基于深度学习的二维人体姿态估计:现状及展望

李佳宁 王东凯 张史梁
(北京大学计算机学院 北京 100871)

摘 要 二维人体姿态估计旨在从摄像机拍摄的图像中识别并定位每个行人的人体关键点.作为行人分析和理解领域的基础任务之一,人体姿态估计能够为多个下游任务和应用提供支持.近年来,随着深度学习技术的进步,人体姿态估计的研究迎来快速发展.基于图像包含的行人数量,人体姿态估计可以分为单人姿态估计和多人姿态估计两大类.本文首先介绍人体姿态估计的研究背景、问题定义、任务难点以及当前方法中的关键点表示方法.在此基础上,本文进一步总结和介绍了具有代表性的单人姿态估计和多人姿态估计方法.单人姿态估计方法包括回归法和检测法,主要关注于网络结构设计、热力图编解码、多任务学习等.对于多人姿态估计,本文分别介绍了基于热力图预测的方法和基于向量场回归的方法.随后,本文总结了当前常用的代表性数据集和性能度量方法,总结了代表性方法在几个常用数据集上的性能,对它们的预测错误的场景进行了详细分析和对比.最终,本文分析了现有二维人体姿态估计算法仍未有效解决的难题,对未来研究进行了展望.

关键词 单人姿态估计;多人姿态估计;深度学习;自顶向下;自底向上;向量场回归
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.00231

Deep-Learning-Based 2D Human Pose Estimation: Present and Future

LI Jia-Ning WANG Dong-Kai ZHANG Shi-Liang
(*School of Computer Science, Peking University, Beijing* 100871)

Abstract 2D human pose estimation aims to identify and locate the human body keypoints of each person in images. As a fundamental task in human analysis and understanding field, human pose estimation can support multiple downstream tasks and can be applied to many real-world applications. In recent years, thanks to the developments of deep learning techniques, significant progresses have been made to human pose estimation. Based on the number of persons in image, human pose estimation tasks can be summarized into single-person pose estimation and the more challenging multi-person pose estimation, respectively. This paper first introduces the research background, problem definition, task difficulty and keypoint representation of human pose estimation task. Next, we introduce the representative single-person and multi-person pose estimation methods, respectively. The single-person pose estimation section introduces regression-based and detection-based methods, including network structure designing, heatmap encoding/decoding and multi-task learning categories. The multi-person pose estimation section introduces heatmap based methods and regression-based methods. We further summarize the widely-used datasets, benchmark metric, and the performance of representative methods on these datasets. This paper also selects representative methods from each category, and analyzes and compares the failure cases of these methods. Finally, this paper discusses the remaining challenges and promising research directions in human pose estimation.

收稿日期:2022-06-20;在线发布日期:2023-05-16.本课题得到国家自然科学基金(U20B2052,61936011)、国家重点研发计划(2018YFE0118400)资助.**李佳宁**,博士,主要研究方向为人体姿态估计、行人重识别.E-mail:ljn-vmc@pku.edu.cn.**王东凯**,博士研究生,主要研究方向为人体姿态估计、行人重识别.E-mail:dongkai.wang@pku.edu.cn.**张史梁**(通信作者),博士,长聘副教授,中国计算机学会(CCF)会员,主要研究方向为多媒体信息检索、计算机视觉.E-mail:slzhang.jdl@pku.edu.cn.

Keywords single-person pose estimation; multi-person pose estimation; deep learning; top-down; bottom-up; regression

1 引 言

人体姿态估计(Human Pose Estimation)旨在从单目摄像机拍摄的二维图像中预测每个行人的人体关键点坐标.作为行人分析与理解领域的基础任务,人体姿态估计技术能够为计算机视觉领域的多个下游任务提供支持,比如行人重识别[1-3]、人体解析[4]和动作识别[5-6]等.由于基于计算机视觉的人体姿态估计不需要额外的穿戴设备,该技术比传统的穿戴式动作捕捉技术成本更加低廉且灵活性更高,因此被广泛地用于各类现实应用,包括虚拟现实、人机交互和数字娱乐等.基于以上原因,人体姿态估计任务已经成为计算机视觉领域重要的基础任务之一,吸引了许多研究者们的关注.

根据图像中出现行人的数量,人体姿态估计任务可以分为单人姿态估计(Single Person Pose Estimation,SPPE)和多人姿态估计(Multiple Person Pose Estimation,MPPE)两大类.单人姿态估计处理的图像仅包含单一行人且场景较为简单,只需预测人体关键点的位置,无需预测关键点的归属行人.多人姿态估计往往面向开放场景,处理的图片包含不定数量的行人,需要同时预测每个人体关键点的位置和其归属行人.与单人姿态估计相比,多人姿态估计面临行人遮挡、表观相似行人难以区分等难题,更具挑战性.

早期的研究者们主要通过设计手工特征的方式实现单人姿态估计[7-8].然而,基于手工特征方法的准确率和泛化性明显不足,难以应用于实际场景.近年来,随着深度学习技术(Deep Learning,DL)的兴起和大规模数据集的发布,计算机视觉领域迎来了迅速的发展.深度学习被应用在计算机视觉的各个任务当中,包括图像分类[9-10]、目标检测[11-12]、图像分割[13-15]、图像检索[16-17]和图像生成[18-19]等.深度学习方法可以基于任务导向的目标函数从大规模数据中自动学习特征,在多个计算机视觉任务上均取得了优异的性能.研究者们同样将深度学习技术应用在人体姿态估计当中,并在多个公开数据集上持续刷新着最佳性能.图 1 总结了近年来具有代表性的人体姿态估计方法.

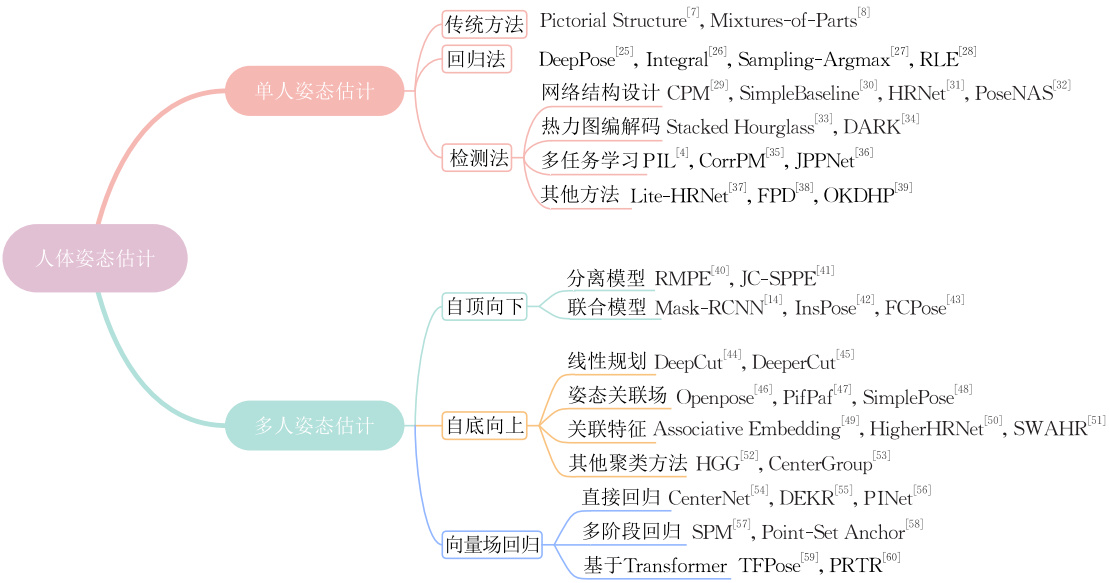


图 1 当前单人/多人姿态估计算法总结

近几年,基于深度学习的二维人体姿态估计研究取得了新进展.然而,学术界却缺乏中文综述性文章对近期进展进行总结和讨论.现有的综述文章可以划分为两类.第一类综述[20-21]重点关注二维人体姿态估计,如其中的单人姿态估计[20].此类综述的特点是方法介绍较为详细,有优缺点分析,但是未能关注最新的研究进展,没有介绍近期流行的单阶段回归方法和基于 Transformer 的方法.另一类综述关注于总结人体动作捕捉技术[22-24].该类综述不仅总结了二维人体姿态估计方法,还涵盖了三维人体

姿态估计以及基于人体模型的方法.由于篇幅限制，该类综述往往只会做简单的方法介绍，没有详细的优缺点分析.此外，现有的综述文章主要停留于方法介绍阶段，没有通过深入分析和对比来指出难点问题.本文聚焦于二维人体姿态估计领域，对近期的深度学习方法做了系统介绍和总结.同时，本文旨在解决现有综述文献不足，通过分析和对比代表性工作指出该任务仍未解决的难点问题，并对未来研究方向进行展望.

基于上述分析，本文对近期的二维人体姿态估计任务工作进行了总结，从每类方法中选取了有代表性的工作进行了详细的分析和对比，并对人体姿态估计领域未来可能的研究方向进行了展望.本文各章节具体组织如下：第 1 节介绍人体姿态估计任务的研究背景和意义；第 2 节介绍人体姿态估计的任务定义，当前面临的难点问题和人体关键点的不同表示方法；第 3 节和第 4 节分别按类别总结单人姿态估计和多人姿态估计的近期工作；第 5 节介绍当前常用的人体姿态估计数据集和人体姿态估计的性能度量标准，对最近方法在多个数据集上的性能进行对比；第 6 节对每一类中的代表性方法在不同场景下的性能进行详细对比，并分析不同方法的优缺点；最后，第 7 节和第 8 节对人体姿态估计领域的未来研究进行展望，并对全文进行总结.

2　任务介绍

本节主要介绍人体姿态估计任务的定义、难点和人体关键点的不同表示方式.

2.1　问题定义

人体姿态估计任务定义人体姿态由一系列关键点坐标组成，对于一个定义了 n 个人体关键点的数据集，其中每个行人的关键点集合 $\mathcal{P}$ 可以表示为

$$\mathcal{P}=\{k_1,k_2,\cdots,k_n\} \tag{1}$$

其中 $k_i=(x_i,y_i,v_i)$ 为第 i 个关键点，(x_i,y_i) 为关键点的二维坐标，v_i 为关键点的可见性，通常情况下 $v_i>0$ 表示关键点可见，$v_i=0$ 表示在图像中不可见.单人姿态估计任务需要预测图像中单个行人的 n 个关键点坐标，多人姿态估计任务除了预测图像中所有行人的关键点坐标外，还需要预测每个关键点的行人归属.人体姿态估计任务的优化目标可以表示为最小化预测关键点坐标和关键点真实坐标之间的距离，

$$\min\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{(\bar{x}_j^i-x_j^i)^2+(\bar{y}_j^i-y_j^i)^2} \tag{2}$$

其中 (x_j^i,y_j^i) 和 $(\bar{x}_j^i-\bar{y}_j^i)$ 分别为第 i 个行人的第 j 个关键点的预测坐标和真实坐标，n 是关键点数量，m 为图像中的行人数量.在单人姿态估计任务中 $m=1$，多人姿态估计任务中 $m\geq 1$.

2.2　任务难点

人体姿态估计面临一系列的难点问题.首先，单人姿态估计中，复杂背景和干扰会增加关键点定位难度.由于人体是非刚体，人体姿态变化会使关键点难以被精准定位.此外，姿态估计任务的应用场景对实时性要求较高，而准确率和速度通常互相矛盾，因此目前的方法难以在准确率和效率之间取得平衡.例如，准确率较高方法往往速度较低，而专注速度提升的轻量化模型难以取得高准确率.

