激光与光电子学进展 先进成像

基于深度学习的人体姿态估计方法综述

卢健，杨腾飞*，赵博，王航英，罗毛欣，周嫣然，李哲
西安工程大学电子信息学院，陕西 西安 710048

摘要 全面综述了基于深度学习的人体姿态估计方法的研究进展。在比较分析各类单人姿态估计方法的基础上，从自上而下和自下而上两个方法角度总结了多人姿态估计算法。在自上而下方法中，着重介绍了局部区域重叠、关节点混淆、人体非典型部位关节点难以检测等问题的解决方案；在自下而上的方法中，重点关注聚类方法对关节点检测的贡献。对目前公共数据集上取得优异性能的代表性方法进行了对比和分析。这样做的目的是使研究者了解和熟悉该领域已有的研究成果，拓展研究思路和方法，并展望未来可能出现的研究方向。

关键词 机器视觉；深度学习；人体姿态估计；关节点检测；公共数据集
中图分类号 TP391.41；TP183 **文献标志码** A **doi**：10.3788/LOP202158.2400005

Review of Deep Learning-Based Human Pose Estimation

Lu Jian, Yang Tengfei*, Zhao Bo, Wang Hangying, Luo Maoxin, Zhou Yanran, Li Zhe
School of Electronics and Information, Xi'an Polytechnic University, Xi'an, Shaanxi 710048, *China*

Abstract The research progress of human pose estimation method based on deep learning is comprehensively summarized. On the basis of comparison and analysis of various single-person pose estimation methods, a variety of multi-person pose estimation algorithms are summarized from the top-down and bottom-up approaches. In the top-down approach, the solutions to local area overlap, articulation point confusion, and difficulty in detecting the articulation point of atypical parts of human body are mainly introduced. In the bottom-up approach, the contribution of clustering method to articulation point detection is emphasized. Representative methods to achieve excellent performance on current public datasets are compared and analyzed. The review enables researchers to understand and familiarize themselves with the existing research results in this field, expand research ideas and methods, and look forward to the possible research directions in the future.

Key words machine vision; deep learning; human pose estimation; articulation point detection; public dataset
OCIS codes 150.1135; 110.2970

1 引 言

人体姿态估计(human pose estimation)是一种以人体骨骼关节点为研究对象，通过检测关节点的位置信息，估计关节点之间的联系进而重构人体肢干的方法，是完成人体行为识别[1-2]、姿态跟踪[3-4]、人物图像生成[5]和人机交互[6]等高级任务的基础环节，相关研究受到广泛关注。传统的人体姿态估计方法[7]依赖于人工标注特征，将姿态估计问题看作回归问题，直接回归出关节点的坐标，估计精度不高。其主要缺点有两方面：一方面局限于站立、静坐等单帧简单姿态，对摔倒、弯腰等复杂连续姿态的鲁棒性较差；另一方面，所使用的回归模型可扩展性较差，很难适应人体图像的多尺度变化。随着深度学

收稿日期：2020-11-16；**修回日期**：2021-02-06；**录用日期**：2021-03-03
基金项目：国家自然科学基金(61971339，61471161)、陕西省自然科学基金(2018JQ4016)、西安市碑林区应用技术研发项目(GX2007)
通信作者：*1062021763@qq.com

习广泛应用于人体姿态估计领域，以沙漏模型及其变体为代表的卷积神经网络（CNN）[8]在处理人体姿态估计问题时表现卓越。相关综述类报告总结了姿态估计方法，研究人员初步地了解相关研究的基本方法和进展，但是未考虑相关文献之间的关联性。

本文按照时间顺序从方法、模型层面系统地对姿态估计处理单人和多人问题进行了综述，整合并摘要人体姿态估计领域内已知的研究成果，并重点介绍了姿态估计中突破性、开创性以及具有里程碑意义的工作。目的是使研究者了解和熟悉所要研究问题的已有研究成果，拓展研究思路和研究方法，并展望未来可能出现的研究方向。表 1 展示了多个常用的人体姿态估计数据集的基本信息。

表 1　国内外数据集信息

Table 1　Information of domestic and foreign datasets

Type	Dataset name	Data source	Number of samples	Number of nodes	Status of use
Multiplayer/Single	MPII Human Pose[9]	Extract from YouTube video	＞250000	16	Research dataset
Multiplayer	Common Objects in Context[10]	Yahoo Web Albums	＞300000	18	Research dataset
Multiplayer	AI Challenge	Provided by Meitu	＞270000	14	Competition dataset
Multiplayer	MSCOCO	Provided by Microsoft	＞300000	18	Research dataset
Multiplayer	PoseTrack	Provided by the PoseTrack team	＞1356	15	Latest dataset
Single	Frames Labeled in Cinema[11]	30 Hollywood movies	＞20000	9	Basic deprecation
Single	Leeds Sports Pose dataset[12]	Sports people on Flickr	＞20000	14	Basic deprecation

2　单人姿态评估方法

对于单人姿势估计，传统的机器学习方法主要采用线性判别函数，往往难以在大量复杂相似的样本上取得好的检测效果。而深度学习凭借强大的自主学习能力和高度的非线性映射特性，可以得到语义信息更为丰富的特征，能够获得不同感受野下多尺度多类型的人体关节点特征向量和每个特征的全部上下文（contextual）信息，摆脱对部件模型结构设计的依赖。

2.1　基于卷积神经网络的姿态估计

使用 CNN 进行人体姿态估计的方法可以分为基于检测[13-14]和基于回归[15-16]的方法。基于检测的方法用一种概率分布图即热图（heatmap）的方式来表示关节点位置，估算图像中每一个像素所对应的概率值，像素越接近真实关节点，概率值就越大，反之则越小。而基于回归的方法以二维坐标的形式表示关节点的位置，通过学习从身体特征到身体部位位置的映射，训练网络，直接得到每个关节点的坐标。

Toshev 等[15]使用 AlexNet 作为基本网络结构对关节坐标进行回归，通过设计的级联网络初步计算得到一个关节点的坐标，然后根据这个坐标，重新在原始图片中获得局部图片，利用局部图像信息进行更高精度的坐标计算。虽然级联网络提高了姿态估计回归网络的精准度，但是并不适用于输入图像分辨率比较小的情况。同时，由于采用了级联的方法，对于每一张输入图像，多次使用 CNN 进行卷积操作时计算复杂度过高。针对 DeepPose[15]在姿势估计中不考虑局部外观，难以估计复杂的人体姿势等问题，Fan 等[17]提出了一种基于双源深度卷积神经网络（DS-CNN）的人体姿态估计方法，该方法将原始的单源 R-CNN 扩展为双源模型 DS-CNN，为研究人员提供了整体视图姿势估计。DS-CNN 采用从输入图像中检测到的一组与类别无关的图像块进行训练，与许多以前的人体姿势估计方法中用作输入的滑动窗口或完整图像相比，图像块可以在多个尺度上捕获具有更好语义含义的局部身体部位。作者将局部（身体）部位图像和每个局部部位的整体视图作为单独的输入，共同输入到 DS-CNN 中统一学习，以实现联合检测，确定图像块是否包含人体关节及关节位置，找到图像块中关节点的确切位置。最后，对来自所有图像块的这些联合检测/定位结果进行组合，以实现更准确的人体姿势估计。Pfister 等[18]提出了一种在深度卷积网络中使用光流进行关节点定位的新方法，该方法将姿态估计看作是检测问题，以热图的形式进行输出，通过使用光流，在多个帧中组合信息，以达到从上下文中受益的目的。

如图 1 所示，将给定帧 t 及上下 n 帧时间邻域图像作为一组输入帧，通过空间卷积(SpatialNet)分别回归每组输入帧的热图，然后使用密集的光流将每组热图单独翘曲(warped)至当前帧 t，将翘曲的热图汇集到每个关节上形成单个热图，取其中置信度最大的热图来进行姿态估计。回归热图而不是直接回归坐标(x,y)的好处：1)可以充分理解关节点定位的失败过程和可视化网络的“思维”过程；2)网络输出多个空间位置的置信度，而不是只考虑单个位置的置信度，使得学习更加立体。在 BBC pose 数据集中，Pfister 等[18]提出的方法成效显著，特别是对手腕识别的精准度，当定位的关节位置与真值位置的距离在标定的联合中心的 6 个像素距离内时，提出的方法的精准度超过之前最佳方法[16]的 10%。

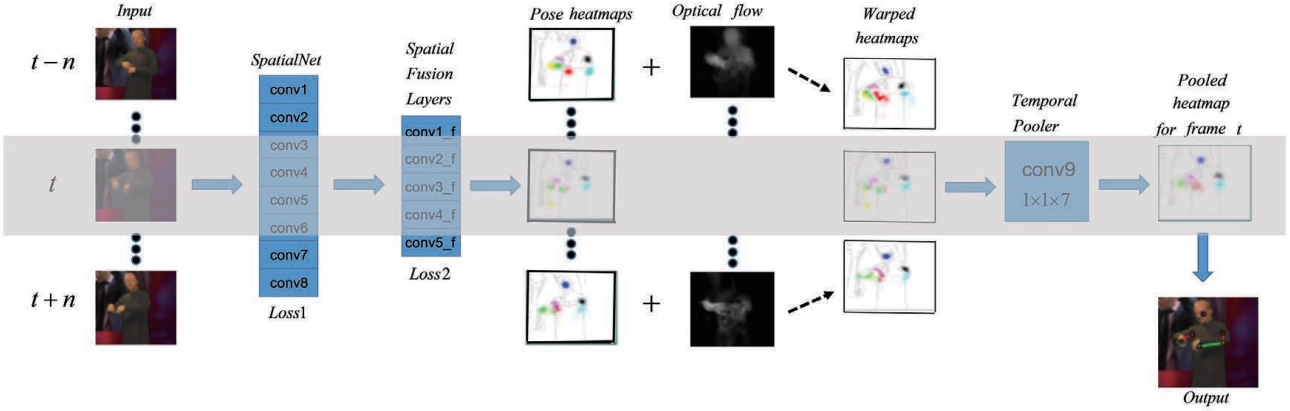

图 1　时态姿态估计网络[18]

Fig. 1　Temporal pose estimation network[18]

在人体姿态估计问题中，对于被遮挡身体部位的位置估计是一个较难解决的问题。因此，Bulat 等[19]提出了一种检测-跟随-回归的级联网络，如图 2 所示。该网络通过两个子网络组成的级联网络对人体遮挡/被遮挡的关节点进行检测，其中部件检测网络(part detection network)[20]用来检测身体各个关节位置，并输出各关节点位置的热图。但是，输出的被遮挡的关节点热图的置信度普遍较低。因此，将原始图像与部件检测网络输出的热图一起堆叠并输入到深度回归网络(regression network)中，通过回归被遮挡关节点真实位置附近的一组置信度图来预测被遮挡关节点的位置，进一步提高检测器在检测特定人体部件被遮挡时的鲁棒性。为了解决在同一区域中存在多个人体部位使得关节点位置难以预测的问题，文献[19]遵循文献[20]的方法，将真值位置周围特定半径内的信息设置为 1，剩余的背景设置为 0，得到一种每个人体部位都对应一个标签信息的二进制映射，使得关节点与人体之间能够正确关联。检测网络仅使用可见的身体部位进行训练。使用内核大小为 1 的卷积层对 VGG-16 网络的全连接层进行替换，并加入 ResNet-152 网络[21]作为部件检测基础网络进行实验，结果证明，这种使用残差网络作为网络架构的方法在 MPII 数据集上可以对被遮挡关节点达到最高检测精准度。Lifshitz 等[10]提出一种使用深度卷积神经网络进行关节点投票而不是关节点检测的新方法，每个身体部位的

