计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(10):295−300 [doi: 10.15888/j.cnki.csa.008143]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于堆叠沙漏网络的人体姿态估计①

吴佳豪, 周　凤, 李亮亮

(贵州大学 计算机科学与技术学院, 贵阳 550025)
通讯作者: 周　凤, E-mail: 41544782@qq.com

摘　要: 人体姿态估计在许多计算机视觉任务中起着重要的作用, 然而, 由于姿态的多变、光照、遮挡和分辨率低等因素, 它仍然是一个具有挑战性的问题. 利用深层卷积神经网络的高级语义信息是提高人体姿态估计精度的有效途径, 本文提出了一种改进的堆叠沙漏网络, 设计了一个大感受野残差模块和预处理模块来更好地获得人体结构特征, 以此获得丰富的上下文信息, 对部分遮挡、大姿态变化、复杂背景等有较好的效果, 此外, 还对不同阶段的结果进行了融合, 以进一步提高定位精度, 在 MPII 数据集和 LSP 数据集上对本文提出的模型进行实验和验证, 结果证明了本文模型的有效性.

关键词: 人体姿态估计; 神经网络; 特征融合

引用格式: 吴佳豪,周凤,李亮亮.基于堆叠沙漏网络的人体姿态估计.计算机系统应用,2021,30(10):295–300. http://www.c-s-a.org.cn/1003-3254/8143.html

Human Pose Estimation Based on Stacked Hourglass Network

WU Jia-Hao, ZHOU Feng, LI Liang-Liang

(College of Computer Science and Technology, Guizhou University, Guiyang 550025, China)

Abstract: Human pose estimation plays an important role in many computer vision tasks. However, it remains challenging due to complex pose changes, illumination, occlusion, and low resolution. The high-level semantic information from deep convolutional neural networks provides an effective way to improve the accuracy of human pose estimation. In this study, an improved stacked hourglass network is proposed. A large-receptive-field residual module and a preprocessing module are designed to better outline structural features of a human body so that rich contextual information can be obtained. The network performs well in the presence of partial occlusion, large pose change, complex background, etc. In addition, the positioning accuracy is further enhanced by the fusion of results from different stages. Experiments on MPII data sets and LSP data sets prove the effectiveness of this model.

Key words: human pose estimation; neural network; feature fusion

人体姿态估计是从输入图像中确定人体关节的像素位置, 再将这些关键点按照人体关节的方式相连, 进而得到人体的姿态. 它是计算机视觉中一项基本而具有挑战性的任务, 是行为识别[1,2]、运动捕捉[3,4]、人机交互[5] 和视频监控[6] 等其他相关任务的研究基础. 遮挡点的可见性、视角等因素都会影响检测精度, 解决这些问题的关键是如何提取强大的低级和中级外观特征来捕获每个尺度上的相关上下文信息, 以及如何建立复杂的部件关系以实现有效的姿态推断.

传统的人体姿态估计方法基于图结构模型, 利用树结构来建立人体的空间关系[7]. 图形结构模型通过探索身体各部分与连接肢体的运动先验之间的空间相关

① 基金项目: 贵州省自然科学技术基金 (黔科合基础 [2019]1088)
Foundation item: Natural Science and Technology Fund of Guizhou Province ([2019]1088)
收稿时间: 2021-01-05; 修改时间: 2021-02-03; 采用时间: 2021-03-02

性, 构建了一个经典的树结构图形框架, 从而捕获各部件之间的关系. 最近, 利用深度神经网络[8–11] 取得了很大的进展, 其中, 沙漏网络[10] 可以很好地捕捉各种尺度的信息, 用于鲁棒的人体姿态估计, 此外, 几个沙漏网络被堆叠起来, 用于自上而下的推理, 并能够重新评估初始估计结果和特征. 与堆叠沙漏网络不同, CPN[12] 设计了另一种级联多个阶段的策略, 即 GlobalNet, GlobalNet 的目的是通过在上采样过程中的每个元素和过程之前应用 1×1 卷积核来获得与 FPN[13] 不同的初始人体姿态. 除了上述的自顶向下方法外, 像 OpenPose[14] 这样的自底向上方法也可以在实时人体姿态估计方面取得较好的结果.