除了以上问题，多人姿态估计还面临一系列新问题.首先，为了区分图像中不同的人，多人姿态估计需要设计额外算法分离不同人.此外，多人场景中往往存在拥挤和遮挡的现象，使不同人的关键点位置接近且互相遮挡.同时，不同行人可能外观相似，难以区分.表 1 总结了每个难点问题对应的代表性方法.

表 1　当前单人/多人姿态估计算法所关注的问题和代表性方法

单人/多人	针对的难点问题	代表性方法
单人	关键点定位精度	Integral[26], Sampling-Argmax[27], RLE[28], HRNet[31], DARK[34], PIL[35]
	模型推理速度	PoseNAS[32], Lite-HRNet[37], FPD[38], OKDHP[39]
多人	区别不同人	Mask-RCNN[14], DeepCut[44], Openpose[46], PifPaf[47], Associative Embedding[49]
	遮挡和密集场景	JC-SPPE[41], PINet[56]
	关键点定位精度	HigherHRNet[50], SWAHR[51], DEKR[55], SPM[57], Point-Set Anchor[58], PRTR[60]
	模型推理速度	InsPose[42], FCPose[43]

2.3　人体姿态表示形式

研究者提出了多种关键点 $\mathcal{P}$ 表示方法.现有方法可以分为三类：二维坐标表示、空间热力图表示以及空间向量场表示.这些关键点的表示形式特点不同，各有优劣.

二维坐标. 人体关键点可以简单的利用二维坐

标表示，即(x,y). 此类方法形式简洁，无需后处理，可以直接提供给后续任务或应用. 然而从输入图像到估计关键点坐标需要建立高度非线性映射，使得模型难以被优化，限制了关键点定位精度.

空间热力图. 为了降低关键点坐标估计难度，同时更好地利用卷积神经网络的空间特征学习能力，研究者提出了利用空间热力图来编码二维关键点坐标，即$(x,y)\rightarrow\mathcal{H}$. 其中空间热力图$\mathcal{H}\in\mathbb{R}^{H\times W}$是一个编码关键点出现概率的概率图，通常采用高斯核在对应关键点坐标位置生成. 该方法可以将复杂的回归问题转化为较为简单的分类问题，对卷积神经网络更加友好. 因此，这种表示形式被广泛应用于关键点定位中，并取得了很高的关键点定位准确率. 然而空间热力图也有其缺点. 首先，它存在量化误差，即定位的精度受到热力图空间大小的限制，为了提高定位准确率需要使用高分辨率的热力图. 其次，相比于二维坐标，热力图占用的存储更多，所需的计算量也更大，处理高分辨率输入数据的效率较低.

空间向量场. 空间向量场表示可以认为是上述两种表示方法的融合. 该方法采用一个向量场$\mathcal{V}\in\mathbb{R}^{H\times W\times C}$来编码关键点的位置信息. 向量场中每个空间位置是一个向量$\boldsymbol{v}=\mathcal{V}(i,j)=(c,\delta x,\delta y)$，其中$c$表示点$(i,j)$的置信度，$(\delta x,\delta y)$表示关键点坐标相对于当前点$(i,j)$的偏置. 基于上述信息，关键点可以通过寻找向量场上置信度c最大的位置(i^*,j^*)，然后加上其对应偏置求得，即$(i^*+\delta x^*,j^*+\delta y^*)$. 向量场表示结合了坐标表示和热力图表示的优点，避免了热力图中的量化误差，同时向量场本身的空间分辨率无需很大，在计算量上比热力图更具优势. 但是向量场表示法形式复杂，需应对回归和分类优化冲突问题，在关键点定位精度上往往稍逊于空间热力图表示法.

3 单人姿态估计

本节主要介绍单人姿态估计的相关工作. 单人姿态估计的假设是每张输入图像I只包含单个行人，因此研究重点是如何准确定位该行人的关键点，即

$$\mathcal{P}=\mathrm{SPPE}(I) \tag{3}$$

其中$\mathrm{SPPE}(\cdot)$为单人姿态估计模型，$\mathcal{P}$为模型预测的n个人体关键点. 单人姿态估计的早期方法大都使用基于模板匹配的传统方法，近年来随着深度学习的兴起，研究者们开始专注深度学习方法，本节将分别介绍这两类方法.

3.1 传统方法

早期研究者使用手工设计特征实现单人姿态估计. Fischler 等人[7]提出了用于单人姿态估计的图形结构(Pictorial Structure). 该方法在空间先验的基础上利用模板匹配思想进行关键点预测. 对于模板关系，提出了著名的弹簧形变模型，合理约束了整体模型和人体部件间的空间相对位置. 图 2(a)展示了如何使用弹簧形变模型对人体关键点进行预测. 在后续研究中，Yang 等人[8]基于弹簧形变模型提出将每个肢体结构切分成更小的区域，使其能够模拟更复杂的姿态变化，从而提高模版匹配效果. 但由于实际场景中行人姿态变化复杂，行人尺度变化大，且容易发生遮挡，导致早期基于手工特征的方法准确率和鲁棒性较差.

3.2 深度学习方法

近期的研究主要关注于深度学习方法. 根据对关键点建模方式的不同，深度学习方法可以进一步分为直接回归法和检测法两大类.

3.2.1 直接回归法

直接回归法将关键点的坐标作为模型输出，使用深度模型建立从图像到坐标的非线性映射. DeepPose[25]是第一个使用深度神经网络直接回归人体关键点的工作. 该方法提出了一个级联模型来回归关键点二维坐标，同时构建了一个大型的单人姿态估计数据集 MPII[61]，为后续的人体姿态估计工作奠定了基础. 为了缓解从图像到坐标高度非线性映射难以优化的问题，研究者们进行了许多尝试与改进. Sun 等人[26]提出了一个基于热力图积分的回归方法 soft-argmax. 该方法首先通过归一化函数将热力图归一化为概率分布，然后使用积分计算关键点坐标，在一定程度上降低了关键点回归难度，提升了回归方法准确率. 之后的研究中，Li 等人[27]发现使用 soft-argmax 会导致学习到多峰的概率分布，降低关键点回归的准确率和置信度，因此提出了基于采样的积分方法 sampling-argmax. 该方法通过使用吉布斯采样来约束关键点概率分布的学习，取得了更好的性能.

此外，直接使用 L1 或 L2 损失函数约束回归模型的学习可以认为是假设在高斯分布或拉普拉斯分布下的极大似然估计. 而这些预先定义的分布往往不能正确反映关键点的真实分布，导致极大似然估计失败. 为了解决这个问题，研究者提出了残差对数似

然估计(RLE)[28]，通过使用归一化流(normalizing flow)来估计关键点的真实分布，以此来调整极大似然估计的先验，在人体姿态估计任务上取得了优异的性能.

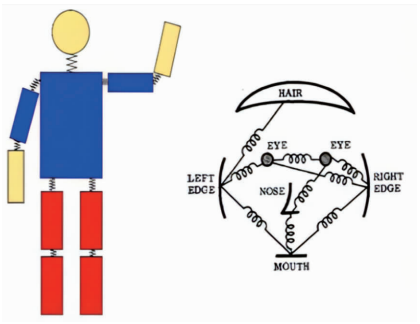
近期有一些工作基于 Transformer 来进行关键点的回归和估计，利用 Transformer 的全局注意力机制来缓解直接回归困难问题. TokenPose[62] 使用关键点标志(keypoint token)表示关键点，使用自注意力机制将 token 和输入图像进行交互来回归对应关键点的坐标. 后续工作中，Ludwig 等人[63] 使用可读的关键点向量来代替 token，通过插值来回归任意的中间关键点，解决了之前方法只能检测固定个数关键点的问题.

3.2.2　检测法

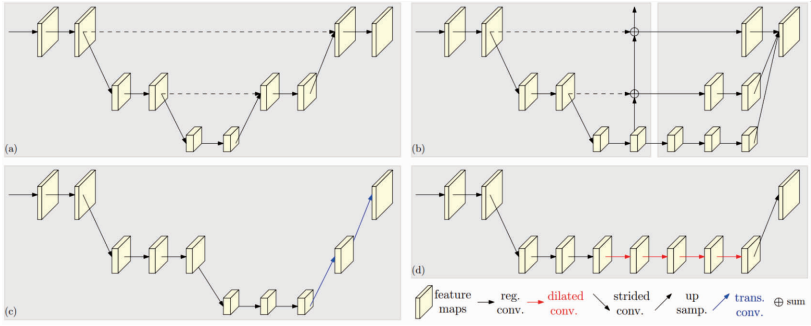
基于检测的方法不直接回归关键点坐标，而是首先在图像中寻找有可能存在关键点的区域，进而通过检测这些区域来解码出关键点的坐标. 其中一个代表性工作是热力图预测[29,64]. 其核心思想是将二维关键点坐标编码为空间概率分布图，即预测每个像素位置存在某个类型关键点的概率，进而选择概率最大的位置作为该类关键点的空间坐标. Wei 等人[29]在 2014 年提出的卷积姿态机(Convolutional Pose Machine，CPM)，是首个将人体姿态估计定义为热力图预测任务的方法. 图 2(b)展示了 CPM 的框架图，CPM 通过全卷积网络预测一个多通道的热力图来表示图像中人体关键点的位置. 由于人体姿态估计需要大范围的上下文信息，因此 CPM 设计了序列结构. 该结构包含多阶段卷积神经网络，每个阶段的感受野逐步增大，使网络输出层可以具有全局感受野，能够更加准确地预测图像中的人体关键点. 此外，该方法还在网络的中间层加入了监督信息，来应对深度网络训练过程中的梯度消失问题.

后续工作进一步研究如何基于热力图提升关键点坐标定位准确率，主要思路包括(1)设计更好的网络结构、(2)引入多任务优化、(3)设计更好的热力图编解码方法等. 本节按照每个方法的主要特点将这些方法分为网络结构设计、多任务学习辅助、热力图编码三类. 由于许多方法同时涉及多个类别，难以进行完全互斥的分类. 同时，一些方法关注模型的测试时微调和推理效率等，难以单独分类，被分为其他方法.