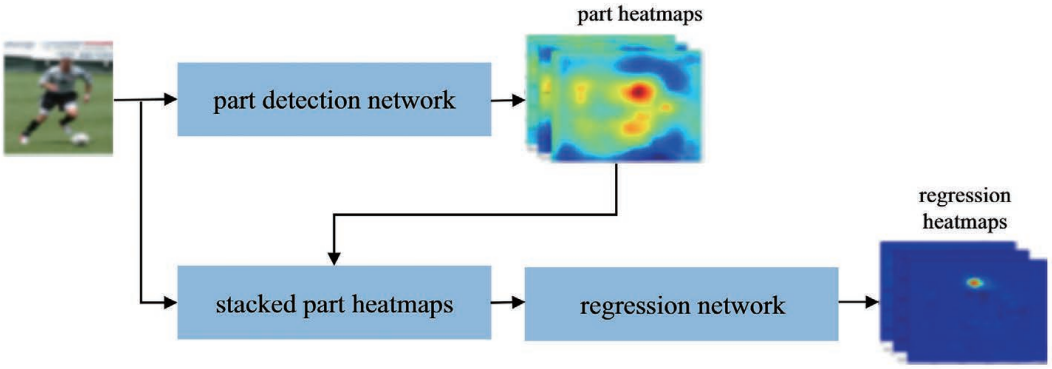

图 2　检测-跟随-回归的 CNN 结构[19]

Fig. 2　CNN structure of detection-following-regression[19]

位置都是由多个位置的投票信息通过卷积聚合来确定的。与关节点检测方法相比，关节点投票方法具有以下优势：1）通过随机利用被检测图像的人体区域内的多个像素点进行投票，产生可靠的关节点预测；2）人体图像的任何位置信息都可能为检测多个不同的关节点做出贡献。如图 3 所示，给定一个像素点 y 并在其周围划分 50 个对数极化分组（log-polar bins），通过关节点所落入分组的位置，预测每一个关节点相对于像素点 y 的方位，以此类推，由多个像素点投票得出表示每个关节点在每个位置概率的热图，最后将这些投票出的热图组合成单人姿态估计结果。

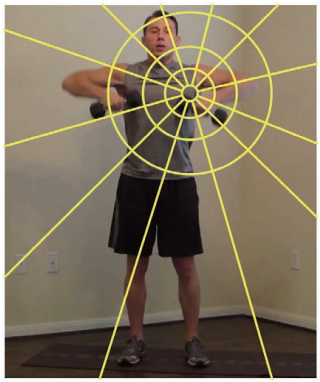

图 3　对数极化分组

Fig. 3　Logarithmic polarization grouping

对于背景杂乱和自我遮挡等问题，Chu 等[22]首次提出多语境注意力（multi-context attention）模型，该模型将图像分为人体和人体局部关节两个部分，通过堆栈沙漏网络（stacked hourglass networks）生成具有不同分辨率特征的注意力图（attention maps），不同分辨率特征对应着不同的语义。将多语境注意力机制整合到 CNN，通过端到端进行人体姿势估计。网络整体框架如图 4 所示，其中每层沙漏网络输入不同分辨率的图像，使得各结构间的输出产生尺度上的差异，得到代表不同尺度特征的热图。底层堆栈（stack 1～stack 4）通过整体注意力模型（holistic attention model）得到图像的全局注意力图；高层堆栈（stack 5～stack 8）通过分层注意力机制对人体局部关节点进行缩放处理，以得到人体局部注意力图。通过学习的全局注意力图来区分背景区域与人体区域，然后在不考虑背景的情况下，基于人体区域来得到人体关节点的局部注意力图；通过注意力图，网络关注难样本（hard negative samples），最后通过沙漏堆栈网络得出人体关节点热图，进而估计人体姿态。依靠 CNN 进行人体姿态估计的方法着重于关节的检测，通过探索目标图像中关节之间的几何形状来进行姿态的估计，缺点是需要独立的分支来检测框架中人类目标的边界框。Artacho 等[23]基于“瀑布式”萎缩空间池架构（WASP）[24]，提出统一的人体姿态估计框架（UniPose）。作为对以前工作的改进，UniPose 不需要单独的分支来进行边界框和关节检测，结合上下文分割和联合定位功能，既确定了关节点的位置，又确定了用于人体检测的边界框，采用单阶段，在不依赖统计后处理方法的情况下，高精度地估计人体姿势。该框架主要组成部分是 waterfall atrous spatial pooling（WASP）模块，得益于 WASP 模块，将空洞卷积的级联方法与从空洞空间金字塔池化（ASPP）模块[25]的并行结构中获得的较大视野（FOV）相结合；使用上下文信息来预测关节的位置，包含整个框架的信息，不需要基于统计或几何方法的后分析。此外，相关文献将此方法已扩展到 UniPose-LSTM 以进行多帧处理，并在视频姿势估计中获得了较优的结果。

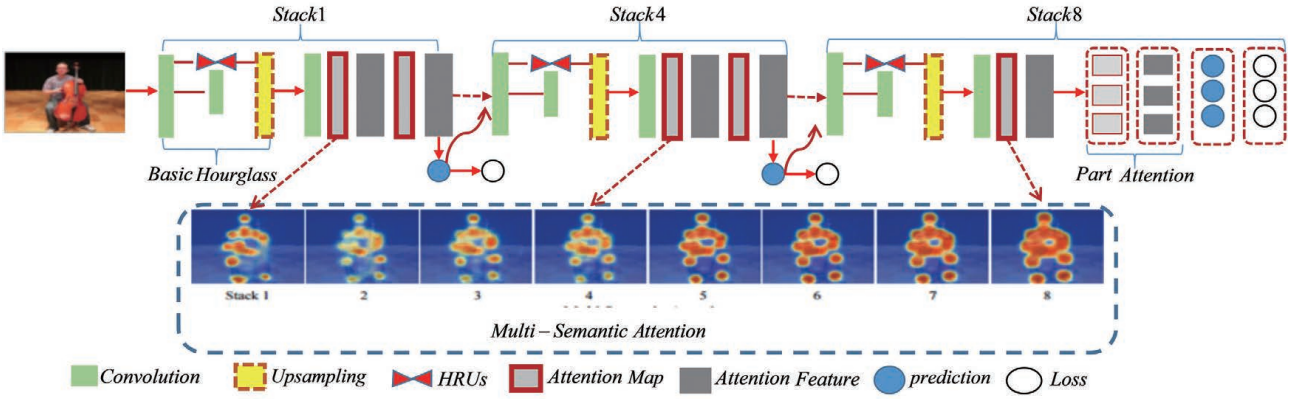

图 4　多语境注意力机制框架[22]

Fig. 4　Multi-context attention mechanism framework[22]

2.2　基于生成对抗网络的姿态估计

针对人体姿态估计缺乏有效特征表示等问题，Chou 等[26]提出了一种使用生成对抗网络(generative adversarial networks)的解决方法。其建立了生成器和鉴别器两个具有相同架构的堆栈沙漏网络框架，如图 5 所示。生成器网络是一个具有残差块的全卷积网络，输入图像通过生成器向前馈送后，得到一组代表每个关节点在每个位置的置信度得分的热图。将生成器网络生成的热图和真实热图馈送到鉴别器网络，重建两组热图，以分别计算出鉴别器的输出与真实热图及其与生成器生成的热图之间的损失 L_{real} 和 L_{fake}。为了解决生成器与鉴别器之间训练不平衡导致过拟合的问题，作者使用变量 k_t 来控制发生器和鉴别器之间的平衡。自适应项 k_t 可表述为

$$k_{t+1}=k_t+\lambda_k(\gamma L_{\text{real}}-L_{\text{fake}}), \tag{1}$$

式中：k_t 表示对 L_{fake} 的重视程度；t 为迭代次数；γ 和 λ_k 为超参数。当 $L_{\text{fake}}<\gamma L_{\text{real}}$ 时，说明生成器生成的热图足够真实，可以欺骗鉴别器，因此 k_t 值将增加使 L_{fake} 更受重视。同时对鉴别器进行更多的训练以识别生成的热图，其中训练的速度取决于鉴别器与生成器之间的距离，即 $\gamma L_{\text{real}}-L_{\text{fake}}$ 值。同样，当鉴别器优于生成器时，即 $L_{\text{fake}}>\gamma L_{\text{real}}$ 时，k_t 将减小以放慢对网络的训练，以便生成器的输出可以跟上鉴别器输出。

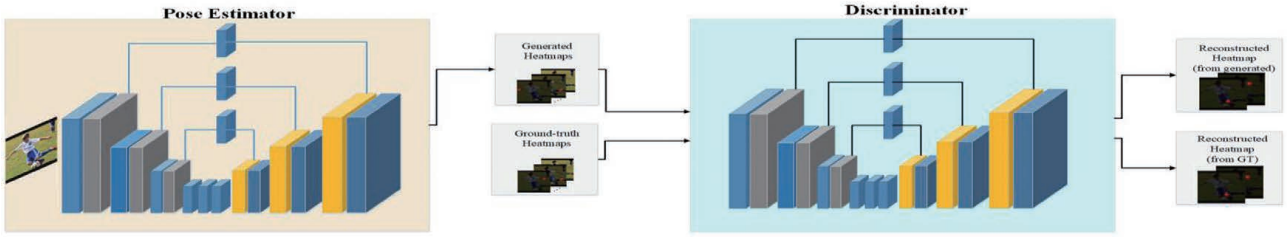

图 5　生成对抗网络的框架[26]

Fig. 5　Framework of generative adversarial networks[26]

估计被周围人的身体部位重度遮挡或看起来与身体部位相似背景遮挡的身体部位时，必须考虑人体关节结构的先验模型，如图 6 所示。解决这个问题的关键是从大量的训练数据中学习真实的身体关节点分布。但是，直接学习这种结构先验有时会产生生物学上难以理解的估计姿态。Chen 等[27]结合人体结构先验知识，提出了一种结构感知卷积网络来估计被遮挡的人体部位，并且运用一种具有两个鉴别器的新型条件对抗网络(adversarial PoseNet)来训练姿态生成器。如图 7 所示，条件对抗网络模型由三部分组成：姿态生成网络 G、姿态鉴别网络 P 和置信鉴别网络 C。姿态生成网络输入 RGB 图像，针对每张输入图像，输出每张输入图像的 32 个热图，其中一半姿态热图针对 16 个关节点的姿态进行估计，而另一半闭塞热图针对相应的遮挡进行预测，每个热图中的值都代表[0,1]范围内的置信度分数。姿态鉴别器网络 P 有两个作用：1)判别预测的姿态在人体几何学上的合理性；2)判别预测姿态热图的置信度得分。由于实际预测的热图往往不是高斯分布的，因此作者通过设计的置信鉴别网络 C，将低的