本文提出并设计了一种有效的人体姿态估计方法. 在堆叠沙漏网络中使用两种不同的残差模块作为沙漏网络的基本构建块, 在分辨率较高时, 使用大感受野残差模块捕获人体结构信息, 当达到最低分辨率时为了更好的捕获深层信息, 因此仍使用原本的残差模块; 同时为了更好的获得结构信息和进行信息传递, 设计一个预处理模块提取包含丰富局部信息的低级特征, 并通过跳过连接将得到的结果传递给沙漏块. 将身体结构作为先验信息, 可以在可见的基础上为推断困难关键点的位置提供指导, 增加预处理模块来捕获包含丰富局部信息的低级特征, 可以有效地提取关于不同关键点之间相对位置的结构信息; 另一方面, 在重复的上采样和下采样操作中会逐渐丢失低级特征中的信息, 为了保证信息传递, 设计了一个分层连接的方式, 不仅可以补偿每个沙漏块的信息损失, 而且可以重用结构信息, 进一步提高局部特征表示的能力.

1 堆叠沙漏网络

1.1 沙漏网络

图 1 是一个单独的沙漏网络的模型图, 它由池层、上采样层和残差模块组成. 图片从 C1 输入, 逐步经过池化层后降低特征图的分辨率, 在抵达最低分辨率 C7 后, 网络开始逐步地对特征图进行上采样操作和跨尺度的特征融合, 沙漏网络的结构是对称的, 因此最后输出时达到与输入时相同的分辨率.

在人体姿态估计中局部特征对于关节点估计更为重要, 最终的姿态估计需要对整个身体的连贯理解. 在人体姿态估计中, 不同的特征图上对不同的关节点的识别精度也不同, 所以利用沙漏网络能更好地进行特征提取, 进行多尺度特征融合后达到网络的输出分辨率 C1b 之后, 输出最后的估计结果,得到一组热图, 是每个像素关节点存在的概率. 热图包括了所有的关节点, 也就意味着第 2 个沙漏网络可以利用前一个沙漏网络的输出结果, 即利用关节点间的相互关系, 从而提升了关节点的预测精度.

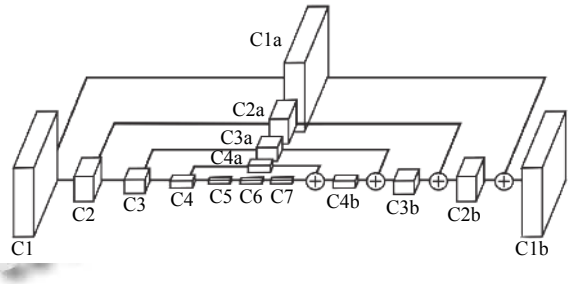

图 1 沙漏网络模型图

1.2 残差模块

残差模块[15] 是沙漏网络中的基本块, 其结构如图 2 中所示. 残差模块由两个分支组成: 第 1 个分支是标识映射分支, 第 2 个分支由卷积层、批归一化层和 ReLU 激活层组成. 使用残差模块不仅减少了网络的计算量, 同时解决了深层网络可能会出现的问题, 提高了网络的性能.

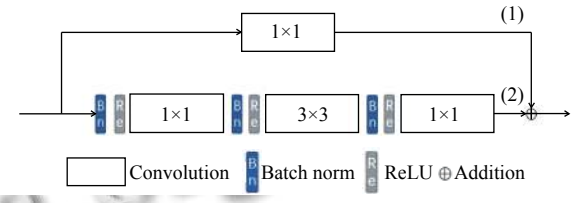

图 2 残差模块图

2 改进的堆叠沙漏

2.1 大感受野残差模块

在人体姿态估计中, 较大的感受野[16] 已被证明对局部身体部位的定位很重要. 原残差模块的感受野较小, 为了更好地获得局部信息, 进一步了解各关节之间的相关性, 对原残差模块进行改进, 并设计了一个大感受野残差模块. 图 3 是其原理图.

该模型主要由卷积层、Batch norm 层、ReLU 激活层、池化层、上采样层和 Spatial dropout 层组成. 该模块在原残差模块的基础上添加了一个分支, 即图中的分支 (3), 包括两个 3×3 卷积层、上下采样, 另外在该模块中引入了 Spatial dropout 层加入分支, 在模型中

有防止激活变强和防止过度拟合的作用, 提高了模型的精度. 这 3 个分支中通过分支 (1) 和分支 (2) 来获得高分辨率信息, 同时使用分支 (3) 增加特征提取时的感受野, 最后将特征融合后进行输出, 传递至下一模块. 但并不是所有的残差模块都需要大感受野, 在图 1 中的 C5、C6、C7 中仍然使用原残差模块, 获取深层的语义信息, 其余的残差模块则使用大感受野的残差模块, 能更好的获取结构信息. 这种卷积分支的加入可以有效地增加图像在特征提取时的感受野, 所以能在人体姿态估计中发挥相当好的作用.