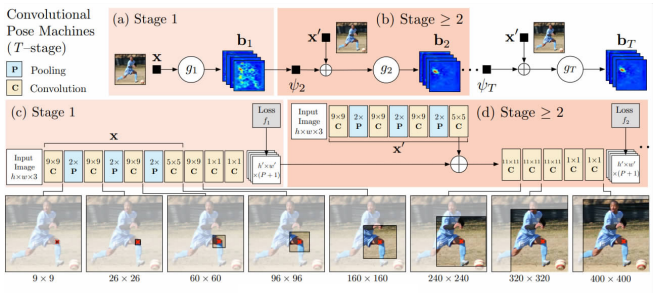
网络结构设计. 该类方法主要关注于设计更好的深度网络实现更加准确的热力图预测，以应对复杂背景和姿态变化等问题. 图 2(c)展示了单人姿态估计中的代表性神经网络结构，包括沙漏网络(Hourglass)[33]、级联网络[65]简单基准网络(SimpleBaseline，SBL)[30]以及高分辨率网络(High Resolution Network，HRNet)[31]等. Newell 等人设计的对称沙漏网络[33]用于恢复较高分辨率的特征，并通过跳跃链接(skip connection)将同分辨率的浅



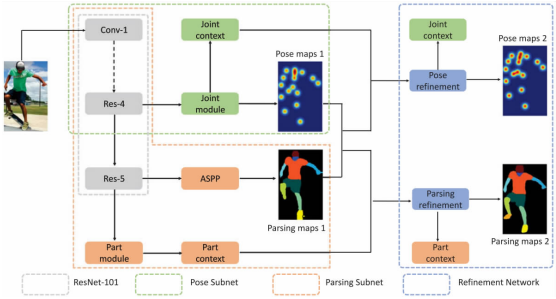
(a) 用于行人关键点检测的弹簧形变模型示意图[7]



(c) 沙漏网络、级联金字塔网络、简单基准网络以及高分辨率网络示意图[31]



(b) 基于热力图的卷积姿态机模型[29]



(d) 多任务联合训练模型[36]

图 2　单人姿态估计方法

层特征与深层特征相连.该网络虽然性能优异,但是参数量巨大,推理效率低.Xiao 等人[30]为了简化网络结构,提出在通用图像分类网络结构后加入三个级联的反卷积层用于恢复语义和高分辨率特征.该结构简单且高效,取得了较好的性能.Chen 等人[65]设计了级联金字塔网络,该网络包含一个全局网络分支和一个精炼网络分支,分别用于关键点预测和预测结构改良,兼顾了人体关键点的局部信息以及全局信息.为了生成更高分辨率的热力图,Sun 等人[31]设计了高分辨率网络(High Resolution Network,HRNet).该网络的核心思想是整个模型维持一个高分辨率的分支,同时利用并联的低分辨率分支来扩大感受野并增强高分辨率分支的语义信息.HRNet 最终使用高分辨率的热力图定位图中的人体关键点.相比于之前的模型,该模型显著降低了低分辨率热力图带来的量化误差,提高了关键点预测准确率.此外,神经网络搜索方法也被用于姿态估计任务中的网络设计[32].

多任务学习辅助.多个任务协同学习可以有效提升特征的表达能力,进而提高每个任务的性能.受此启发,研究者们设计了多任务学习模型[4,35-36,66].Nie 等人[35]提出了一个人体解析(Human parsing)指导的学习器用于提取人体部件信息,然后动态生成姿态模块的参数用于人体姿态估计,取得了更高的姿态估计准确率.Zhang 等人[4]提出将姿态估计、人体解析和边缘检测三个任务统一到一个模型当中,通过学习 3 个任务之间的相关性同时提升多个任务的性能.

热力图编解码.由于检测类方法所输出热力图的分辨率往往小于输入图像的分辨率,此类方法往往只能实现较低分辨率下的关键点定位,无法在亚像素级别下精准定位关键点坐标.提高网络的输出分辨率可以在一定程度上缓解该问题,但模型计算量会显著增加,且不能精确定位连续值的坐标.为了降低量化误差,研究者设计了一系列热力图解码方法来提高关键点定位的准确率.Newell 等人[33]提出了一个经验性的改进方案.他们在热力图上寻找最大值和第二大值的坐标,然后将最大值坐标向次大值坐标偏移 0.25 得到处理后的关键点坐标.Zhang 等人[34]进一步提出了无偏的关键点编解码方法(Distribution-Aware coordinate Representation of Keypoints,DARK),构建了无偏的关键点热力图用于模型训练.DARK 编码缓解了热力图量化误差对关键点编码的影响,使得关键点预测更加精准.

其他方法.考虑到人体关键点之间的关联以及人体结构信息,Tang 等人[64]提出将人体关键点划分为几个组合,然后每一个组合单独学习一个特征用于对应的关键点热力图估计,而不像之前的方法对所有关键点共享一个特征.作者首先在数据集 MPII[61]上使用互信息估计每两个关键点之间的信息熵,进而将一个人的关键点划分为 5 个组,并采用一个多分支的网络来为每个组关键点学习特定特征,从而估计对应的热力图.

Kamel 等人[67]提出了一个增强和纠正混合估计姿态的方法 Hybird-Pose.该方法使用两个独立的沙漏网络进行热力图估计.其中一个网络保持正常的上下采样,用于关键点增强估计.另一个网络使用不同的上下采样策略,来探索不同特征用于关键点纠正.该方法通过两个网络的协同实现了姿态估计准确率的显著提升.

有一些方法通过在测试时微调模型来提升其泛化性能.常用方法大多是在训练集上训练模型,在测试时将模型固定,但难以应对测试数据和训练数据之间的领域鸿沟.为此 Li 等人[68]提出了一个基于 Transformer 的测试微调框架.该框架通过监督学习和自监督学习得到两个姿态估计结果,利用 Transformer 建立联系,并在测试时微调自监督模块来提升模型性能.

此外,研究者们还考虑了模型的推理效率问题.为了提升模型的推理速度并维持原有的准确率,研究者将模型蒸馏[38-39]引入到热力图估计中.通过使用较深的教师模型来指导轻量级学生模型训练,这类方法能有效提升模型推理速度.另外,Yu 等人[37]提出了轻量级的高分辨率网络 Lite-HRNet,在保持高分辨率的同时,使用随机交换模块和条件通道加权模块来进行高效的信息交互,从而提升关键点定位性能.Wang 等人[69]通过收缩实验发现了高分辨率网络中的冗余,提出了一个轻量级的网络 LitePose.该网络包含融合反卷积头,并使用了大尺寸卷积核来提升姿态估计性能,在显著减少参数量的同时维持了较高的姿态估计准确率.

4 多人姿态估计

多人姿态估计需要同时预测人体关键点的位置和归属行人.根据关键点的建模方式,当前基于深度

学习的多人姿态估计方法可以分为两类：基于热力图预测的方法和基于向量场回归的方法.

4.1　基于热力图预测的方法

多人姿态估计需要预测人体关键点位置和其归属. 根据这两个步骤的先后顺序，基于热力图预测的方法可以分为自顶向下和自底向上两类方法. 这两类方法采取不同策略来区分不同人体. 自顶向下方法首先使用行人检测器检测图像中出现的行人，生成行人位置信息如检测框，然后从原图中剪裁出行人区域并应用单人姿态估计算法进行姿态估计. 自底向上方法首先预测出图像中所有行人的关键点，然后通过后处理聚类等方法将关键点组合成不同的行人. 本节将分别介绍两类方法.

4.1.1　自顶向下的方法

自顶向下的方法[14,26,31,33,40,65,70-71]首先使用行人检测器检测图像中出现的所有行人，并利用检测信息得到单个行人信息用于单人姿态估计. 具体来说，给定一张输入图像 I，自顶向下的方法可以表示为

$$\begin{aligned} Loc &= \mathrm{Det}(I), \\ \mathcal{P}_i &= \mathrm{SPPE}(I, Loc_i) \end{aligned} \tag{4}$$

其中 Det 为行人检测器，$Loc=\{Loc_1, Loc_2, \cdots, Loc_n\}$ 是行人检测器预测的 n 个行人位置信息，如检测框或中心点，$\mathcal{P}_i$ 为第 i 个人的姿态估计结果，SPPE 是单人姿态估计模型. 根据是否联合训练行人检测器 Det 和姿态估计模型 SPPE，已有方法可以分为分离模型和联合模型两大类.

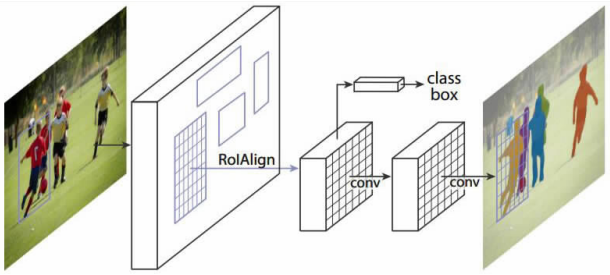
分离模型. 早期方法[72]使用分离的行人检测器和单人姿态估计算法实现多人姿态估计. 常用的行人检测器包括 Faster RCNN[11]、YOLO[73]等. 基于检测出的行人检测框，该类方法使用第 3 节中的单人姿态估计算法进行姿态估计. 行人检测器产生的误差会使行人身体被截断，导致身体部件丢失. 为了应对行人检测器的检测误差，Fang 等人[40]提出了对称空间变换网络(Symmetric Spatial Transformer Network，SSTN)，通过预测仿射变换参数对检测框进行修正. 此外，为了消除拥挤场景下检测框内的冗余关键点，Fang 等人[40]还设计了参数化的非极大值抑制方法，通过参数化的人体关键点距离去除冗余预测. Li 等人[41]针对拥挤场景下行人重叠问题，设计了候选关键点损失函数(Joint Candidates loss)，对包含多个人的检测框预测多峰热力图，防止关键点漏检，然后建立行人-关节图对关键点进行匹配，消除冗余的关键点. 该方法的缺点是需要后处理的匹配操作，无法进行端到端优化.

与单人姿态估计相比，多人姿态估计的场景更加复杂，不同行人之间往往存在遮挡，降低了行人检测准确率. 此外，分离模型中的行人检测器和单人姿态估计模型需要单独训练，不能端到端地联合优化，限制了分离模型的性能和推理效率. 针对以上问题，研究者们设计了自顶向下的联合模型用于多人姿态估计.

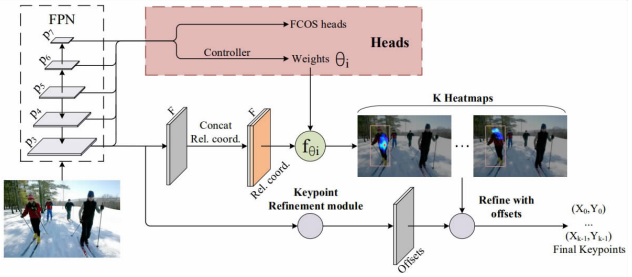
联合模型. 联合模型使用单一模型同时实现行人检测和姿态估计，通过共享特征来联合优化检测和姿态估计任务，从而可以使模型能够学习更好的特征. 此外共享特征能够减少模型的计算量，提升模型的效率. He 等人[14]提出了 Mask-RCNN，在行人检测器 Faster R-CNN[11]的基础上添加了用于单人姿态估计的网络分支，可以对行人检测任务和单人姿态估计任务进行联合优化. 图 3(a)展示了 Mask-RCNN 模型的示意图. Mask-RCNN 的优势是降低了分离模型的参数量，使多人姿态估计能实时进行. 然而该方法依然使用检测框来分离不同的人，因此也难以处理前文提到的遮挡等问题.

上述方法需要感兴趣区域池化(ROI-Pooling)或者感兴趣区域对齐(ROI-Align)操作从特征图中剪裁出单人特征图用于单人姿态估计. 但是裁剪操作容易丢失图像中的上下文信息. 此外，ROI-Pooling 和 ROI-Align 操作的引入会增加模型的部署难度. 因此有研究者提出无检测框的自顶向下方法. Shi 等人[42]设计了全卷积的自顶向下姿态估计方法 InsPose. 该方法通过全卷积架构为图像中的每个行人生成 3 个动态卷积核. 动态卷积核作用于全局特征图生成对应行人的关键点热力图用于姿态估计. 相似地，Mao 等人[43]提出了基于全卷积架构的 FCPose. 在动态卷积的基础上，FCPose 在每个行人的热力图计算过程中并联了额外的相对坐标为模型提供空间先验.