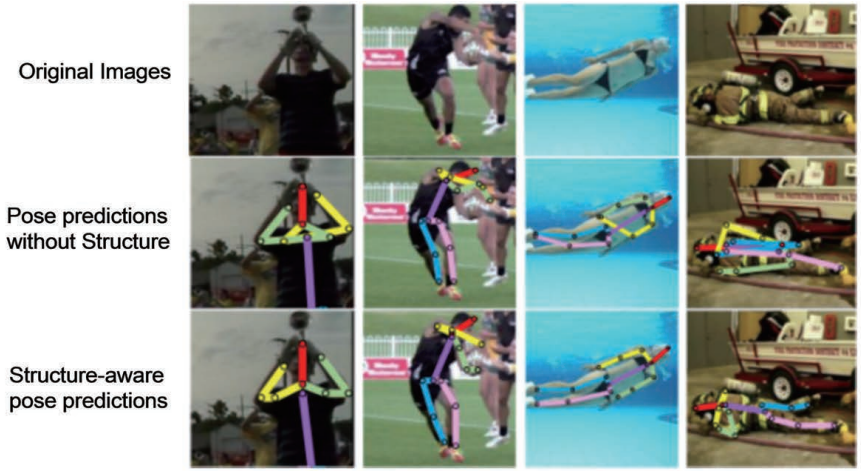

图 6　网络训练期间使用人体结构先验和不使用的结果对比[27]

Fig. 6　Comparison of the results with and without human structure prior during network training[27]

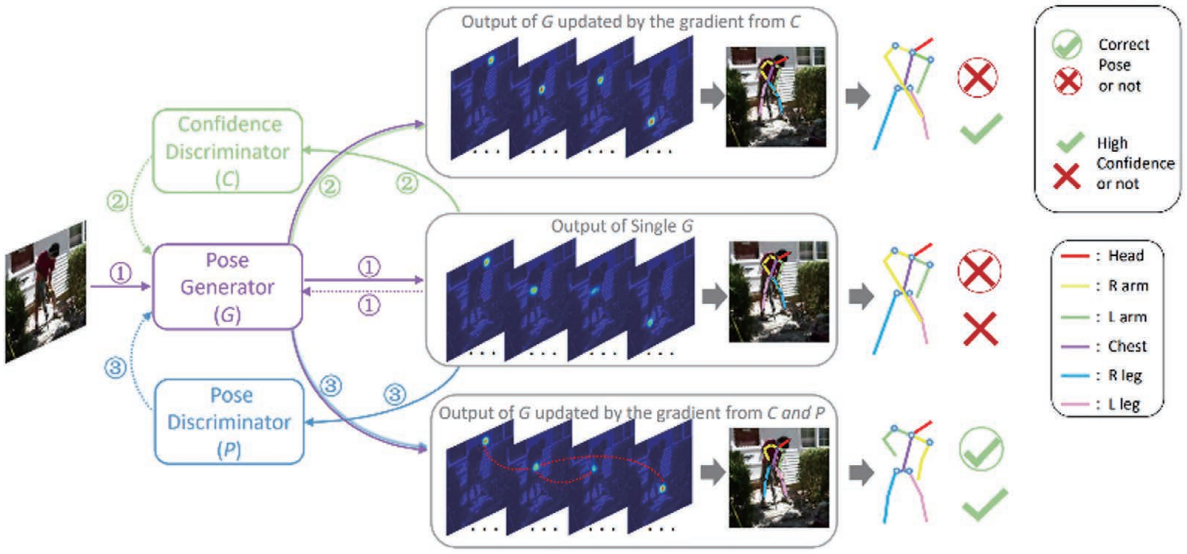

图 7　Adversarial PoseNet 模型结构[27]

Fig. 7　Adversarial PoseNet model structure[27]

置信度值预测和高的置信度值预测区分开，采用具有高、低置信度分数的热图作为姿态鉴别器 P 网络的输入。当预测部位存在遮挡时，姿态生成网络 G 会产生置信度低的热图，置信鉴别网络 C 会将这些热图设置为赝品，同时逼迫姿态生成网络 G 产生更高置信度的热图，从而提升对遮挡部分定位的精准度。对于原始 RGB 输入图像，作者以目标人物为中心对图像进行裁剪，并将图像块调整到 256×256 像素大小。通过对图像进行旋转（+/−30°）和缩放（0.75～1.25），训练网络对不同尺度和方向的图像更具鲁棒性。实验结果表明，此网络对人体的遮挡、重叠和扭曲问题更具鲁棒性，即使网络出现故障，输出也更接近人眼的真实预测，而不是“机器”预测。LSP 数据集中，在归一化距离为 0.2 的 PCK（关键点正确估计的比例）性能下，相较之前的方法，此方法的精准度平均提高了 2.4%。在 MPII 数据集中，此方法获得了当时 92.1%的最佳 PCKh 得分。

所谓尺度不稳定性，来自人体检测器的输入边界框的轻微扰动，导致姿态估计结果改变。对于人体姿态估计尺度不稳定问题，一般的沙漏结构方法会在特定尺度下过度拟合身体关节点，导致单一尺度“统治”。解决方法是在不同尺度下进行多次重复训练，进行姿态估计，并输出准确度最高的结果，但是这种方法缺乏一致的比例表示。Yang 等[28]也注意到了人体形状变化和视角改变，身体部位的比例不一致，导致身体部位检测器难以正确定位身体部位关节的问题，提出了一种金字塔残差模型（PRM），如图8所示，采用堆栈沙漏网络作为基本

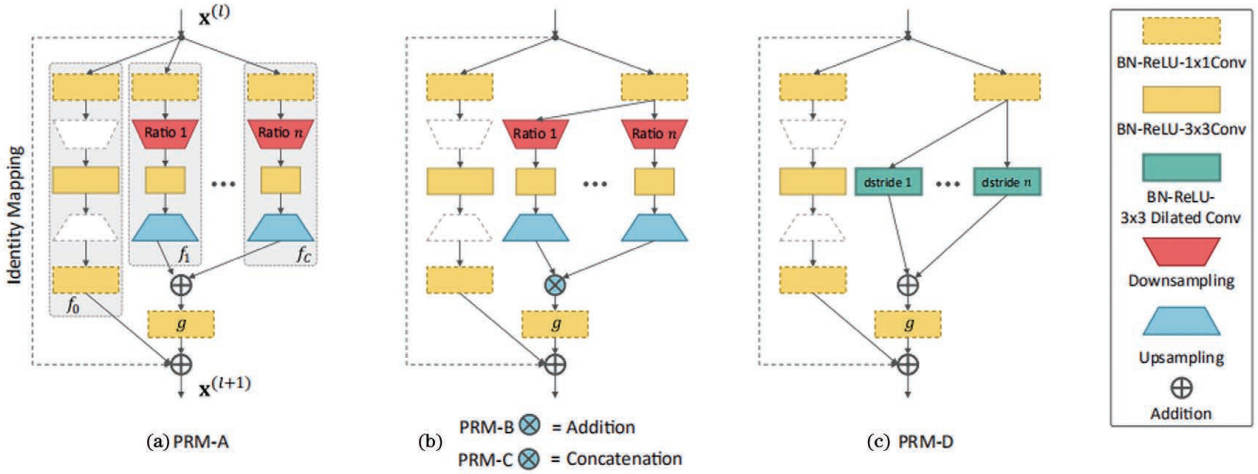

图 8　PRM 结构[28]

Fig. 8　PRM structure[28]

网络结构，构建具有学习功能的多尺度特征金字塔，以实现增强网络尺度不变性的目的。Ke 等[11]提出了一种基于多尺度结构感知神经网络的方法，该方法将多尺度监督网络(MSS-Net)与多尺度回归网络(MSR-Net)结合，以匹配多尺度的特征，提高关节点定位的鲁棒性。

多尺度监督网络和多尺度回归网络都使用结构感知损失，从多尺度特征中学习人体骨骼结构特征，这些特征可以为恢复复杂场景下被遮挡身体部位提供强大先验。而这种将多尺度监督与回归网络紧密结合的方法有效地使用多尺度特征定位关节点，通过多个关节点之间的结构关系进行全局姿态估计。多尺度结构感知网络如图 9 所示，通过微调网络来生成有效的训练样本，重点放在复杂场景中被遮挡的关节点上。该方法相较于深度卷积沙漏模型有 4 个关键改进：1)组合不同尺度特征的热图，加强身体关节点的特征学习；2)通过多尺度回归网络，全局优化多尺度特征的结构匹配；3)在中间监督和回归中使用结构感知损失，改善各相邻关节点之间的匹配，以推断出更高阶的匹配模型；4)通过相邻关节点之间的关系，定位被遮挡的关节点。在多尺度监督网络中，在每个反卷积层上添加分层丢失项，以允许对网络中每层特定比例的特征进行监督。这种方法可以有效地帮助模型多尺度地学习人体关节点的局部特征，促进网络对多尺度特征有效的学习，更好地捕获人体关节点的局部特征。此外，沿着分辨率金字塔由粗到细的解卷积也遵循类似于注意力机制的方法，关注局部关节点的定位并且对关节点之间的匹配进行改进。多尺度回归网络从多尺度监督网络堆栈中获取输入，通过融合多个关节点热图来确定姿态输出，从而进行全局关节点回归。从文献[29]可以看出，尽管沙漏堆栈越深姿态估计结果越好，但是越深的沙漏堆栈越容易导致网络训练过程中出现梯度消失问题。为此，文献[11]引入了一个人体骨架图来定义结构感知损失。如图 9 左下图所示，根据人类骨骼图确定出结构上相互连接的关节点，可以更好地捕获人体中关节点的物理连通性，从而获得结构先验。该方法在 FLIC 数据集中的 PCK 达到 99.2%(其中肘部、手腕为 97.3%)，在 MPII 数据集中取得了 92.1%的精准度。

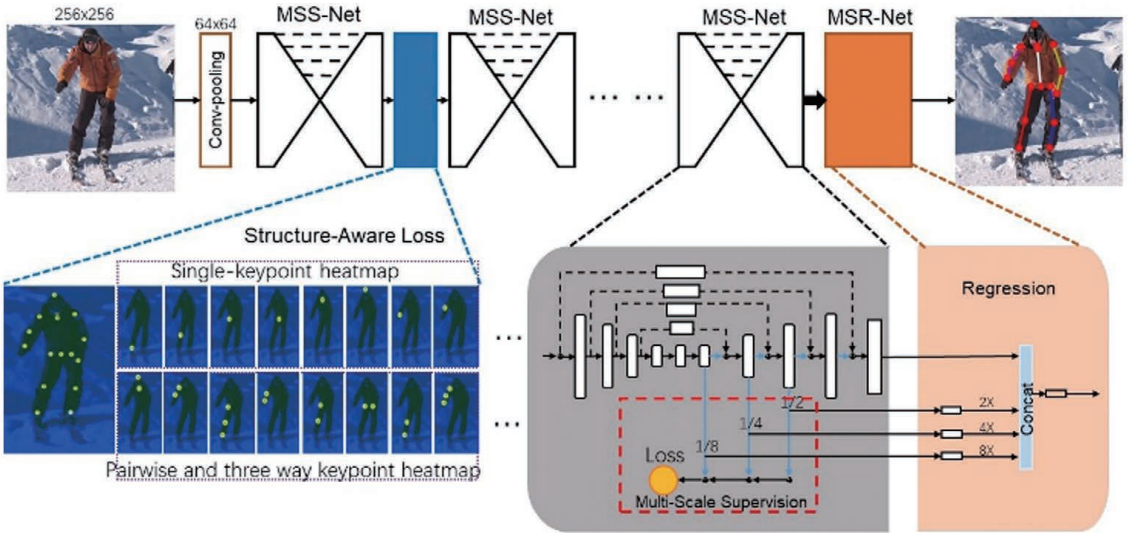

图 9　多尺度结构感知网络[11]