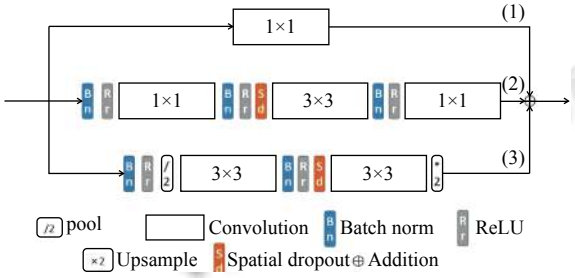

图 3 大感受野残差模块图

2.2 预处理模块

以往的研究表明, 高层特征包含语义等全局信息, 而局部信息往往存在于低层特征中, 设计合理的预处理模块可以有效地提取关节结构作为先验, 有助于更好地处理遮挡、大姿态变化、复杂背景等比较困难的图像. 而对于每个单独的沙漏模块, 采用跳过层进行信息传递和特征融合, 避免在重复的自下而上和自上而下的过程中逐渐丢失低层的信息. 预处理模块如图 4 所示.

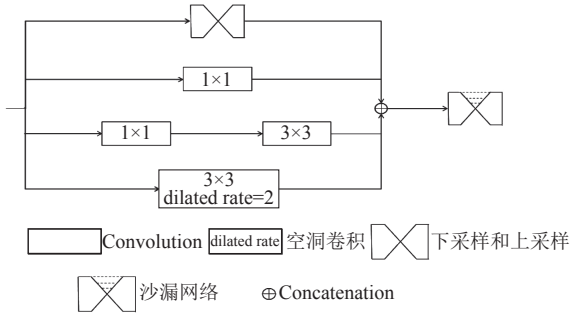

图 4 预处理模块图

现阶段的大多数方法在预处理阶段都是通过卷积池层降低了输入图像的分辨率, 然后经过一个简单的残差模块. 浅层特征对于人体姿态识别更为重要, 因此本文使用了一个多分支并行模块在预处理过程中, 以取代原来的残差模块, 能更好的捕捉浅层特征, 并通过该模块来提取多尺度特征. 不同大小的卷积核不仅可以使网络提取多尺度特征, 对关键点定位任务具有重要意义, 此外, 感受野的大小直接决定了关节结构能否有效地获得, 在此基础上使用 3×3 空洞卷积代替普通卷积作为分支之一, 空洞卷积能在不丢失分辨率下扩大感受野, 相较其他方法能更好的获得浅层特征. 结构特征可以通过连接 4 个分支的输出来捕获, 然后将它们输入到堆叠沙漏网络中.

2.3 分层连接结构

对于每个单独的沙漏块, 采用跳过连接的方式进行信息传递和特征融合, 但每个沙漏网络之间没有联系, 每个沙漏网络都经历自下而上和自上而下的这一过程, 这将导致在反复操作的过程中丢失所获得的信息. 为了使网络充分利用人体结构信息并执行消息传递, 本文设计了不同沙漏块之间的分层连接结构, 如图 5 所示.

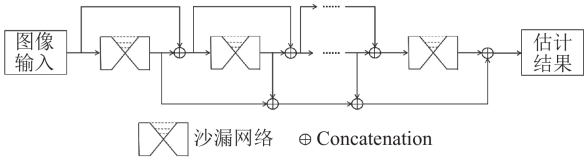

图 5 分层连接结构图

具体而言, 就是将上一阶段得到的浅层和深层的特征进行融合, 以获得局部详细信息和全局上下文信息, 能更好的进行人体姿态估计. 堆叠沙漏网络每个沙漏块之间没有联系, 因此容易丢失所获得的信息, 现阶段的许多方法将各个沙漏模块的输入与输出结果进行特征融合, 较好的提高了估计精度. 通过建立沙漏块之间的连接, 可以有效地将结构特征传递到每个沙漏块的开始, 这样得到的特征可以为整个网络提供有价值的结构信息作为先验, 由于从不同阶段聚集特征的优点, 有助于更好地处理遮挡、大姿态变化、复杂背景等具有挑战性的场景, 并更具鲁棒性. 在此基础上, 本文还融合了每个阶段的结果, 通过平均每个阶段的输出热图, 进一步的特征融合, 充分利用了人体结构信息, 改善了人体姿态估计结果.

3 数据集和评价指标

3.1 数据集

为了评估本文提出的网络模型的性能, 本文使用

的数据集为 LSP (Leeds Sports Pose) 和 MPII 这两种数据集.

LSP 数据集[17] 是从 Flickr 收集的图像样本, 每个图像都注释了 14 个身体关节点, 并包含一些难以估计的姿势. 数据集由 11 000 个训练图像样本和 1000 个测试图像样本组成, 其左侧和右侧关节始终以人为中心进行标记.