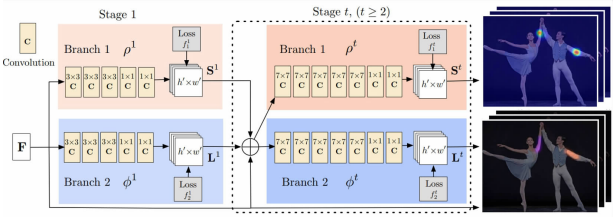
自顶向下的方法在多个姿态估计数据集上取得了较高的性能. 然而，自顶向下的方法需要额外的行人检测器，并且大多数方法会对检测到的行人进行裁剪、缩放等操作，然后对每个行人进行单人姿态估计. 此过程显著增加了自顶向下模型的计算量和推理时间，使得自顶向下方法效率低下. 此外，在拥挤场景下难以准确检测行人，进一步影响了后续的单人姿态估计准确性.



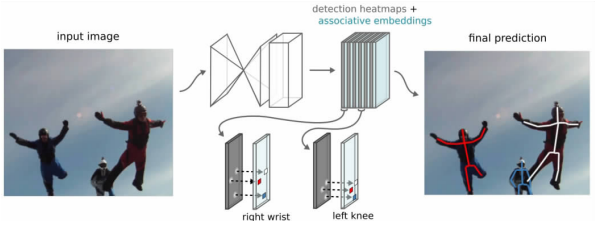
(a) 自顶向下的多人姿态估计算法Mask-RCNN示意图[14]



(b) 无外接框自顶向下的算法FCPose示意图[43]



(c) 自底向上的姿态估计算法Openopse示意图[46]



(d) 自底向上的姿态估计算法AE示意图[49]

图 3 基于热力图的多人姿态估计方法

4.1.2 自底向上的方法

与自顶向下方法的流程相反，自底向上的方法[46,49-50,53]首先定位图像中所有的人体关键点，然后通过聚类将关键点组合成不同行人. 具体来说，给定一张输入图像 I，自底向上的方法可以表示为

$$K=\mathrm{MPPE}(I),$$
$$\mathcal{P}=\mathrm{Group}(K) \quad (5)$$

其中 MPPE 为关键点检测算法，用于检测图中出现的所有人体关键点，Group 为关键点聚类算法，用于将所有检测到的关键点分组为不同行人. 自底向上方法的核心是设计更加准确高效的聚类算法对关键点进行聚类. 自底向上的方法不需要额外的行人检测和单人姿态估计，可以在一定程度上将模型的推理时间与图片中的行人数量解耦，提升推理速度.

整数线性规划. Pishchulin 等人提出的 DeepCut[44]是最早的自底向上的方法之一. DeepCut 首先使用改进的 Fast R-CNN[74]检测图像中行人的身体部件，然后使用整数线性规划(Integer Linear Program，ILP)算法对关键点进行匹配. 基于DeepCut，Insafutdinov等人设计了DeeperCut[45]. DeeperCut 使用了更深的 ResNet[10]作为人体部件检测器，并设计了额外的图像条件化的匹配项(Image-Conditioned Pairwise Terms)，通过引入额外的空间信息降低候选部件的数量，提升匹配效率. 然而，整数线性规划算法计算复杂，且不能和关键点检测模型联合优化. 因此，后续方法设计了可学习的关键点匹配策略，包括人体部件关联场(Part Affinity Fields，PAFs)和关联特征.

部件关联场. 基于部件关联场的方法将关键点之间的关联定义为身体部件，基于每对关键点之间身体部件的置信度对关键点进行聚类. Cao 等人在 2016 年提出了自底向上的多人姿态估计算法 Openpose[46]. Openpose 包含两路网络分支分别预测行人关键点和部件关联场. Openopse 通过在姿态关联场中计算每一对关键点连线路径上的积分得到该对关键点的相关系数用于聚类. Kreiss 等人[47]提出了 PifPaf，同时预测了部件强度场(Part Intensity Field，PIF)和部件关联场(Part Association Field，PAF). 其中 PIF 用于在低分辨率下精确定位人体关键点，PAF 用于关键点聚类. 基于 PAF，Hidalgo 等人[75]提出了首个全身姿态估计方法，通过自底向上的方式同时预测行人的身体、脸部、手部和脚的关键点. Li 等人[48]提出新的身体部件编码方式用于编码关键点之间的联系，并设计了恒等映射的沙漏网络用于热力图的预测，在模型训练过程中加入了多尺度的中间监督信号. 部件关联场能快速确定关键点之间的连接信息，因此具有较快的组装速度. 例如，Openpose 是典型的实时多人姿态估计方法. 然而这一类方法需要设计复杂的匹配算法进行关键点组装，同时也容易受到遮挡等问题的影响.

关联特征. 基于关联特征的方法对每个关键点预测一个额外的向量或标量特征用于关键点分组. Newell 等人[49]在 2017 年提出了基于关联特征(Associative Embedding，AE)的双路模型，用于同

时预测关键点的热力图和关联特征. 图 3(d)展示了 AE 算法的具体流程. AE 算法首先从关键点分支选取高响应值的关键点，并将关联特征分支中关键点对应位置的关联特征用于关键点聚类. 相比于部件关联场方法，关联特征的计算更加简洁高效. 行人尺度变化会对姿态估计产生影响. Cheng 等人[50]设计了能够适应行人尺度变化的超高分辨率网络 HigherHRNet. HigherHRNet 包含多个不同分辨率的网络分支，网络的每个阶段对不同分辨率分支的特征进行交互，并全程保持了高分辨的特征图，最终将多个分支特征融合用于关键点热力图预测. HigherHRNet 能够更好地适应行人的尺度变化，可以通过输出的高分辨率热力图降低量化误差，取得了优异的性能. 由于图像中每个人的关键点的尺度不同，且人工标注存在歧义，使用固定的均值和方差对关键点进行编码会影响模型的训练. Luo 等人[51]针对这一问题提出了尺度权重自适应热力图回归(Scale Weight-Adaptive Heatmap Regression，SWAHR). 该方法在训练过程中自适应地调整关键点编码的高斯核标准差和训练权重，不需要额外的尺度标注，能够有效减小尺度变化和标注误差的影响.

其他聚类方法. Jin 等人[52]提出了用于关键点聚类的层次图聚类(Hierarchical Graph Grouping，HGG)算法. HGG 将人体关键点作为结点构建图，并训练边判别器(Edge Discriminator)判定 2 个结点是否来自同一个人. HGG 通过结点合并和图修建操作将所有结点组合成不同行人. 相比较于 AE，HGG 的聚类通过网络前传完成，因此具有较高的推理效率. Brasó 等人[53]提出了基于 Transformer 的特征聚类方法 CenterGroup. CenterGroup 使用 Transformer 对人体中心点和关键点位置的特征向量进行交互，然后基于交互后的中心点特征和关键点特征之间的相似度进行聚类. CenterGroup 的聚类过程只需要计算中心点特征和关键点特征之间的相似度，因此具有更高的效率.

4.2　基于向量场回归的方法

基于向量场回归的方法[54,56-58]将行人的关键点表示成人体中心点和人体关键点相对于中心点之间的偏移，进而设计单阶段的模型以同时预测人体中心点和回归偏移值. 该类方法中人体关键点可以表示为

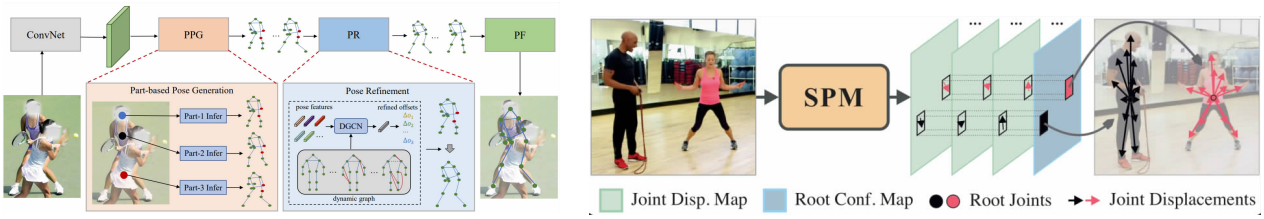
$$(x_i,y_i)=(x_c,y_c)+(\delta x_i,\delta y_i) \tag{6}$$

其中(x_i,y_i)是行人第 i 个关键点坐标，(x_c,y_c)为行人中心点的坐标，$(\delta x_i,\delta y_i)$是行人第 i 个关键点相对于行人中心点(x_c,y_c)的偏移值.

直接回归方法. Zhou 等人[54]提出了直接回归关键点偏移的 CenterNet 模型. CenterNet 模型包含三个分支，分别用于预测图像中行人的中心点、关键点相对于中心点的偏移以及关键点的热力图. 在模型推理阶段，CenterNet 将回归结果匹配到热力图上最近的关键点预测作为最终结果. CenterNet 仅基于人体中心点位置的特征回归关键点，难以关注到不同关键点附近丰富的上下文信息. 因此 Geng 等人[55]提出了解耦关键点回归的方法(DisEntangled Keypoint Regression，DEKR). DEKR 对每个关键点使用单独的分支进行回归，并加入了变型卷积[76]使每个分支更好地关注对应关键点附近的上下文信息. 在拥挤场景下，人体中心点可能被遮挡，导致人体中心点无法准确检测出. Wang 等人[56]提出了基于直接姿态推理的多人姿态估计模型(Pose-level Inference Network，PINet). 图 4(a)展示了 PINet 整体框检图. PINet 基于多个部件中心点对关键点进行回归，并将多个中心点的回归结果组合成最终的姿态估计结果. 该方法有效解决了密集场景下人的区分和关键点漏检问题.

多阶段回归方法. 图像中行人的尺度变化增大了人体关键点偏移的变化范围，增加了回归难度，使基于向量场回归的方法难以准确回归关键点的坐标. Nie 等人[57]提出了单阶段多人姿态机(Single-stage multi-person Pose Machines，SPM)，将关键点偏移的回归任务分解为多层次短距离回归. 图 4(b)，展示了 SPM 算法的整体框架图. SPM 将人体关键点划分为多个层次，从人体中心点开始逐层次回归，从而降低了每次回归的长度. Wei 等人[58]提出了设置关键点锚点(Point-Set Anchor，PSA)的方法降低回归难度. PSA 将数据集中关键点聚类生成多组锚点，对每个检测到的行人中心点，从最近的锚点回归关键点偏移. 此外，PSA 还通过多阶段精炼模块对回归的结果进一步精炼，使得回归的结果更加准确.

基于 Transformer 的回归方法. 最近的一些工作将特征变换器(Transformer)引入姿态估计任务用于直接回归关键点坐标. Li 等人[60]设计了端到端的多人姿态估计 Transformer，同时进行行人检测和关键点回归. 相比于基于热力图的姿态估计方法，基于回归的方法可以同时预测行人的中心点和关键点偏移. Shi 等人[77]提出了一个多层次 Transformer



(a) 直接回归方法PINet示意图[56] (b) 层次回归方法SPM示意图[57]

图 4 基于向量场回归的多人姿态估计方法

模型,包括一个姿态解码器来得到粗粒度的人体姿态以及一个关键点解码器来对粗粒度姿态进行增强与优化. 得益于 Transfomer 大模型的强大表征能力,该模型在多个姿态估计数据集上取得了优异的性能.