Fig. 9　Multi-scale structure-aware network[11]

2.3　基于组合模型的姿态估计

在过去的十年中，组合模型已经在多个人体姿态估计方法[30-32]中被采用。人体组合模型意味着将整个人体表示为部分和子部分的层次结构，并且满足一些关节约束，这种层次结构能够捕捉人体各部件之间的结构关系。组合模型使用一组离散型变量来模拟部件之间的兼容性，不仅包括部件的方向和比例信息，还包括跨语义类(如直臂和弯屈臂)信息。由于部件的不同类型与其所有子部件类型的组合非常多，由此形成的更高级别部件的状态空间可能呈指数增长，这对于计算和存储的要求都很苛刻。为了最大可能地解决这一问题，Tang 等[33]提出一种通过学习人体复杂的结构关系以进行姿态估计的深度学习成分模型(DLCM)，同时还提出了一种基于人体骨骼组成的空间本地信息摘要(SLIS)表示模型，精确编码每个部件的比例、方向和形状，以避免出现错误的部件组合方式，分别包括 16、12 和 6 个部分的三种语义级别的组合模型。与先前的姿态估

计网络相比，DLCM 具有跨多个语义级别的分层结构，并且针对人体姿态的估计具有类似于多人姿态估计的自下而上/自上而下网络阶段。在此之前所有的基于 CNN 的人体姿态估计结构模型都没有将实体分解为有意义和可重用的层次结构，也没有在不同的语义级别之间进行推断。文献[30]中的方法与它们的不同之处在于：1)网络更具层次性；2)学习了身体部位之间的成分关系；3)拥有包括跨越多个语义级别的自下而上和自上而下的网络阶段；4)利用新颖的部件表示来监督网络的训练。

DLCM 通过网络来学习身体部位之间的成分关系。它具有分层组合架构和自下而上/自上而下的推理阶段。在自下而上阶段，目标关节点的热图直接从图像中回归，通过人体关节子节点递归出更高级别部件的热图。在自上而下阶段，较低级别部分的热图使用与其关联的父关节得分图以及自下而上阶段的自身得分图进行递归细化描述；使用文献[34]中应用的均方误差(MSE)损失来区分预测的热图与真实热图，以引导网络学习身体部位之间的正确连接关系。在 FLIC 数据集上，DLCM 在手腕识别上的精准度较文献[11]提升了 1.5%。在 MPII 数据集上，分别在踝关节，膝关节，髋关节，腕关节和肘关节的精准度提升了 2.6%，2.0%，1.7%，1.6%和 1.4%。在这两个数据集中，该方法的精准度超越了之前最先进方法的精准度，同时与 8 层沙漏堆栈模型[31]及特征金字塔模型[28]相比，3 级 DLCM 具有明显更少的参数量和更低的计算复杂性。

基于深度学习的单人姿态估计方法更多的是作为人体姿态估计的基础方法。基于回归的方法通过多分辨率提取图像特征学习从身体特征到身体部位位置的映射，利用多阶段网络直接得到每个关节点的坐标，进而估计人体姿态。但是这种直接回归关节点坐标的方法不能很好地学习到人体关节点之间的结构信息。基于输出热图的方法通过构建概率图模型或者利用多尺度感受野学习到了关节点的结构信息，通过设计的级联网络得出精确的关节点坐标。由于这种方法能充分理解关节点定位的失败过程，网络输出多个空间位置的置信度，而不是只考虑单个位置的置信度，使得学习更加立体。因此主流的姿态估计都是采用基于热图检测的方法的。

3　多人姿态评估方法

相比于单人姿态估计只需对被检测图像中的单个人体进行检测，多人姿态估计需要检测出图像中的所有人体，并且存在人与人之间彼此遮挡、截断等问题，同时被检测出来的所有关节均需要与正确的人体相关联。多人姿态估计一般有两种方法：一种是自上而下的方法[35-37]，即首先使用人体检测器检测出图像中的所有单个人体，然后对检测出的单个人体分别进行关节点检测，进而估计出每个人的姿态；另外一种是自下而上的方法[38-40]，即先检测出图像中所有人体关节点，根据人体结构先验对检测出的关节点进行重新组合，使得人体与自身关节点正确匹配，进而实现对每个人的姿态估计。通常情况下，自上而下方法只需对图像中被检测出的单个人体进行姿态估计，不需要区分身体部件和解决关节点在全图中的归属问题，相比较自下而上的方法，更加容易实现。但是，正因为自上而下方法的这一特性，当图像中人体出现相互遮挡时，更容易出现重复检测和错误估计等问题。

3.1　自上而下的评估方法

对于多人图像中存在的人体相互遮挡或被截断问题，一般方法分为人体检测和姿态估计两个阶段来解决。此类方法依赖于简单的人体几何部位关系，强调个体间的分割与检测。比如文献[38，41-43]都是首先检测出单个人体，然后再独立估计出单人的姿态。Iqbal 等[44]提出了一种进行多人姿态估计的新思路，将多人姿态估计视为人体局部关联问题，通过使用人体检测器检测出图像中的人体，生成一组联合候选框。对于每个检测到的人体，使用卷积姿态机(CPM)分别对每个候选框进行单个人体姿态的估计。但是在单独对每个候选框进行估计时，都不会考虑人体之间相互遮挡或截断的问题，这使得估计的关节点不能与正确的人体相关联，从而导致错误的估计。因此，为了将每个关节与正确的人体相关联并且移除错误的候选框，该方法使用整数线性规划(ILP)在每个人的完全连通图上进行推断，以得出最终的姿态估计结果，具体方法过程如图 10 所示。给定图像 $\boldsymbol{I}$ 中的人，定义其姿态为一个集合 $\chi=\{\boldsymbol{X}_j\}_{j=1,\cdots,J}$，其中 J 为 14 个人体关节，用向量 $\boldsymbol{X}_j\in\chi$ 表示图像中第 j 个 a 关节的位置(u,v)。网络中的 CPM 由多阶段 CNN 架构，每个阶段的多标签分类器 $\phi_t(\boldsymbol{X})$为每个关节 j 提供置信度 $\boldsymbol{S}_t^j\in\mathbf{R}^{w\times h}$，其中 w 和 h 分别是图像的宽度和高度，t 表示第 t 个阶段。网络第一阶段使用局部图像特征来提供置信度分数：

$$\varphi_t=1(\boldsymbol{X}\,|\,\boldsymbol{I})\rightarrow\{s_1^j(\boldsymbol{X}_j=\boldsymbol{X})\}_{j=1,\cdots,J+1}。\quad(2)$$

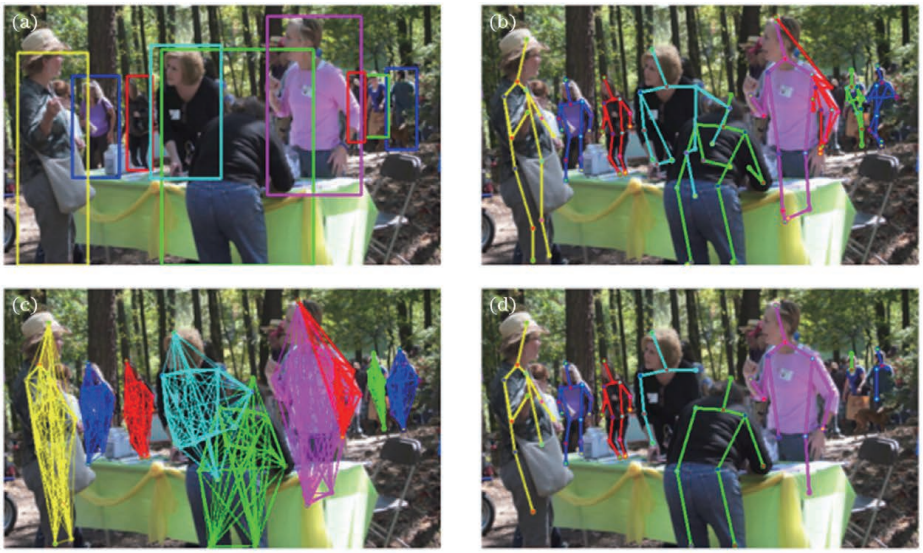

图 10　联合人体关联具体过程[44]

Fig. 10　Specific process of joint human body association[44]

所有后续阶段都利用前一阶段的上下文信息产生新的置信度得分：

$$\varphi_t > 1[\boldsymbol{X}\,|\,\boldsymbol{I},\psi(\boldsymbol{X},\boldsymbol{S}_{t-1})] \rightarrow \{s_t^j(\boldsymbol{X}_j=\boldsymbol{X})\}_{j=1,\cdots,J+1}, \tag{3}$$

式中：$\boldsymbol{S}_t \in \mathbf{R}^{w\times h\times(J+1)}$ 对应于所有身体关节的置信度得分图和阶段 t 的背景；$\psi(\boldsymbol{X},\boldsymbol{S}_{t-1})$ 表示从置信度图 $\boldsymbol{S}_{t-1}$ 到位置 x 的特征映射。由于假设所有关节点在图像中都是可见的，忽略了隐形关节点，可能错误地将其他人的关节与当前人体相关联。因此文献[44]不采取置信度分数的最大值，而是从每个推断得分图 $\boldsymbol{S}_t^j$ 中抽取 N 个候选者，通过 ILP 解决人体关节之间的关联以及错误连接的去除问题。在 MPII 数据集上，该方法的精准度比 DeepCut[38] 提高了 0.8%，并且运算时间从 57995 s 提升到了 10 s。为了评估人体检测器的检测精度对姿态估计的影响，当给出检测人体的真实位置时，精确度从 49.3%显着提高到了 76.9%，表明更好的人体检测器将进一步改善人体姿态估计结果。

