiropoulos G, *et al*. Toward fast and accurate human pose estimation via soft-gated skip connections. arXiv:2002.11098, 2020.

[21] Nie X, Feng J, Zuo Y, *et al*. Human pose estimation with parsing induced learner. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 2100−2108.

[22] Nie X, Feng J, Yan S. Mutual learning to adapt for joint human parsing and pose estimation. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 502−517.

[23] Bin Y, Cao X, Chen X, *et al*. Adversarial semantic data augmentation for human pose estimation. In: Proc. of the European Conf. on Computer Vision. Cham: Springer, 2020. 606−622.

[24] Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Advances in Neural Information Processing Systems, 2014, 27: 2672−2680.

[25] Zhang F, Zhu X, Dai H, *et al*. Distribution-aware coordinate representation for human pose estimation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 7093−7102.

[26] Huang J, Zhu Z, Guo F, *et al*. The devil is in the details: Delving into unbiased data processing for human pose estimation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 5700−5709.

[27] Ge DH, Li HS, Zhang L, Liu RY, Shen PY, Miao QG. Survey of lightweight neural network. Ruan Jian Xue Bao/Journal of Software, 2020, 31(9): 2627−2653 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5942.htm [doi: 10.13328/j.cnki.jos.005942]

[28] Zhang ZK, Pang WG, Xie WJ, Lü MS, Wang Y. Deep learning for real-time applications: A survey. Ruan Jian Xue Bao/Journal of Software, 2020, 31(9): 2654−2677 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5946.htm [doi: 10.13328/j.cnki.jos.005946]

[29] Zhang F, Zhu X, Ye M. Fast human pose estimation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 3517−3526.

[30] Chen Y, Wang Z, Peng Y, *et al*. Cascaded pyramid network for multi-person pose estimation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7103−7112.

[31] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 466−481.

[32] Ren S, He K, Girshick R, *et al*. Faster *r*-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137−1149.

[33] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770−778.

[34] Su K, Yu D, Xu Z, *et al*. Multi-person pose estimation with enhanced channel-wise and spatial information. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 5674−5682.

[35] Zhang X, Zhou X, Lin M, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6848−6856.

[36] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7132−7141.

[37] Chen L, Zhang H, Xiao J, *et al*. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 5659−5667.

[38] He K, Gkioxari G, Dollár P, *et al*. Mask R-CNN. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2961−2969.

[39] Fang HS, Xie S, Tai YW, *et al*. RMPE: Regional multi-person pose estimation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2334−2343.

[40] Papandreou G, Zhu T, Kanazawa N, *et al*. Towards accurate multi-person pose estimation in the wild. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 4903−4911.

[41] Wang M, Tighe J, Modolo D. Combining detection and tracking for human pose estimation in videos. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 11088−11096.

[42] Li J, Wang C, Zhu H, *et al*. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 10863−10872.

[43] Qiu L, Zhang X, Li Y, *et al*. Peeking into occluded joints: A novel framework for crowd pose estimation. arXiv:2003.10506, 2020.

[44] Pishchulin L, Insafutdinov E, Tang S, *et al*. DeepCut: Joint subset partition and labeling for multi person pose estimation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4929−4937.

[45] Insafutdinov E, Pishchulin L, Andres B, *et al*. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In: Proc. of the European Conf. on Computer Vision. Cham: Springer, 2016. 34−50.

[46] Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 7291−7299.

[47] Newell A, Huang Z, Deng J. Associative embedding: End-to-end learning for joint detection and grouping. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 2277−2287.

[48] Kocabas M, Karagoz S, Akbas E. MultiPoseNet: Fast multi-person pose estimation using pose residual network. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 417−433.

[49] Papandreou G, Zhu T, Chen LC, *et al*. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 269−286.

[50] Kreiss S, Bertoni L, Alahi A. Pifpaf: Composite fields for human pose estimation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 11977−11986.

[51] Cheng B, Xiao B, Wang J, *et al*. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. arXiv:1908.10357, 2019.

[52] Nie X, Feng J, Xing J, *et al*. Pose partition networks for multi-person pose estimation. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 684−699.

[53] Varamesh A, Tuytelaars T. Mixture dense regression for object detection and human pose estimation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 13086−13095.

附中文参考文献:

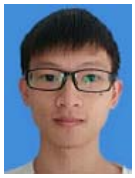
[2] 王磊, 吴俊, 周志敏, 赵旭, 刘允才. 人体检测部分响应特征映射的人体动作识别. 软件学报, 2015, 26: 128−136. http://www.jos.org.cn/1000-9825/15023.htm

[27] 葛道辉, 李洪升, 张亮, 刘如意, 沈沛意, 苗启广. 轻量级神经网络架构综述. 软件学报, 2020, 31(9): 2627−2653. http://www.jos.org.cn/1000-9825/5942.htm [doi: 10.13328/j.cnki.jos.005942]

[28] 张政馗, 庞为光, 谢文静, 吕鸣松, 王义. 面向实时应用的深度学习研究综述. 软件学报, 2020, 31(9): 2654−2677. http://www.jos.org.cn/1000-9825/5946.htm [doi: 10.13328/j.cnki.jos.005946]

张宇(1986－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为计算机视觉, 机器学习, 深度学习.

张敏灵(1979－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为机器学习, 数据挖掘.



温光照(1997－), 男, 硕士, 主要研究领域为计算机视觉.

耿新(1978－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为机器学习, 模式识别, 计算机视觉.

米思娅(1988－), 女, 博士, 讲师, 主要研究领域为用于网络安全的数据处理, 计算机视觉.