4.3 数据集

目前广泛使用的人体姿态估计数据集有 6 个,包括单人数据集 LSP[78]、LSP-Extended[79]、MPII-Single[61]和多人数据集 MPII-Multiple[61]、COCO[80]以及包括拥挤场景的 OCHuman[81]和 CrowdPose[41]. 表 2 按照发布时间汇总了当前人体姿态估计数据集的信息. 其中 4 个多人姿态估计数据集的详细信息如下:

(1) MPII-Multiple[61]数据集一共包含 5500 张图片,共标注了 14 184 个行人,每个行人被标注了 16 个关键点,包括头顶、脖子、肩膀、手肘、手腕、胸腔、骨盆、臀部、膝盖和脚踝等. 每个关键点标注了空间坐标和可见性. MPII-Multiple 数据集的图片被划分为训练集和测试集,分别包含 3800 和 1700 张图片. MPII-Multiple 数据集的测试基准为 PCKh.

(2) COCO[80]数据集一共包含约 82 000 张图片,共标注了 230 000 个行人,是当前最大的多人姿态估计数据集. 每个行人被标注了 17 个关键点,包括鼻子、眼睛、耳朵,肩膀、手肘、手腕、臀部、膝盖和双脚等. 每个关键点标注了空间坐标和可见性. COCO 数据集的图片被划分为训练、验证和测试 3 个子集,分别包含约 57 000 张,5000 张和 20 000 张图片. COCO 数据集的测试基准为基于关键点相似度 OKS(Object Keypoint Similarity)的平均准确率 AP(Average Precision).

(3) OCHuman[81]数据集是最近新提出的多人姿态估计数据集,包含更具挑战性的场景. 行人检测框的平均交并比(IoU)为 67%. OCHuman 数据集和 COCO 一样标注了 17 个人体关键点. 与 COCO[80]和 CrowdPose[41]数据集不同,OCHuman 数据集不包含训练集,仅提供了验证集和测试集. 该数据集一共包含 4731 张图像,其中 2500 张为验证集,2231 张为测试集. 该数据集的测试基准为 AP.

(4) CrowdPose[41]数据集是针对更具有挑战性的拥挤场景下的多人姿态估计任务提出的. 该数据集包含 20 000 张图片,一共标注了约 80 000 个行人. 每个行人标注了 14 个关键点,包括头顶、脖子,肩膀、手肘、手腕、臀部、膝盖和双脚等. 每个关键点标注了坐标和可见性. CrowdPose 数据集被划分为训练、验证和测试 3 个子集,分别包含 10 000 张、2000 张和 8000 张图片. 该数据集的测试基准同样为 AP.

表 2 人体姿态估计数据集信息对比

数据集	发布年份	类型	图片数/k			行人数量	关键点数量	标注信息	测试标准
			训练	验证	测试				
LSP[78]	2010	单人	1.0	—	1.0	2	14	坐标+可见性	*PCP*
LSP-Extended[79]	2010	单人	10.0	—	—	10	14	坐标+可见性	*PCP*
MPII-Single[61]	2014	单人	29.0	—	12.0	41	16	坐标+可见性	*PCKh*
MPII-Multiple[61]	2014	多人	3.8	—	1.7	14	16	坐标+可见性	*AP*
COCO[80]	2016	多人	57.0	5.0	20.0	230	17	坐标+可见性	*AP*
OCHuman[81]	2019	多人	—	2.5	2.2	8	17	坐标+可见性	*AP*
CrowdPose[41]	2019	多人	10.0	2.0	8.0	80	14	坐标+可见性	*AP*

图 5 展示了 4 个多人姿态估计数据集中的图片样本. 从图中可以看到,OCHuman[81]和 CrowdPose[41]数据集中的图片包含了更加拥挤的场景,行人间的遮挡严重,给姿态估计带来了更大的挑战.

4.4 测试基准

基于人体关键点坐标 $\mathcal{P}$,研究者们定义了多种性能度量方式,本节将详细介绍 3 种常用的人体姿态估计的性能度量.

图 5　多人姿态估计数据集图片样例：MPII[61]、COCO[80]、OCHuman[81]和 CrowdPose[41]

5　数据集和性能度量标准

5.1　身体部件正确百分比

早期的研究工作主要使用身体部件正确百分比 *PCP*(Percentage of Correct Parts)衡量单人姿态估计算法的准确率. *PCP* 以肢体长度的一半作为参考值，计算关键点预测位置和真实位置之间距离小于该参考值的关键点的比例，主要用于衡量人体四肢关键点的预测准确率，通常表示为 *PCP*@0.5. 对于每张包含单个行人的图片，其 *PCP* 的计算过程为

$$PCP@0.5=\frac{\sum_{i=1}^{n}\delta\left(\frac{d_i}{limb_i}\leq 0.5\right)\delta(v_i>0)}{\sum_{i=1}^{n}\delta(v_i>0)} \tag{7}$$

其中 d_i 为第 i 个关键点的预测坐标和真实坐标的距离，通常为欧式距离，$limb_k$ 为对应的肢体长度. 对于每个行人，*PCP* 通常只计算其可见的关键点. *PCP* 的数值越高表示性能越好. 由于不同视角下看到的人体肢体长度会发生变化，导致 *PCP* 度量稳定性较差.

5.2　关键点正确百分比

在 MPII[61] 数据集上 研究者提出了关键点正确百分比 *PCK*(Percentage of Correct Keypoints)用以衡量姿态估计算法的准确性. *PCK* 同样将预测位置和真实位置距离小于一定阈值的关键点视为正确预测，并计算正确预测关键点的比例. *PCK* 通常选用躯干长度作为阈值，例如 *PCK*@0.2 表示阈值为 0.2 倍躯干长度. *PCK* 的计算过程为

$$PCK@\tau=\frac{\sum_{i=1}^{n}\delta\left(\frac{d_i}{T}\leq\tau\right)\delta(v_i>0)}{\sum_{i=1}^{n}\delta(v_i>0)} \tag{8}$$

其中 T 是该行人的躯干长度，τ 为阈值，*PCK* 同样只计算可见关键点. *PCKh*@0.5 为 *PCK* 的改进版本，其阈值选取为行人头部长度的 0.5 倍. 通过对 *PCK* 设置不同的阈值，可以得到包围曲线 *AUC*(Area Under the Curve)用以衡量算法的性能.

5.3　平均准确率

为了统一目标检测、实例分割和姿态估计三个任务的性能度量，研究者们在 COCO 数据集上提出了基于目标关键点相似度 *OKS*(Object Keypoint Similarity)的平均准确率 *AP* 和平均召回率 *AR*(Average Recall). 对于关键点 P，其 *OKS* 定义为

$$OKS(P)=\frac{\sum_{i=1}^{n}\exp(d_i^2/2S\alpha_i^2)\delta(v_i>0)}{\sum_{i=1}^{n}\delta(v_i>0)} \tag{9}$$

其中 S 为行人在图像中的面积，α_i 是每个关键点的归一化系数.

基于 *OKS*，一些工作定义了不同阈值下的准确率和召回率，对于数据集中第 i 张图像，其准确率 *Prec* 和召回率 *Recall* 为

$$Prec_\tau(i)=\frac{TP}{TP+FP},$$
$$Recall_\tau(i)=\frac{TP}{TP+FN} \tag{10}$$

其中 τ 为 *OKS* 的阈值，*TP*、*FP* 和 *FN* 的含义具体如下：

真实的正样本 *TP*(True Positive)：表示预测关键点中与标注关键点之间 *OKS* 大于或等于阈值 τ 的数量. 该数值的取值范围为[0，N]，其中 N 为图片中行人的数量.

错误的正样本 *FP*(False Positive)：表示预测关键点中与标注关键点之间 *OKS* 小于阈值 τ 的数量. 理论上由于预测的关键点的数量可以有无限个，所以该数值的取值范围为[0，$+\infty$].

错误的负样本 *FN*(False Negative)：表示标注关键点中与预测关键点之间的 *OKS* 小于阈值 τ 的

数量.该数值的取值范围为$[0,N]$.

整个数据集在阈值τ下的平均准确率AP_τ和平均召回率AR_τ定义为

$$AP_\tau=\frac{1}{M}\sum_{i=1}^{M}Prec_\tau(i),$$
$$AR_\tau=\frac{1}{M}\sum_{i=1}^{M}Recall_\tau(i) \tag{11}$$

其中M为数据集中图片的数量.

在COCO数据集上的评价指标AP和AR定义为不同OKS阈值下的平均准确率和平均召回率的均值,

$$AP=\frac{1}{10}\sum_{i=0}^{10}AP_{0.5+i\times0.05},$$
$$AR=\frac{1}{10}\sum_{i=0}^{10}AR_{0.5+i\times0.05} \tag{12}$$

即AP为阈值τ从0.5开始到0.95一共10个阈值下平均准确率的均值.同样的,AR为10个阈值下平均召回率的均值.

此外,研究者在COCO数据集上还定义了不同尺度下的平均准确率AP_M和AP_L,分别表示中等尺度的行人和大尺度行人的姿态估计平均准确率.其中中等尺度的行人定义为该行人的检测框面积$area\in[32^2,96^2]$,大尺度行人的检测框面积为$area\in(96^2,+\infty)$.在CrowdPose数据集上研究者还定义了不同场景下的平均准确率AP_E、AP_M和AP_H,分别用于衡量算法在简单场景、中等难度场景和拥挤场景下的平均准确率.

为了综合对比不同姿态估计算法的性能,本节总结了最近所提出的算法在3个大规模数据集上的性能,包括MPII、COCO和CrowdPose数据集,并对结果进行分析.

5.4 性能对比和分析

5.4.1 MPII

表3总结了最近所提出的方法在MPII数据集上的性能.MPII数据集中的行人遮挡较少,姿态估计相对简单.因此,现有基于深度学习的方法在该数据集上的准确率较高,早期深度学习方法如DeepPose[25]和IEF[84]实现了近80%的$PCKh@0.5$准确率.从表中还可以看出,2016年CPM[29]的提出显著提升了MPII数据集上的性能.当前性能最高的方法为BPN[64],实现了92.7%的$PCKh@0.5$准确率.