Papandreou 等[45]在没有提供人体的真实位置或比例的情况下，重新审视了自上而下的方法，并表明它可以有效解决更具挑战性的基于“野外”场景下的姿势估计问题。该方法分为两个阶段：在第一阶段，使用 Faster-RCNN 检测出图像中的多个人体，由于检测出的人体边界框大小不同，在保持人体纵横比为 1.37 的情况下对每个人体边界框进行裁剪，得到的裁剪图像大小为 353×257；在第二阶段，采用全卷积残差网络(fully convolutional resnet)预测每个人体边界框中人体的热图和弥补图(offset)，通过热图和弥补图的融合得到关节点的精确定位。其中弥补图代表热图的每个像素位置到正确关节点位置的偏移(每个弥补图对应一个通道，分别表示关节点的 x 和 y 坐标)，然后对三个通道进行融合，从预测的偏移量中投票出真实的关节点位置。融合方式为

$$f_k(x_i)=\sum_j \frac{1}{\pi R^2}G[x_j+F_k(x_j)-x_i]h_k(x_j), \tag{4}$$

式中：$G(\cdot)$ 为双线性内插算法；h_k 是输出的热图通道。计算出热图上所有位置相对于 x_i 的距离，再除去一个 πR^2，之所以除去 πR^2 是假设网络的预测特别准，那么除了在以真正关节点为圆心，半径为 R 的圆内应该有值的话，其余地方都是 0，所以才只除去 πR^2 而不是除去整个热图的大小，不然会分散真正有值的地方的值。该过程先预测单独的热图和弥补图，并将二者融合，产生高度本地化的激活图，从而可以提高关节点定位的精准度。作者在基于 ResNet101 的 Faster-RCNN 上利用空洞卷积(atrous convolution)对卷积操作进行了修改。空洞卷积的卷积核是稀疏的，其有效性基于如下假设：紧密相邻的像素点的属性几乎相同，将其全部纳入计算则过于冗余，不如跳 H(空洞尺寸)个像素取一个像素进行卷积计算，如图 11 所示，为普通卷积到空洞卷积的变化过程。空洞卷积减小了卷积核的大小，起到节省内存和缩短运算时间的作用。为了重点解决关节点被遮挡以及复杂背景下多人姿态估计问题，Chen等[46]提出了一种新型级联金字塔网络

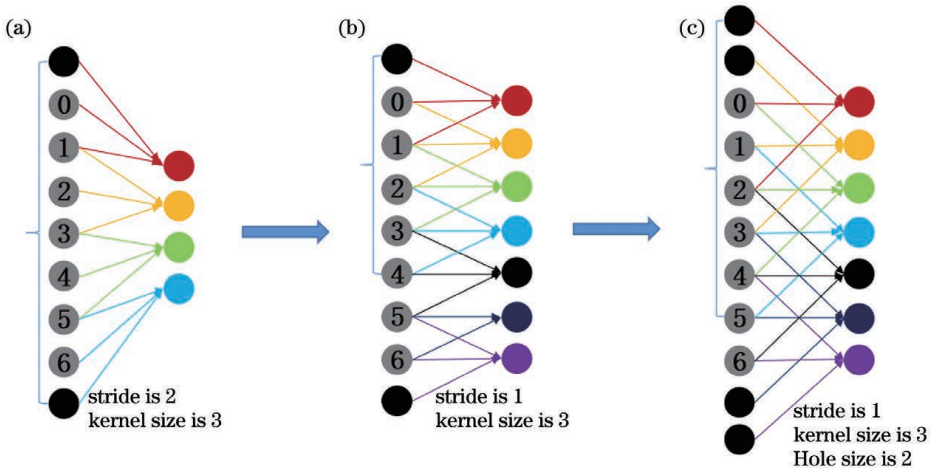

图 11　空洞卷积示意图

Fig. 11　Schematic of dilated convolution

(CPN)，整体框架采用从上而下的检测策略，首先使用人体检测框架得出人体候选框，然后使用 CPN 对每一个人体候选框中的人体关节点进行回归，进而得出人体姿态。网络整体结构如图 12 所示，CPN 包含两个子网络：全球金字塔网络(GlobalNet)和金字塔精炼网络(RefineNet)。GlobalNet 是一个功能金字塔网络，可以定位“简单”关节点，如眼睛和手，但无法精确识别被遮挡或不可见的关节点。它采用残差网络作为基础网络提取特征，不同的卷积特征的最后一个残差模块分别为 4 个不同分辨率的特征图 C2，C3，C4 和 C5。其中 C2 和 C3 的特征图分辨率较大，但语义信息较少，而C4和C5的特征图语义信息较多，但分辨率较低。因此通过不同倍率的上采样对底层特征与其上一层特征进行相加，实现不同尺度特征的融合，再对融合后的四层特征进行通道转换并进行上采样，一方面用于计算该阶段的 L2 loss，另一方面作为 RefineNet 的输入。一些被遮挡或不可见的关节点在 GlobalNet 中的输出误差会比较大，因此需要通过 RefineNet 将 GlobalNet 得到的所有层次的特征融合到一起，以实现对 GlobalNet 的预测结果进行修正的目的。文献[43]中探讨了可能影响多人姿态估计性能的各种因素。在 RefineNet 中选取 M 个 loss 反馈至 GlobalNet 中进行训练，从图13(a)可以看出，不同的M值下，

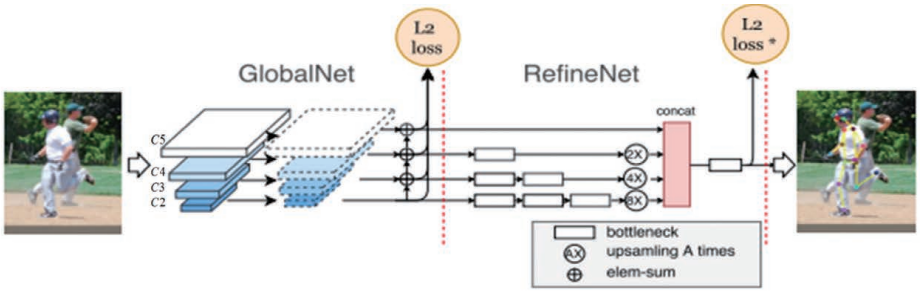

图 12　CPN 整体结构[46]

Fig. 12　CPN overall structure[46]

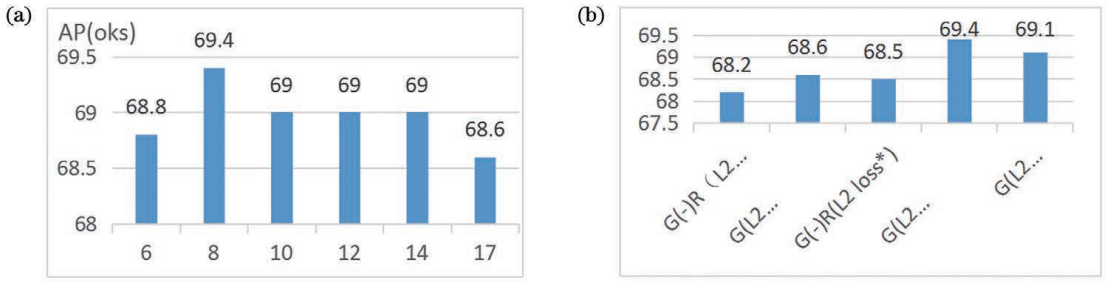

图 13　影响多人姿态估计性能的各种因素实验结果对比[46]

Fig. 13　Comparison of experimental results of various factors affecting the performance of multi-person pose estimation[46]

会得到不同的精准度 AP 值；从图 13(b)可以看出，在 CPN 中，采用不同的 loss 形式会带来不同的检测效果。当 GlobalNet 中采用 L2 loss 时，RefineNet 中采用 L2 loss *，精准度最高。

虽然以上方法有较高的精准度，但是前提是要有正确的人体检测。Fang 等[47]使用基于 Faster-RCNN 的人体检测器和沙漏堆栈网络模型[31]对这些问题进行解释并加以优化。因为单个人的姿态估计对被检测出的人体框位置很敏感，无论被检测出的人体框是否正确，都会对每个人体框进行姿态估计，这种重复检测产生了冗余姿态。图 14 分别说明了边界框定位错误问题和冗余检测问题。在图 14(a)中，虚线框是真实边界框，实线框是交并比(IoU)大于 0.5 的边界框，即使实线框被视为“正确”检测，也未能得到人体姿态。如图 14(b)所示，对于每个检测到的人体边界框，无论正确与否，都会得到一种人体姿态估计，因此会出现单个人体存在多个姿态的问题。为了解决以上问题，文献[47]提出了一种区域多人姿势估计(RMPE)框架。该框架由三部分组成，包括对称空间变换网络(SSTN)、参数姿势非最大抑制(NMS)和姿态引导建议生成器(PGPG)。将设计的 SSTN 结合 SPPE，从不准确的边界框中提取高质量的单人区域，引入 NMS 来比较姿态之间的相似性，以消除冗余姿势。通过 PGPG，学习人体检测器对不同姿态的输出分布，模拟人体边界框的生成，产生大量的训练样本数据，以实现增强训练样本的目的。虽然该方法针对单个人体的姿态估计，但是在多人姿态估计中也显示出不错的效果，在 MPII(多人)数据集上实现了 76.7% 的平均精度均值(mAP)。

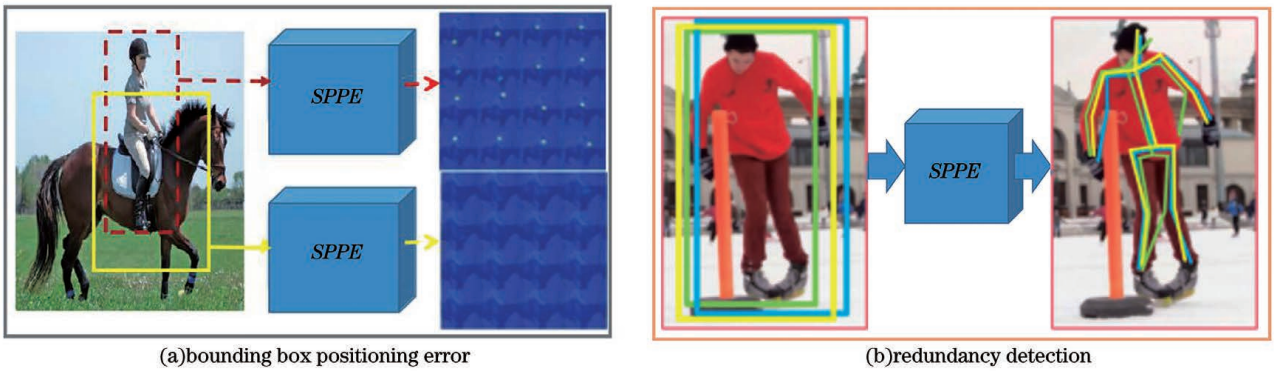

图 14　人体检测问题[47]

Fig. 14　Human detection problem[47]

但是随着检测人数的增多，文献[47]中的方法所花费的计算时间也变得更长，此类问题也是自上而下方法的通病，对于每次检测，都要运行单人姿势估计器，其中检测的人数越多，计算成本就越高。

3.2　自下而上的评估方法

尽管近年来检测器精度得到了很大的提升，检测错误的发生还是不可避免的，比如边界框定位错误以及检测冗余等，会严重地阻碍自上而下方法精度的提升。与自上而下的方法不同，自下而上方法摆脱了自上而下方法对个体进行精确检测这一要求的依赖。首先检测出图像中所有的人体关节点，对这些关节点进行聚类，使之与正确的人体匹配，最终实现人体的姿态估计。Pishchulin 等[38]采用自下而上方法解决图像中多人人体姿态估计问题，提出了联合子集划分和标记(SPLP)模型。与自上而下先检测出人体再进行姿态估计的两阶段方法不同，SPLP 模型可以同时检测出图像中的人体数量和姿态，并对人体遮挡情景具有良好的适应性。SPLP 模型使用两个网络分支。其一，使用 CNN 提取身体部位候选区，将每一个候选区域所对应的一个关节点作为一个节点，所有的这些候选节点组成一幅完整的节点图，如图 15(a)所示。以各节点之间的关联性作为权重，将关节之间的关联问题看作是一个整数线性规划问题，通过数学方法求解正确的连

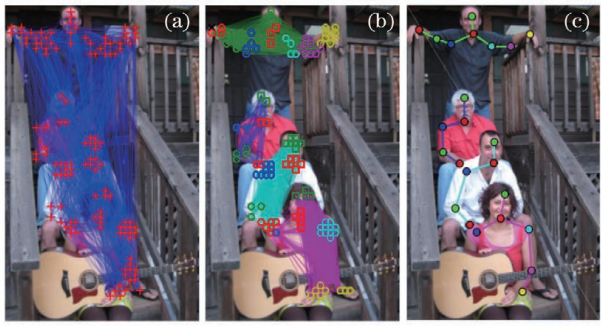

图 15　SPLP 网络姿态估计过程[38]

Fig. 15　Pose estimation process of SPLP network[38]

接关系，将属于同一个人的关节点归为一类，则每一个人成为一个单独的类。其二，对检测出来的节点进行标记，确定节点属于人体的哪一个部分，如图 15(b)所示。最后，通过网络对属于同一人体的关节点和各关节点的标记类型进行结合，如图 15(c)所示。在适用于多人姿态估计的 WAF 数据集上，SPLP 模型的精准度比传统的两阶段方法提高了 30%。但是，所采取的改进 Fast R-CNN 结合 ILP 的方法增加了算法复杂度，运算时间达到了 57995 s。针对这一问题，Insafutdinov 等[39]在文献[38]工作的基础上，提出了 DeeperCut 算法，最主要的改进为：1)使用残差网络(ResNet-50、ResNet-101 和 ResNet-152)[31]进行身体部件的提取；2)使用图像条件成对术语(ICPT)方法将得到的丰富的候选区域节点压缩至一定数量，以减少全连接网络的运算时间。在 MPII 数据集上，DeeperCut 的精准度相较于文献[38]中的方法提升了 12.1%，运行时间减少到 230 s。但是从时间效率上看，还是不令人满意。因此，Cao 等[48]基于卷积姿态机(CPM)提出了部件亲和域场(PAFs)，在保证关节点检测精度的前提下进一步缩短了运行时间。先利用 VGG-19 网络的前 10 层对原始图像进行处理得到特征图，通过改进的 CPM 网络对特征图进行多阶段处理，每个阶段的输出分别为部件置信度图(PCM)和 PAFs。其中，PCM 关注人体关节点，例如头、肩膀等；而 PAFs 关注肢体段，例如大臂、大腿等。双分支多级 CNN 的网络结构如图 16 所示，将得到的图像特征图 F 作为 Stage 1 的输入，输出原始图像对应的部件置信度图 PCM 和 PAFs；将 Stage 1 的输出和原始图像的特征图 F 作为 Stage 2 的输入，输出对应的 PCM 和 PAFs；Stage 3 及其后面各个阶段的网络结构及功能和 Stage 2 相似，这样反复迭代，直到网络收敛。图 17(a)为输入图像，图 17(b)为检测的关节点的置信度图，图 17(c)代表人体躯干的 PAFs。在 PAFs 的帮助下，把 PCM 检测出的关节点坐标连接起来，实现一段躯干与两端关节的快速匹配，从而形成人体姿态骨架。为了防止训练时出现梯度消失的问题，在每个阶段中间加入了中层

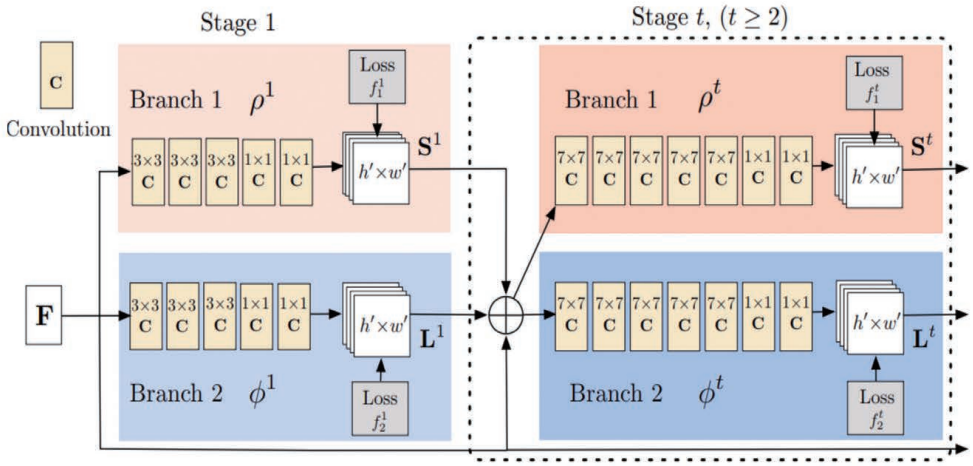

图 16　双分支多级 CNN 体系结构[48]

Fig. 16　Two-branch multi-stage CNN architecture[48]

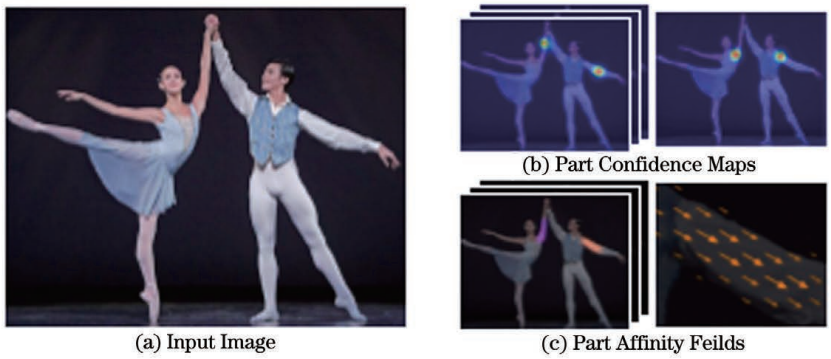

图 17　PAFs 方法示意图[48]

Fig. 17　Schematic of PAFs method[48]

损失以加强反向传播。在 MPII 数据集测试中，该方法的 mPA 比 DeeperCut 提升了 13%。相比于自上而下方法，自下而上方法具有更好的稳健性，更适合解决图像中复杂场景下的人体姿态估计问题。

4　方法技术指标及对比分析

4.1　评价指标

1）PCK

关节点正确估计的比例(PCK)是指检测关节点与其相对应的真实标注数据(ground truth)间的归一化距离小于设定阈值的比例。

2）OKS

目标关节点相似度(OKS)在关节点检测中用来代替 IoU，计算关节点位置的加权欧氏距离，对检测关节点与其对应的真实标注数据间的相似性进行评估，具体评估方式为

$$O_{\mathrm{OKS},p}=\frac{\sum_i \exp\{-d_{pi}^2/2s_p^2\delta_i^2\}\delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)},\quad(5)$$

式中：p 为检测人体框编号；i 为数据集标注时各关节点编号；d_{pi} 为测试时检测关节点位置与数据标注关节点位置的欧氏距离；s_p 为检测人体框面积的平方根(人体尺度因子)；δ_i 为骨骼关节点归一化因子(人工标注关节点位置偏移的标准差)；v_{pi} 指编号为 p 的人的第 i 个关节点的状态(可见、不可见或者不在框内等)；$\delta(v_{pi}=1)$ 为克罗内克函数。(5)式中只将可见关节点($v=1$)计入对姿态评估的评价指标。

3）AP

在整个测试数据集上，每一关节点检测结果的平均准确率(AP)为

$$A_{\mathrm{P}}=\frac{\sum_p \delta(O_{\mathrm{OKS}}>s)}{\sum p_1},\quad(6)$$

式中：s 为 OKS 阈值。

4）mAP

mAP 是指在不同阈值下，将每一关节点的检测 AP 都单独“拎”出来，再计算所有关节点 AP 的均值[49]。计算方式为

$$P_{\mathrm{mAP}}=\mathrm{mean}\{A_{\mathrm{P}}@s(0.50:0.05:0.95)\},\quad(7)$$

在给定阈值 s 下，AP 为每一关节点在整个测试数据集上检测结果的平均准确率。

5）IoU

IoU 是一个作为目标检测算法性能 mAP 计算的非常重要的函数。IoU 表示产生的候选框(candidate bound)与原标记框(ground truth bound)的交叠率，也就是它们的交集与并集的比值。

4.2　对比分析

在人体姿态估计中，由于人体比较灵活，人体关节点通常存在可见、不可见、不在图内(不可推测)3 种状态，还存在不同部位的关节点的检测难易程度差异、部位相似使得关节点区分性较差等问题，数据图像中关节点选择数量和关节点位置对模型训练效果有直接的影响。图 18、图 19 及图 20 分别对比

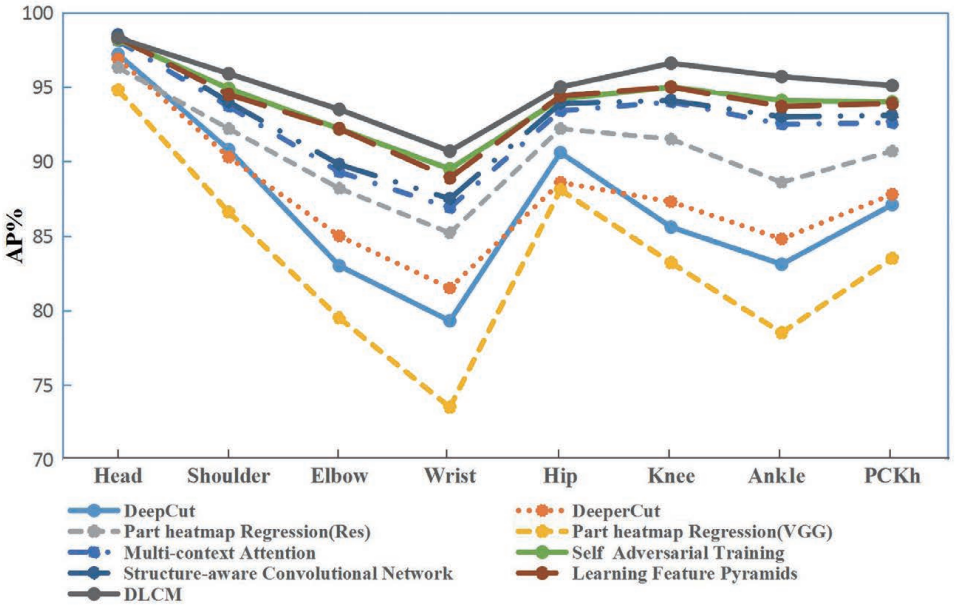

图 18　单人姿态估计实验对比结果(LSP)

Fig. 18　Comparison results of single person pose estimation experiments (LSP)