表3 MPII数据集上的性能对比

方法名称	发表时间	主干网络	输入尺寸	PCKh0.5	关键词
基于回归的方法					
Deeppose[25]	CVPR2014	AlexNet	220×220	79.6	直接回归,多阶段精炼
Integral[26]	ECCV2018	PointNet++	256×256	91.0	可微积分回归
RLE[28]	ICCV2021	ResNet50	256×256	**85.5**	残差对数似然估计
基于检测的方法					
HybridConvNet[82]	NeuIPS2014	AlexNet	220×220	79.6	马尔可夫随机场,多尺度输入
CascadeConvNets[83]	CVPR2015	AlexNet	256×256	82.0	空间随机失活,由粗到细级联结构
IEF[84]	CVPR2016	GoogLeNet	224×224	81.3	迭代错误反馈
RGs[85]	CVPR2016	VGG	224×224	82.4	层次矫正高斯模型,二次规划
CPM[29]	CVPR2016	自己设计	368×368	90.9	热力图预测,中间层监督
StackedHourglass[33]	ECCV2016	Hourglass	256×256	90.9	沙漏模型,中间层监督
PoseNet[86]	ICCV2017	Encoder/Decoder	256×256	91.9	人体结构信息,对抗学习
PRM[87]	ICCV2017	Hourglass	256×256	92.0	多尺度特征金字塔
MSS-Net[88]	ECCV2018	Hourglass	256×256	92.1	多尺度监督,多尺度回归网络
DLCM[89]	ECCV2018	Hourglass	256×256	92.3	层次组合模型,骨架分割
HRNet[31]	CVPR2019	HRNet-W32	256×256	92.3	高分辨率,多尺度融合
BPN[64]	CVPR2019	Hourglass	256×256	**92.7**	数据驱动的关键点分组,身体部件分支

5.4.2 COCO

表4总结了最近的方法在COCO test-dev数据集上的性能.从表格中可以看出,自顶向下分离模型方法在COCO数据集上实现了最好的性能.可能的原因包括(1)COCO数据集中图像所含的拥挤场景较少,行人检测器能够准确检测出没有被遮挡的行人;(2)分离模型方法在剪裁放大后的单人图像上进行姿态估计.基于高分辨图像进行姿态估计能够降低热力图预测的量化误差,使结果更加准确.联合模型(自顶向下、自底向上和向量场回归)的输入为包含多人的原始图像.用于定位关键点的热力图分辨率相对较低,因此在准确率上不具有优势.然而,联合模型不需要额外的行人检测器,且可以将运行效率和图像中行人的数量解耦,因此效率更高.从表格中还可以看出,最近的方法大多使用HRNet[31]作为主干网络,并实现了优异的性能.这说明了高分辨率的热力图更加有助于关键点的定位.图6展示了典型的自顶向下方法在COCO数据集上的可视化结果对比.

表 4　COCO 测试集上单尺度测试的性能对比

方法名称	发表时间	主干网络	输入尺寸	AP	AP_{50}	AP_{75}	AP_M	AP_L	关键词
自顶向下									
分离模型									
G-RMI[72]	CVPR2017	ResNet101	353×257	64.9	85.5	71.3	62.3	70.0	亚像素回归，关键点相似度 NMS
RMPE[40]	ICCV2017	Hourglass	320×256	61.8	83.7	69.8	58.6	67.6	对称空间变换网络，参数化 NMS
SimpleBaseline[30]	ECCV2018	ResNet152	384×288	73.7	91.9	81.8	70.3	80.0	基线模型，反卷积
Integral[26]	ECCV2018	ResNet101	256×256	67.8	88.2	74.8	63.9	74.0	可微积分回归
CPN[65]	CVPR2018	ResNet-Inception	384×288	72.1	91.4	80.0	68.7	77.2	级联金字塔网络，在线难例挖掘
JC-SPPE[41]	CVPR2019	ResNet-101	320×256	70.9	—	—	—	—	候选关键点
HRNet-W32[31]	CVPR2019	HRNet-W32	384×288	74.9	92.5	82.8	71.3	80.9	高分辨率，多尺度特征融合
HRNet-W48[31]	CVPR2019	HRNet-W48	384×288	75.5	92.5	83.3	71.9	81.5	高分辨率，多尺度特征融合
DARK[34]	CVPR2020	HRNet-W48	384×288	76.2	92.5	83.6	72.5	82.4	无偏热力图编解码
RSN[90]	ECCV2020	RSN50	384×288	78.6	94.3	86.6	75.5	83.3	残差阶梯网络，层内融合
RLE[28]	CVPR2021	HRNet-W48	256×256	75.7	92.3	82.9	72.3	81.3	残差似然估计
TokenPose[62]	ICCV2021	HRNet-W48	256×192	**75.8**	**90.3**	**82.5**	**72.3**	**82.7**	自注意力，关联约束
联合模型									
Mask-RCNN[14]	ICCV2017	ResNet50+FPN	800	63.1	87.3	68.7	57.8	71.4	多任务学习，联合模型
FCPose[43]	CVPR2021	ResNet101	800	65.6	87.9	72.6	62.1	72.3	全卷积架构，端到端
InsPose[42]	ACM MM2021	HRNet-W32	800	69.3	90.3	76.0	64.8	76.1	动态卷积，端到端
自底向上									
Openpose *[46]	CVPR2017	自己设计	368	61.8	84.9	67.5	57.1	68.2	部件关联场，多阶段模型
AE *[49]	NeurIP2017	Hourglass	512	62.8	84.6	69.2	57.5	70.6	关联特征，聚类后处理
PersonLab[91]	ECCV2018	ResNet152	1401	66.5	88.0	72.6	62.4	72.3	多尺度偏移预测，偏移聚类
MultiPoseNet[92]	ECCV2018	ResNet101	480	69.6	86.3	76.6	65.0	76.3	多任务学习，姿态残差网络
PifPaf[47]	CVPR2019	ResNet152	401	66.7	—	—	62.4	72.3	部件强度场，部件关联场
HGG[52]	ECCV2020	Hourglass	800	60.4	83.0	66.2	—	—	层次图聚类
HigherHRNet[50]	CVPR2020	HRNet-W48	640	68.4	88.2	75.1	64.4	74.2	超高分辨率，多尺度融合
CenterGroup[53]	ICCV2021	HRNet-W48	512	69.6	89.7	76.0	64.9	76.3	基于注意力的聚类
SWAHR[51]	CVPR2021	HRNet-W48	640	**72.0**	**90.7**	**78.8**	**67.8**	**77.7**	自适应尺度的热力图回归
向量场回归									
CenterNet[54]	Arxiv2019	Hourglass104	512	63.0	86.8	69.6	58.9	70.4	基于中心点的直接回归
SPM *[57]	ICCV2019	Hourglass	384	66.9	88.5	72.9	62.6	73.1	基于中心点的层次回归
End2end PRTR[60]	CVPR2021	HRNet-W48	—	64.9	87.0	71.7	60.2	72.5	级联 Transformer
Point-Set Anchor[58]	ECCV2020	HRNet-W48	800	66.3	87.7	73.4	64.9	70.0	关键点锚点，多阶段回归
PINet[56]	NeurIP2021	HRNet-W32	512	66.7	88.0	74.0	61.9	74.8	部件姿态回归，拥挤场景
DEKR[55]	CVPR2021	HRNet-W48	640	**70.0**	**89.4**	**77.3**	**65.7**	**76.9**	关键点独立分支，自适应卷积

注：分离模型的输入尺寸为单人姿态估计模型输入图像尺寸，其他方法为原始视频帧短边长度.

* 表示使用额外单人姿态估计模型改善预测结果.

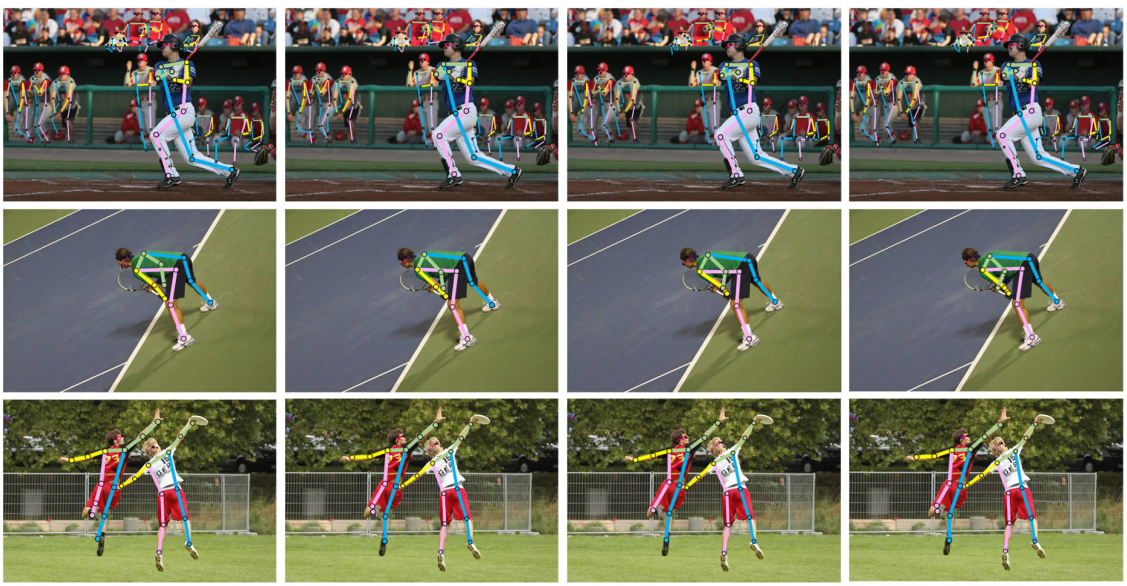


图 6　不同方法在 COCO[80] 数据集上的可视化对比（从左到右依次是：直接回归、RLE[28]、SimpleBaseline[30] 以及 HRNet[31]）

5.4.3 CrowdPose

CrowdPose 数据集上近期方法的性能总结如表 5 所示. 由于 CrowdPose 数据集是最近发布的数据集,因此在该数据集上汇报性能的方法相对较少. 表 5 总结了截至目前为止,所有在该数据集上汇报了性能的方法. 从表 5 中可以看到,这些方法在 CrowdPose 上的性能低于在 COCO 上的性能. 这显示出 CrowdPose 数据集更具挑战性. 从表格中还可以看出,针对遮挡场景设计的 PINet[56] 虽然在 COCO 数据集上性能明显低于 DEKR[55](*AP* 低了 3.3%),但是在 CrowdPose 数据集上的性能实现了反超.

表 5 CrowdPose 数据集上的性能对比

方法名称	发表时间	主干网络	输入尺寸	*AP*	AP_{50}	AP_{75}	AP_E	AP_M	AP_H
自顶向下									
Mask-RCNN[14]	ICCV2017	ResNet50+FPN	800	57.2	83.5	60.3	69.4	57.9	45.8
RMPE[40]	ICCV2017	Hourglass	320×256	61.0	81.3	66.0	71.2	61.4	51.1
SimpleBaseline[30]	ECCV2018	ResNet152	384×288	60.8	84.2	71.5	71.4	61.2	51.2
JC-SPPE[41]	CVPR2019	ResNet-101	320×256	66.0	84.2	71.5	75.5	66.3	57.4
自底向上									
Openpose[46]	CVPR2017	自己设计	368	—	—	—	62.7	48.7	32.3
HigherHRNet[50]	CVPR2020	HRNet-W48	640	65.9	86.4	70.6	73.3	66.5	57.9
SWAHR[51]	CVPR2021	HRNet-W48	640	**71.6**	**88.5**	**77.6**	**78.9**	**72.4**	**63.0**
向量场回归									
DEKR[55]	CVPR2021	HRNet-W48	640	67.3	86.4	72.2	74.6	68.1	58.7
PINet[56]	NeurIPS2021	HRNet-W32	512	68.9	88.7	74.7	75.4	69.6	61.5

注:分离模型的输入尺寸为单人姿态估计模型输入图像尺寸,其他方法为原始帧短边尺寸.

6 代表性方法分析与讨论

本节选取了多人姿态估计的代表性方法,包括自顶向下方法中的 HRNet、Mask-RCNN,自底向上方法中的 HigherHRNet、CenterGroup 以及向量场回归中的 DEKR 和 PINet. 我们首先对比了这些方法的准确率和效率,通过误差分析讨论它们在不同关键点定位上的表现,最后在密集场景下测试它们对于遮挡的鲁棒性.

6.1 性能与误差分析

我们在常用的 COCO 验证集上测试以上方法的性能并分析其误差,结果如表 6 与图 7 所示.