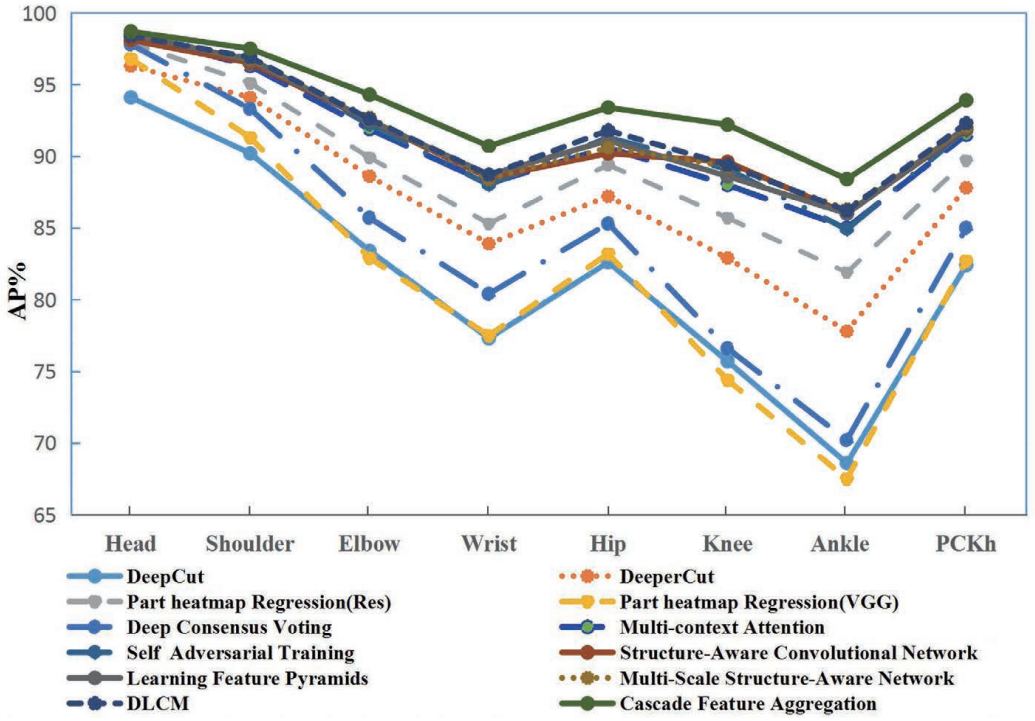

图 19　单人姿态估计实验对比结果(MPII)

Fig. 19　Comparison results of single person pose estimation experiments (MPII)

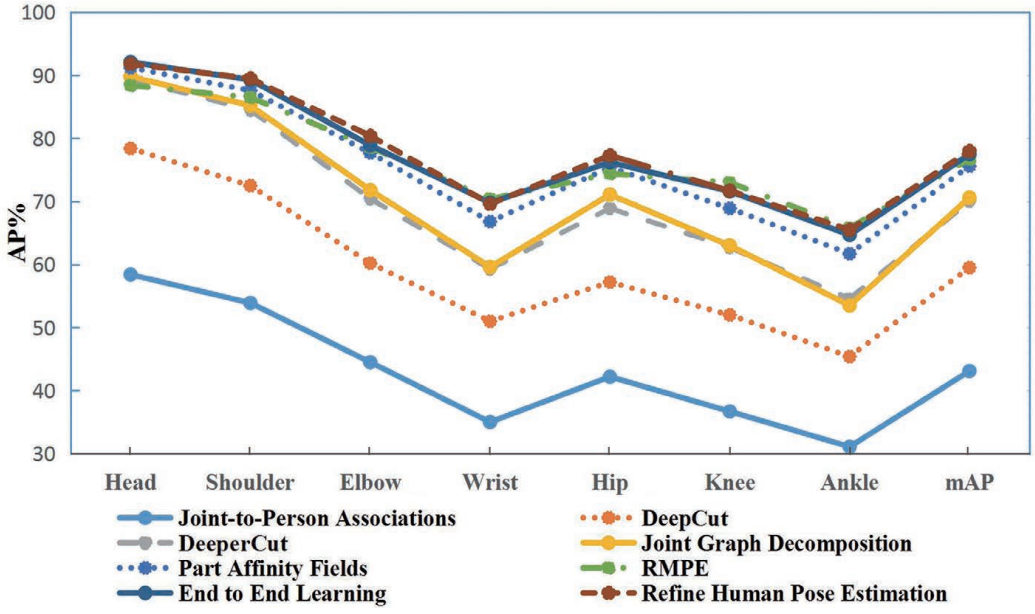

图 20　多人姿态估计实验对比结果(MPII)

Fig. 20　Comparison results of multi-person pose estimation experiments (MPII)

了不同姿态估计方法在 MPII 和 LSP 数据集上的实验结果。在图 18 中，从 Part heatmap Regression 方法[19]的实验结果可以看出，采用残差网络进行回归的方法的整体精准度要比采用 VGG 网络架构进行回归的方法高出 7%，相比于 VGG，残差网络可以进一步减小网络加深过程中出现的梯度消失和梯度爆炸等问题所带来的损失。相比于 DeepCut 方法[38]，DeeperCut 方法[39]通过残差网络提取身体部件，并采用图像条件成对术语方法将候选区域大量的节点数压缩到较少数量，使得网络运算时间从 57995 s 降低到了 230 s，同时网络精准度也提升了 0.7%，对数据量要求降低的同时提高精准度。在图 19 中，各种姿态估计方法的检测精度在典型关节点 Head 处达到了 93%以上，但在非典型关节点 Ankle

处的检测精度却都处在 90%之下，其中 DeepCut[38] 和 Part heatmap Regression[19] 方法在 Ankle 处的检测精度甚至低于 70%，说明各种姿态估计方法在典型关节点和非典型关节点的检测精度上存在较大差距。虽然 Cascade Feature Aggregation 方法[50] 对典型关节点的检测精度与其他方法相比并没有展现出较大的优势，但是对于非典型关节点的检测效果要比其他对比方法更加突出，甚至在 Ankle 上的检测精度比文献[19]高出 20%，对于非典型关节点的正确检测将更有助于提高人体姿态估计的整体检测精度。从图 20 可以看出，相比于单人人体姿态估计方法在各个关节点上的平均检测精度，多人人体姿态估计各种方法的平均检测精度普遍较低，特别是对于非典型关节点的检测精度，二者平均相差 20%。由此也可以看出，图像中需要检测的关节点数量和各关节点之间位置关联关系的复杂性会增加检测难度。相较于关节点标注个数为 14 的 LSP 数据集，MPII 数据集增加了 Pelvis（骨盆）和 Thorax（胸部）两个难以检测的关节点，通过图 18 和图 19 的对比可以看出，相同人体姿态估计方法在 LSP 数据集上的检测精度要略高于在 MPII 数据集上的检测精度，由此可以看出关节点位置对模型训练效果有直接的影响，并且越难区分的关节点，检测精度越低。

5　未来工作展望

根据近年来的相关研究以及各研究之间的关联性，综述了各类姿态估计方法，总结了从各个方面提升性能和精度的相关模型，具体包括基于堆栈沙漏网络、注意力机制、残差网络（ResNet）和生成对抗网络的姿态估计模型。在单人姿态估计的基础上，从自上而下和自下而上两个方法角度总结了多人姿态估计算法，对多人人体姿态估计方法中待检测人数与模型检测精度之间的权衡关系进行了分析，并着重介绍了局部区域重叠、关节点混淆、人体非典型部位关节点难以检测等问题的解决方案，以及在自下而上方法中聚类方法对关节点检测的贡献。由于人体姿态估计作为人体行为识别、行人重识别等机器视觉问题研究的基础性方法，相关研究在受到广泛关注的同时，还存在很多瓶颈性制约因素，限制了其发展。最后对人体姿态估计未来的发展趋势进行了展望。

1）缺乏与实际场景更匹配的训练数据

虽然目前有多个人体姿态估计数据集可供使用，比如 COCO 数据集、MPII 数据集和 FLIC 数据集等，但是这些数据集中大都是一些如站立、静坐等简单姿态图像，缺乏摔倒、翻越等复杂姿态。随着深度学习在解决人体姿态估计问题中的广泛应用，利用这些数据集进行训练的人体姿态估计模型对于简单人体姿态的估计水平已经到达较高的水平。比如，Su 等[50] 通过级联特征聚合方法，在 MPII 数据集上将单人姿态估计的检测精度提升至 93.9%。但是，由于缺乏日常生活中的复杂姿态数据，模型难以适应如智能安防、智能交通以及智慧城市等环境下所要求的实际场景。研究如何在实际场景中有效估计人体复杂姿态具有很好的实用价值。

2）人体检测器存在缺陷

由于存在人体尺度不同、相互遮挡以及背景重叠等问题，容易出现边界框定位错误和检测冗余现象。姿态估计算法对被检测出的人体框位置很敏感，无论被检测出的人体框是否正确，都会对每个框中的人体进行姿态估计，而重复的检测会产生冗余姿态，错误检测会降低姿态估计模型的性能。

3）算法的时效性不高

虽然基于深度学习的人体姿态估计方法的检测精度不断提高，但是这些方法都要构建复杂的网络结构，训练相关网络时需要花费大量时间。对于自上而下方法，由于对检测到的每一个人体都需要进行一次姿态估计，因此随着检测人数的增加所花费的计算时间也变得更长。虽然自下而上检测方法不需要对图像中的人体依次进行检测，但是当拥挤场景下各类需要检测的关节点数目呈爆炸式增长时，算法的时效性也会明显下降。为了提高模型的时效性，研究人员通过减小网络规模设计了一些轻量级网络，但是牺牲了一定的精度。如何保证在高精度检测的前提下提高时效性是一个值得重视的问题。

4）视觉角度和环境的影响

基于目前的研究方法可以看出，无论是在非正常角度如俯视或者侧视角度下，还是当受到光照变化、遮挡等环境因素影响时，人体关节点检测过程中都会存在关节点误判或遗漏等问题，使得人体姿态的识别率明显偏低。

基于对以上限制因素的分析，考虑现实应用需求，未来的工作可从如下几方面开展。

1）针对非典型关节点的检测

无论是自上而下还是自下而上的人体姿态估计方法，非典型的人体关节点的检测精准度普遍低于典型人体关节点的精准度，如图 18 所示，无论哪种

方法，对头、肩膀的检测精度都高于对手腕、脚踝的精度。人体姿态估计方法的精度可以通过对非典型关节点的准确检测得到进一步提升[50]。例如，可以通过引入注意力机制获取人体中需要重点关注的目标关节区域，之后对这一区域投入更多注意力资源，以获取更多需要关注的非典型关节点的细节信息，在提高非典型关节点的检测精度的同时也提高了人体姿态估计的整体精度。如何通过准确估计非典型的人体关节点来进一步提升人体姿态估计精度是一个重要的研究方向。

2）提高对复杂场景的适应性

人体姿态估计方法发展至今，无论是单人姿态估计还是多人姿态估计，无论是 2D 姿态估计还是 3D 姿态估计，能够有效检测的人数一般不会太多，主要是缺乏针对拥挤场景高效的人体检测算法。当图像中人数增多时，会出现人与人之间相互遮挡、相互截断以及复杂背景干扰，这将导致检测器出现漏检、误检和多检等问题，将会影响后续的姿态估计结果。未来可以利用多视角信息融合方法，通过多路匹配算法[51]对多视角中二维位姿进行聚类，然后利用不同视角下的互补信息来解决复杂场景下人体相互遮挡、相互截断等问题。从实际应用的角度来看，人体姿态估计不能只局限于简单的场景，如何提高对复杂场景的适应性已经是迫在眉睫的问题。