表 6 COCO 验证集上的性能对比

方法名称	FPS	*AP*	AP_{50}	AP_{75}	*AR*	AR_{50}	AR_{75}
自顶向下							
HRNet[31]	2.0	**76.4**	**93.6**	**83.7**	**79.3**	**94.4**	**85.7**
Mask-RCNN[14]	11.2	66.4	87.9	72.4	73.1	91.9	78.4
自底向上							
HigherHRNet[50]	2.5	67.2	86.3	73.1	71.8	88.5	76.8
CenterGroup[53]	5.6	69.1	87.7	74.4	73.4	90.7	78.1
向量场回归							
DEKR[55]	**15.8**	68.1	86.8	74.5	73.0	89.8	78.4
PINet[56]	13.4	67.0	86.6	74.0	73.1	90.5	79.0

每个方法在多人检测指标 *AP* 和 *AR* 上的表现如表 6 所示. 除了使用分离模型的 HRNet,其他方法的 *AP* 和 *AR* 性能指标较为相似,集中在 65%~68%之间. 这说明不同的人体区分方法在普通场景下取得了相似性能. 使用分离模型的 HRNet 取得的准确率显著高于联合模型. 一方面是由于分离模型中的网络只需要进行关键点定位这一任务,相比于联合模型中同时进行人体检测和关键点定位的任务会更加简单也更加容易优化. 另一方面,分离模型会将裁剪的单人区域缩放到统一大小,作为模型输入. 因此这类方法能有效处理人的尺度变化问题,对大尺寸和小尺寸的人体都有良好的表现. 而联合模型往往以原始图像作为输入,其中人体的尺度往往变化较大,导致模型性能降低. 因此一个可行的研究方向是将分离模型处理尺度变化的方法结合进联合模型中,提升联合模型对于尺度变化的鲁棒性.

为了进一步分析不同方法对关键点检测的准确率,我们使用 Ronchi 等人[93] 提供的分析工具对每个方法进行关键点定位误差分析,结果如图 7 所示. 我们主要关注四种误差:

(1) 抖动(jitter). 预测的关键点与真实值的距离处于下界 α 与上界 β 之间.

(2) 漏检(miss). 预测的关键点与真实值的距离大于下界 α.

(3) 倒置(inversion). 预测的关键点对于真实值来说是漏检,但是对于这个人的其他关键点来说是处于抖动或者更好的状态.

(4) 交换(swap). 预测的关键点对于真实值来说是漏检,但是对于其他人的关键点来说是处于抖

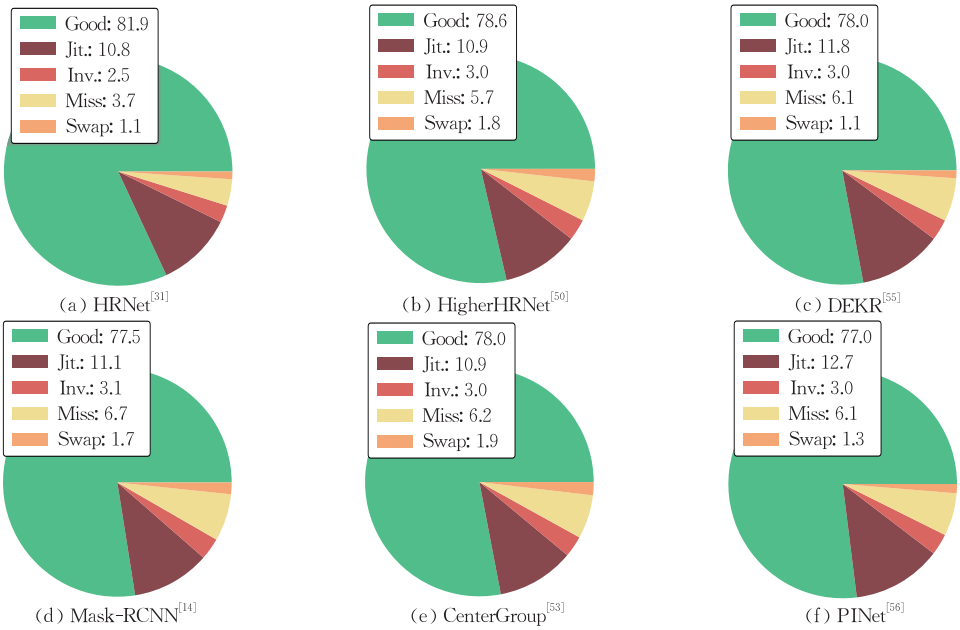


图 7　关键点定位误差分析(其中 Good 表示准确定位的比例,而 Jit. 等表示四种误差所占比例,详情见 6.1 节)

动或者更好的状态.

以上四种误差中,前两个关注模型对于关键点定位的精度,后两个则更加关注模型区分相似人体部件关键点的能力,即模型是否可以克服左右对称以及重叠人的干扰等影响.

从图 7 中我们可以发现,基于热力图的自顶向下和自底向上方法,在抖动和漏检方面要比基于向量场回归的方法要好. 这说明检测法比回归法可以取得更高的关键点定位精度. 在处理交换误差方面,向量场回归的 DEKR 和 PINet 性能显著高于 Mask-RCNN 和 HigherHRNet. 这说明单阶段回归方法将每个人的姿态视为一个整体来进行回归,使其不容易受到相邻人和相似部件关键点的影响. 此外,这类方法处理倒置的能力相对更好,说明回归方法在处理左右对称和重叠人的干扰两方面具有优势.

6.2　效率对比

表 6 比较了三类方法的推理效率,表中的 FPS 值是在 RTX3090 GPU 和单尺度测试场景下测得. 可以发现,基于热力图的两阶段方法,包括自顶向下和自底向上法,推理速率普遍较低. 例如,经典的 HRNet 和 HigherHRNet 的推理速度均低于 3FPS. 这一结果表明两阶段算法复杂度较高,难以实现高效的人体姿态估计. 基于向量场的单阶段回归方法能直接得到多人姿态估计结果,具有更好的推理效率. 例如,DEKR 可以实现 15FPS 以上的推理速度. 后续研究可以结合热力图的高定位准确率以及回归方法的高推理效率.

6.3　遮挡鲁棒性分析

实际应用场景中往往会出现人与人、人与物之间的遮挡,对模型的遮挡鲁棒性提出了更高的要求. 之前的研究主要在非拥挤场景下进行评测,无法很好地反映模型的抗遮挡能力. 我们将现有方法在密集场景下的多人姿态估计数据集 CrowdPose 和 OCHuman 上进行了测试,得到了如表 7 所示的结果.

表 7　密集场景下的性能对比

方法	CrowdPose				OCHuman		
	AP	AP_E	AP_M	AP_H	$AP^{\dagger}_{\text{test}}$	AP_{val}	AP_{test}
自顶向下							
HRNet[31]	69.3	—	—	—	—	37.8	37.2
Mask-RCNN[14]	57.2	69.4	57.9	45.8	20.2	—	—
自底向上							
HigherHRNet[50]	65.9	73.3	66.5	57.9	27.7	**40.0**	**39.4**
CenterGroup[53]	**70.0**	**76.8**	**70.7**	**62.2**	—	—	—
向量场回归							
DEKR[55]	65.7	73.0	66.4	57.5	52.2	37.8	36.4
PINet[56]	68.9	75.4	69.6	61.5	**59.8**	38.4	37.2

注:†表示在 OCHuman 验证集上训练,测试集测试的性能;否则是表示在 COCO 训练集训练,在 OCHuman 上测试的性能.

从表 7 可以看出，在密集场景下自顶向下方法表现比较差. 此类方法使用检测框来区分不同行人，因此对遮挡干扰较为敏感，容易在人和人发生重叠时导致漏检. 而自底向上和基于向量场回归的方法无需单人区域检测，因此在密集场景下也展现出了较好的性能. 其中，向量场方法在 OCHuman 数据集上取得了显著的性能优势. 因此，基于向量场的方法在效率和对遮挡鲁棒性上均有优异的表现. 后续研究可以关注进一步提升此类方法的定位性能.

7 未来研究方向展望

目前人体姿态估计研究虽然取得了许多进展，但是仍面临着很多挑战. 未来工作可以进一步对这些挑战开展深入研究. 包括

(1) 严重拥挤场景. 现实应用往往需要应对和处理多人互相遮挡的拥挤场景. 目前大多数工作仅关注如何在非拥挤场景下提升姿态估计性能，尚未专门考虑和处理拥挤问题. 为提升算法在现实应用中的性能，未来的研究应该关注严重拥挤场景下的多人姿态估计问题，提出新的方法和新的数据集.

(2) 视频中连续姿态估计和跟踪. 现有的基于单帧的方法能直接扩展到视频姿态估计任务中，如 SimpleBaseline[30] 等方法. 目前视频姿态估计的主流方法也大多应用了单帧姿态估计方法. 但是这些方法无法利用视频中的前后帧信息，通过追踪来快速定位关键点，效率较为低下. 此外，视频中存在更多的遮挡与拥挤场景，这对视频中的多人姿态估计提出了新挑战. 如何设计算法来有效利用前后帧关联处理遮挡问题是有价值的研究方向.

(3) 全身姿态估计. 现有的姿态估计方法主要关注预定义的十多个身体关键点，难以预测更加细微的人体姿态，如面部、手部和脚部等关键点. 可以对更加全面的全身姿态估计开展研究，通过预测更多的人体关键点，提供更加细致的人体姿态信息. 在此任务中，如何处理粗粒度关键点和细粒度关键点是一个重要问题. 另外，更大数量的关键点对模型设计也提出了新的挑战，例如需要高效关联关键点，并快速编解码大量关键点的位置.

(4) 算法实时性. 目前很多姿态估计算法已经在多个数据集上取得了较高的准确率. 然而这些算法往往具有较高的复杂度. 如表 6 所示，目前准确率较高的方法在 GPU 上的推理速度仍然难以达到实时的效果，也难以在算力较低的边缘设备上部署. 未来的研究同样需要关注算法的效率问题，通过设计轻量化的姿态估计框架，引入模型压缩和蒸馏等方法提升姿态估计算法的实时性.

(5) 数据域适应. 当前人体姿态估计数据集中的图像均为 RGB 图像，且训练集和测试集来源一致. 实际应用可能需要基于红外图像、绘画和动漫等跨模态的数据实现姿态估计. 领域差异会导致训练数据缺乏和跨域性能降低等问题. 目前，人体姿态估计领域尚未对领域差异问题开展研究. 研究相关领域适应算法，在不同数据分布之间实现姿态估计模型迁移同样是有价值的研究方向.

(6) 三维姿态估计. 目前的二维人体姿态估计技术已经相对成熟. 未来可以考虑更具挑战性的三维姿态估计任务. 三维姿态估计可以基于图像预测人体关键点的三维坐标，能够更好地辅助下游任务和应用. 未来的研究可以将二维姿态估计方法推广到三维姿态估计问题中，解决其面临的尺度变化、开放环境下的遮挡、计算复杂度高等问题.

8 总 结

近年来随着深度学习技术的进步，人体姿态估计方法迎来了快速发展. 本文系统地总结了近些年来二维姿态估计任务的研究进展，将现有方法按照关键点建模方式分类，并对每类方法进行了详细介绍. 本文还介绍了常用的姿态估计数据集和测评标准，总结和对比了近期方法在不同数据集上的性能. 最后，本文选取了代表性多人姿态估计方法，开展了详细的对比和分析，并对不同方法的错误预测和原因进行了讨论. 目前姿态估计研究还面临许多难点和挑战，我们希望这篇综述可以为姿态估计领域的未来研究带来启发.

作者贡献声明 李佳宁、王东凯对本文贡献相同，为共同第一作者.

参 考 文 献

[1] Wei L, Zhang S, Yao H, et al. GLAD: Global-local-alignment descriptor for pedestrian retrieval//Proceedings of the ACM International Conference on Multimedia. Mountain View, USA, 2017: 420-428

[2] Su C, Li J, Zhang S, et al. Pose-driven deep convolutional model for person re-identification//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 3960-3969

[3] Li J, Zhang S, Tian Q, et al. Pose-guided representation learning for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 622-635

[4] Zhang Z, Su C, Zheng L, et al. Correlating edge, pose with parsing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8900-8909

[5] Liu Z, Wu S, Jin S, et al. Towards natural and accurate future motion prediction of humans and animals//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10004-10012

[6] Liu Z, Lyu K, Wu S, et al. Aggregated multi-GANs for controlled 3D human motion prediction//Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 2225-2232

[7] Fischler M A, Elschlager R A. The representation and matching of pictorial structures. IEEE Transactions on Computers(ToC), 1973, 100(1): 67-92

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

[9] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[11] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks// Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91-99

[12] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 21-37

[13] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3431-3440

[14] He K, Gkioxari G, Dollár P, et al. Mask R-CNN//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2961-2969

[15] Tian Z, Shen C, Chen H. Conditional convolutions for instance segmentation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 282-298

[16] Liu H, Wang R, Shan S, et al. Deep supervised hashing for fast image retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2064-2072

[17] Yang H F, Lin K, Chen C S. Cross-batch reference learning for deep classification and retrieval//Proceedings of the ACM International Conference on Multimedia. Amsterdam, Netherlands, 2016: 1237-1246

[18] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014

[19] Tulyakov S, Liu M Y, Yang X, et al. MoCoGAN: Decomposing motion and content for video generation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1526-1535

[20] Zhang F, Zhu X, Wang C. Single person pose estimation: A survey. arXiv preprint arXiv:2109.10056, 2021

[21] Chen H, Feng R, Wu S, et al. 2D human pose estimation: A survey. arXiv preprint arXiv:2204.07370, 2022

[22] Dang Q, Yin J, Wang B, et al. Deep learning based 2D human pose estimation: A survey. Tsinghua Science and Technology, 2019, 24(6): 663-676

[23] Liu W, Mei T. Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. ACM Computing Surveys(CSUR), 2022, 55(4): 1-41

[24] Chen Y, Tian Y, He M. Monocular human pose estimation: A survey of deep learning based methods. Computer Vision and Image Understanding, 2020, 192: 102897

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

[26] Sun X, Xiao B, Wei F, et al. Integral human pose regression //Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 529-545

[27] Li J, Chen T, Shi R, et al. Localization with sampling-argmax. Advances in Neural Information Processing Systems, 2021: 27236-27248

[28] Li J, Bian S, Zeng A, et al. Human pose regression with residual log-likelihood estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 11025-11034

[29] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, USA, 2016: 4724-4732

[30] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and racking//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 466-481

[31] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5693-5703

[32] Bao Q, Liu W, Hong J, et al. Pose-native network architecture search for multi-person human pose estimation//Proceedings of the ACM International Conference on Multimedia. Seattle, USA, 2020: 592-600

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

[34] Zhang F, Zhu X, Dai H, et al. Distribution-aware coordinate representation for human pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7093-7102

[35] Nie X, Feng J, Zuo Y, et al. Human pose estimation with parsing induced learner//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2100-2108

[36] Liang X, Gong K, Shen X, et al. Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(4): 871-885

[37] Yu C, Xiao B, Gao C, et al. Lite-HRNet: A lightweight high-resolution network//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10440-10450

[38] Zhang F, Zhu X, Ye M. Fast human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3517-3526

[39] Li Z, Ye J, Song M, et al. Online knowledge distillation for efficient pose estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 11740-11750

[40] Fang H S, Xie S, Tai Y W, et al. RMPE: Regional multi-person pose estimation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2334-2343

[41] Li J, Wang C, Zhu H, et al. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10863-10872

[42] Shi D, Wei X, Yu X, et al. InsPose: Instance-aware networks for single-stage multi-person pose estimation//Proceedings of the ACM International Conference on Multimedia. 2021: 3079-3087

[43] Mao W, Tian Z, Wang X, et al. FCPose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 9034-9043

[44] Pishchulin L, Insafutdinov E, Tang S, et al. DeepCut: Joint subset partition and labeling for multi person pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4929-4937

[45] Insafutdinov E, Pishchulin L, Andres B, et al. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 34-50

[46] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7291-7299

[47] Kreiss S, Bertoni L, Alahi A. PifPaf: Composite fields for human pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11977-11986

[48] Li J, Su W, Wang Z. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation//Proceedings of the AAAI Conference on Artificial Intelligence. New York City, USA, 2020: 11354-11361

[49] Newell A, Huang Z, Deng J. Associative embedding: End-to-end learning for joint detection and grouping//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 2277-2287

[50] Cheng B, Xiao B, Wang J, et al. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5386-5395

[51] Luo Z, Wang Z, Huang Y, et al. Rethinking the heatmap regression for bottom-up human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13264-13273

[52] Jin S, Liu W, Xie E, et al. Differentiable hierarchical graph grouping for multi-person pose estimation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 718-734

[53] Brasó G, Kister N, Leal-Taixé L. The center of attention: Center-keypoint grouping via attention for multi-person pose estimation//Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada, 2021: 11853-11863

[54] Zhou X, Wang D, Krähenbü hl P. Objects as points. arXiv preprint arXiv:1904.07850, 2019

[55] Geng Z, Sun K, Xiao B, et al. Bottom-up human pose estimation via disentangled keypoint regression//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 14676-14686

[56] Wang D, Zhang S, Hua G. Robust pose estimation in crowded scenes with direct pose-level inference//Proceedings of the Advances in Neural Information Processing Systems. 2021: 6278-6289

[57] Nie X, Feng J, Zhang J, et al. Single-stage multi-person pose machines//Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea, 2019: 6951-6960

[58] Wei F, Sun X, Li H, et al. Point-set anchors for object detection, instance segmentation and pose estimation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 527-544

[59] Mao W, Ge Y, Shen C, et al. TFPose: Direct human pose estimation with transformers. arXiv preprint arXiv: 2103.15320, 2021

[60] Li K, Wang S, Zhang X, et al. Pose recognition with cascade transformers//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 1944-1953

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

[62] Li Y, Zhang S, Wang Z, et al. TokenPose: Learning keypoint tokens for human pose estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 11313-11322

[63] Ludwig K, Harzig P, Lienhart R. Detecting arbitrary intermediate keypoints for human pose estimation with vision transformers//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2022: 663-671

[64] Tang W, Wu Y. Does learning specific features for related parts help human pose estimation?//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1107-1116

[65] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7103-7112

[66] Sánchez D, Oliu M, Madadi M, et al. Multi-task human analysis in still images: 2D/3D pose, depth map, and multi-part segmentation//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Lille, France, 2019: 1-8

[67] Kamel A, Sheng B, Li P, et al. Hybrid refinement-correction heatmaps for human pose estimation. IEEE Transactions on Multimedia, 2020, 23: 1330-1342

[68] Li Y, Hao M, Di Z, et al. Test-time personalization with a transformer for human pose estimation//Proceedings of the Advances in Neural Information Processing Systems. 2021: 2583-2597

[69] Wang Y, Li M, Cai H, et al. Lite pose: Efficient architecture design for 2D human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 13126-13136

[70] Huang S, Gong M, Tao D. A coarse-fine network for keypoint localization//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 3028-3037

[71] Artacho B, Savakis A. UniPose: Unified human pose estimation in single images and videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7035-7044

[72] Papandreou G, Zhu T, Kanazawa N, et al. Towards accurate multi-person pose estimation in the wild//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4903-4911

[73] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779-788

[74] Girshick R. Fast R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440-1448

[75] Hidalgo G, Raaj Y, Idrees H, et al. Single-network whole-body pose estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 6982-6991

[76] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 764-773

[77] Shi D, Wei X, Li L, et al. End-to-end multi-person pose estimation with transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11069-11078

[78] Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation//Proceedings of the British Machine Vision Conference. Aberystwyth, UK, 2010: 5

[79] Johnson S, Everingham M. Learning effective human pose estimation from inaccurate annotation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 1465-1472

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

[81] Zhang S H, Li R, Dong X, et al. Pose2Seg: Detection free human instance segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 889-898

[82] Tompson J J, Jain A, Lecun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimationestimation//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014

[83] Tompson J, Goroshin R, Jain A, et al. Efficient object localization using convolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 648-656

[84] Carreira J, Agrawal P, Fragkiadaki K, et al. Human pose estimation with iterative error feedback//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4733-4742

[85] Hu P, Ramanan D. Bottom-up and top-down reasoning with hierarchical rectified Gaussians//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5600-5609

[86] Chen Y, Shen C, Wei X S, et al. Adversarial PoseNet: A structure-aware convolutional network for human pose estimation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1212-1221

[87] Yang W, Li S, Ouyang W, et al. Learning feature pyramids for human pose estimation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1281-1290

[88] Ke L, Chang M C, Qi H, et al. Multi-scale structure-aware network for human pose estimation//Proceedings of the European Conference on Computer Vision(ECCV). Munich, Germany, 2018: 713-728

[89] Tang W, Yu P, Wu Y. Deeply learned compositional models for human pose estimation//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 190-206

[90] Cai Y, Wang Z, Luo Z, et al. Learning delicate local representations for multi-person pose estimation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 455-472

[91] 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//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 269-286

[92] Kocabas M, Karagoz S, Akbas E. MultiPoseNet: Fast multi-person pose estimation using pose residual network//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 417-433

[93] Ronchi M R, Perona P. Benchmarking and error diagnosis in multi-instance pose estimation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 369-378

LI Jia-Ning, Ph.D. His major research interests include human pose estimation and person re-identification.

WANG Dong-Kai, Ph. D. candidate. His major research interests include human pose estimation and person re-identification.

ZHANG Shi-Liang, Ph. D., associate professor with Tenure. His research interests include large-scale fine-grained image retrieval and recognition.

Background

2D human pose estimation is an important and fundamental task in computer vision. This task aims to identify and locate the human body keypoints of each person in images, which can support multiple downstream tasks and can be applied to many real-world applications. Recent years, with the rapid development of deep learning, significant progresses have been made to human pose estimation. However, there is lack of a survey to summarize recent progresses in pose estimation.

This paper provides a survey which summaries recent progresses in the field of pose estimation. In this paper, we first introduce the research background, problem definition, task difficulty and problem formulation of human pose estimation task. Next, we introduce the representative single-person and multi-person pose estimation methods, respectively. Next, we summarize the widely-used datasets, benchmark metric, and the performance of representative methods on these datasets. Finally, we discuss the remaining challenges and promising research directions in human pose estimation. We believe will promote the development of pose estimation task.

This work is supported in part by the National Natural Science Foundation of China under Grant Nos. U20B2052, 61936011, in part by the National Key Research and Development Program of China under Grant No. 2018YFE0118400.