3）简化运算参数

随着人体姿态估计在实际中的逐步应用，降低运算成本开始成为科研工作者努力的方向之一。卷积神经网络是人体姿态估计工作的核心组件，绝大多数计算量集中在卷积操作上，因此高效的卷积层设计是降低网络复杂度的关键。虽然关于卷积模型的研究层出不穷，产生了如 VGG[19]、ResNet[21]、Xception[52]和 ResNeXt[53]等性能优异的网络结构，并在多个视觉任务上超过了人类水平，然而，这些成功的模型往往伴随着巨大的计算复杂度，模型的训练无法在移动端进行，只能依赖高性能的服务器集群。本着在保证一定的识别精度条件下，尽可能减少网络规模的思想，一些轻量级卷积网络结构被提出。例如，Zhang 等[54]提出的 ShuffleNet 是一种针对移动端低功耗设备的更为高效的卷积模型结构，其在 ResNet 的设计思想基础上做出了一系列改进来提升模型的效率，在大幅降低模型计算复杂度的同时仍然保持了较高的识别精度，使得网络可以在 CPU 上高效运行。Osokin[55]在 OpenPose 算法的基础上，将 VGG[19]替换为 MobileNet[56]网络结构，实现了可以在 CPU 上实时运行的人体姿态估计。可以考虑利用一些包括蒸馏法[57]、分组卷积[58]以及逐点卷积[52]等轻量化网络算法来简化网络结构，将这些轻量级网络结构用于人体姿态估计，将进一步降低算法的落地门槛。

4）基于视频的人体姿态估计

目前主流的人体姿态估计还是通过单帧图像特征提取到人体各个关节点信息完成人体姿态估计。在现实需求中，从提升精度和鲁棒性的角度考虑，在连续多帧图像序列中识别人体姿态将更有效。基于视频的人体姿态估计的核心环节是考虑帧与帧之间的关联性，利用人体连续动作或人体持续姿态等特点，可以将一些失真姿态恢复成真实姿态，能够解决单帧图像中存在的遮挡、背景干扰和比例失调等所带来的模型精度下降问题。例如，通过光流法来估计视频中人体姿态，使用光流在连续的多个帧中组合信息，通过连续多帧图像的上下文信息还原当前人体姿态。

5）模型向半监督、甚至无监督模式发展

现有主流的人体姿态估计模型需要大量的标注好的训练数据。手工标注是一项非常耗时耗力的工作，尤其针对基于视频的人体姿态数据集，现有的做法是对每一帧图像都进行标注，整个过程不仅耗时耗力，而且人工标注的数据存在一定的局限性，降低了模型的泛化能力。可以采用稀疏注释[59]的方法对视频序列中每隔 k 个帧进行一次标注，利用光流法对视频序列中标注的帧和未标记的帧的特征进行翘曲，通过标记帧的人体姿态预测结果来反向优化未标记帧的预测结果。还可以在无标签数据中加入少量标注好的样本，同一簇中的无标签数据与有标签的数据共用同一标签，在减少标注数据工作量的条件下明显提升模型的泛化能力。甚至可以采取无监督训练方法，采用无标签训练数据，通过聚类算法训练模型，省略对数据标注这一工作的同时提升模型对场景的适应性。如何基于弱监督、半监督、甚至无监督模式形成高精度的人体姿态估计模型将会成为未来的重点方向。

参　考　文　献

[1] Wang C Y, Wang Y Z, Yuille A L. An approach to pose-based action recognition[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 915-922.

[2] Liang Z J, Wang X L, Huang R, et al. An

expressive deep model for human action parsing from a single image[C]//2014 IEEE International Conference on Multimedia and Expo (ICME), July 14-18, 2014, Chengdu, China. New York: IEEE Press, 2014.

[3] Cho N G, Yuille A L, Lee S W. Adaptive occlusion state estimation for human pose tracking under self-occlusions[J]. Pattern Recognition, 2013, 46(3): 649-661.

[4] Nie B X, Xiong C M, Zhu S C. Joint action recognition and pose estimation from video[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1293-1301.

[5] Huang Y W, Zhao P, You Y D. Pose-guided human image synthesis based on fusion feature feedback mechanism[J]. Laser & Optoelectronics Progress, 2020, 57(14): 141011.
黄友文, 赵朋, 游亚东. 融合反馈机制的姿态引导人物图像生成[J]. 激光与光电子学进展, 2020, 57(14): 141011.

[6] Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June 20-25, 2011, Colorado Springs, CO, USA. New York: IEEE Press, 2011: 1297-1304.

[7] Ionescu C, Li F X, Sminchisescu C. Latent structured models for human pose estimation[C]// 2011 International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE Press, 2011: 2220-2227.

[8] LeCun Y, Ranzato M. Deep learning[M]. Cambridge: Cambridge University Press, 2011.

[9] Pishchulin L, Andriluka M, Gehler P, et al. Poselet conditioned pictorial structures[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 588-595.

[10] Lifshitz I, Fetaya E, Ullman S. Human pose estimation using deep consensus voting[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9906: 246-260.

[11] Ke L P, Chang M C, Qi H G, et al. Multi-scale structure-aware network for human pose estimation [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11206: 731-746.

[12] Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation[C]// Procedings of the British Machine Vision Conference 2010, August 31-September 3, 2010, Aberystwyth. British: British Machine Vision Association, 2010.

[13] Chen X J, Yuille A L. Articulated pose estimation by a graphical model with image dependent pairwise relations[C]//Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13, 2014, Montreal, Quebec, Canada. New York: Curran Associates, 2014: 1736-1744.

[14] Varamesh A, Tuytelaars T. Mixture dense regression for object detection and human pose estimation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 13086-13095.

[15] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 1653-1660.

[16] Pfister T, Simonyan K, Charles J, et al. Deep convolutional neural networks for efficient pose estimation in gesture videos[M]//Cremers D, Reid I, Saito H, et al. Computer vision-ACCV 2014. Lecture notes in computer science. Cham: Springer, 2015, 9003: 538-552.

[17] Fan X C, Zheng K, Lin Y W, et al. Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1347-1355.

[18] Pfister T, Charles J, Zisserman A. Flowing ConvNets for human pose estimation in videos[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1913-1921.

[19] Bulat A, Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9911: 717-732.

[20] Zhang N, Shelhamer E, Gao Y, et al. Fine-grained pose prediction, normalization, and recognition[EB/OL]. (2015-11-22)[2020-11-10]. https://arxiv.org/abs/1511.07063.

[21] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[22] Chu X, Yang W, Ouyang W L, et al. Multi-context attention for human pose estimation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5669-5678.

[23] Artacho B, Savakis A. UniPose: unified human pose estimation in single images and videos[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 7033-7042.

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

[25] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[26] Chou C J, Chien J T, Chen H T. Self adversarial training for human pose estimation[C]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), November 12-15, 2018, Honolulu, HI, USA. New York: IEEE Press, 2018: 17-30.

[27] Chen Y, Shen C H, Wei X S, et al. Adversarial PoseNet: a structure-aware convolutional network for human pose estimation[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1221-1230.

[28] Yang W, Li S, Ouyang W L, et al. Learning feature pyramids for human pose estimation[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1290-1299.

[29] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]// Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 483-499.

[30] Tian Y D, Zitnick C L, Narasimhan S G. Exploring the spatial hierarchy of mixture models for human pose estimation[M]// Fitzgibbon A, Lazebnik S, Perona P, et al. Computer vision-ECCV 2012. Lecture notes in computer science. Heidelberg: Springer, 2012, 7576: 256-269.

[31] Rothrock B, Park S, Zhu S C. Integrating grammar and segmentation for human pose estimation[C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 3214-3221.

[32] Park S, Zhu S C. Attributed grammars for joint estimation of human attributes, part and pose[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 2372-2380.

[33] Tang W, Yu P, Wu Y. Deeply learned compositional models for human pose estimation[C]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11207: 197-214.

[34] Tompson J J, Jain A, LeCun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[C]// Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13, 2014, Montreal, Quebec, Canada. New York: Curran Associates, 2014: 1799-1807.

[35] Gkioxari G, Hariharan B, Girshick R, et al. Using k-poselets for detecting people and localizing their keypoints[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 3582-3589.

[36] Chen X J, Yuille A. Parsing occluded people by flexible compositions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3945-3945.

[37] Yan F T, Wang P, Lü Z G, et al. Real-time multi-person video-based pose estimation[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021006.
闫芬婷, 王鹏, 吕志刚, 等. 基于视频的实时多人姿态估计方法[J]. 激光与光电子学进展, 2020, 57(2): 021006.

[38] Pishchulin L, Insafutdinov E, Tang S Y, et al. DeepCut: joint subset partition and labeling for multi person pose estimation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4929-4937.

[39] Insafutdinov E, Pishchulin L, Andres B, et al.

DeeperCut: a deeper, stronger, and faster multi-person pose estimation model[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9910: 34-50.

[40] Cheng B W, Xiao B, Wang J D, et al. HigherHRNet: scale-aware representation learning for bottom-up human pose estimation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 5385-5394.

[41] Pishchulin L, Jain A, Andriluka M, et al. Articulated people detection and pose estimation: reshaping the future[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 3178-3185.

[42] Sun M, Savarese S. Articulated part-based model for joint object detection and pose estimation[C]//2011 International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE Press, 2011: 723-730.

[43] Huang J J, Zhu Z, Guo F, et al. The devil is in the details: delving into unbiased data processing for human pose estimation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 5699-5708.

[44] Iqbal U, Gall J. Multi-person pose estimation with local joint-to-person associations[M]//Hua G, Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 627-642.

[45] Papandreou G, Zhu T, Kanazawa N, et al. Towards accurate multi-person pose estimation in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3711-3719.

[46] Chen Y L, Wang Z C, Peng Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7103-7112.

[47] Fang H S, Xie S Q, Tai Y W, et al. RMPE: regional multi-person pose estimation[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2353-2362.

[48] Cao Z, Simon T, Wei S H, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1302-1310.

[49] Lin T Y, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[50] Su Z H, Ye M, Zhang G H, et al. Cascade feature aggregation for human pose estimation[EB/OL]. (2019-02-21)[2020-11-10]. https://arxiv.org/abs/1902.07837.

[51] Dong J T, Jiang W, Huang Q X, et al. Fast and robust multi-person 3D pose estimation from multiple views[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 7784-7793.

[52] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1800-1807.

[53] Xie S N, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5987-5995.

[54] Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 6848-6856.

[55] Osokin D. Real-time 2D multi-person pose estimation on CPU: lightweight OpenPose[C]//Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, February 19-21, 2019, Prague, Czech Republic. Southampton: Science and Technology Publications, 2019: 744-748.

[56] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2019-04-17)[2020-11-10]. https://arxiv.org/abs/1704.04861.

[57] Zhang F, Zhu X T, Ye M. Fast human pose

estimation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 3512-3521.

[58] Zhang T, Qi G J, Xiao B, et al. Interleaved group convolutions for deep neural networks[EB/OL]. (2017-07-10)[2020-11-10]. https://arxiv.org/abs/1707.02725.

[59] Bertasius G, Feichtenhofer C, Tran D, et al. Learning temporal pose estimation from sparsely-labeled videos[EB/OL]. (2019-06-06)[2020-11-10]. https://arxiv.org/abs/1906.04016.