MPII 人体姿势数据集[18] 包括大约 25 000 幅图像, 包括超过 40 000 个带注释的姿势, 这些图像是从 YouTube 视频中收集的, 并对人的 16 个全身关节进行注释.

3.2 评价指标

实验中使用了正确关键点的百分比 (PCK)[19] 标准来评估网络模型的性能, 如果检测到的关节与真实关节之间的归一化距离小于设定阈值的比例, 则认为关节被检测到. 另外, 在 MPII 数据集上的官方基准是中是以头部长度 (head length) 作为归一化参考, 即 PCKh.

4 实验及分析

该网络使用 Torch 实现, 并使用 RMSProp 进行优化, 参数随机初始化. 设置 300 个 epoch 训练本文的模型和批次大小为 6, 初始的学习率为2.5×10^{-4}, 在第 240 次 epoch 后降低到 1/10, 在上述的两个数据集上对模型进行训练.

4.1 大感受野残差模块和预处理模块的性能分析

为了评价本文提出的大感受野残差模块和预处理模块对人体姿态估计的影响, 本文使用不同数量的沙漏网络在 LSP 数据集上进行实验, 并给出了在 PCK 评价标准下的实验结果. 结果如表 1 所示.

表 1 LSP 数据集上 PCK 结果 (%)

沙漏网络数量	头	肩	肘部	手腕	臀部	膝盖	脚踝	PCK
2	98.3	92.7	88.7	85.8	92.3	93.3	92.5	91.9
4	98.3	93.7	90.0	88.0	93.6	94.3	93.2	93.0

因为网络的感受野很小, 使得每个沙漏网络的学习能力有限, 将本文提出的模块放入沙漏网络, 增加网络的感受野, 这样能够学习更多的特征信息, 再增加沙漏网络的数量以提高网络的性能, 最终提高人体姿态估计的精度.

4.2 比较实验

将本文提出的方法与近年来的其他人体姿态估计方法进行了比较, 在 LSP 数据集上给出了 PCK 标准下的比较实验结果, 其中, (+) 表示将 MPII 训练集作为额外的样本进行训练. 如表 2 所示.

在 MPII 数据集的实验中, 选择了近年来与 MPII 数据集相关的几种姿态估计方法, 并与我们的方法进行了比较, 在 PCKh 标准下的比较结果在表 3 中给出.

表 2 LSP 数据集上 PCK 比较结果 (%)

方法	头	肩	肘部	手腕	臀部	膝盖	脚踝	PCK
Pishchulin等[20]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei等[16]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat等[21]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Sun等[22]	97.9	93.6	89.0	85.8	92.9	91.2	90.5	91.6
Chu等[23]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Peng等[24]	98.6	95.3	92.8	90.0	94.8	95.3	94.5	94.5
Chen等(+)[25]	98.5	94.0	89.8	87.5	93.9	94.1	93.0	93.1
Yang等(+)[26]	98.3	94.5	92.2	88.9	94.4	95.0	93.7	93.9
Our model	97.9	95.1	92.5	90.3	93.8	94.4	91.8	93.7
Our model(+)	98.0	94.8	92.0	89.2	94.2	94.8	93.8	94.2

表 3 MPII 数据集上 PCKh 比较结果 (%)

方法	头	肩	肘部	手腕	臀部	膝盖	脚踝	PCKh
Pishchulin等[20]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Wei等[16]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat等[21]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Chu等[23]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chen等[25]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang等[26]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Zhang等[8]	98.6	97.0	92.8	88.8	91.7	89.8	86.6	92.5
Our model	98.4	98.2	93.1	89.4	91.8	90.4	86.9	92.9

4.3 结果分析

根据表 1 中的结果可以看出, 增大感受野对人体姿态估计有较好的影响, 同时增加沙漏子网络的数量可以有效地提高网络的性能. 以往的研究表明, 高层特征包含语义等全局信息, 而局部信息往往存在于低层特征中, 浅层的特征对于姿态估计更为重要, 因此增加特征提取时的感受野能更好的对人体关键点进行定位. 根据表 2、表 3 中的结果验证了本文提出的方法的可行性, 因为我们在训练过程中利用大感受野来学习关节点之间的相关性, 预处理模块来更好的获得结构信息, 连接各个沙漏网络, 并对每个沙漏网络的输出结果进行特征融合, 提高了人体姿态估计的准确性.

5 结语

本文提出了一种用于人体姿态估计的改进堆叠沙漏网络. 该网络利用两种残差模块提取图像特征信息

和了解关节点间的相互关系, 因此能更准确地定位关节点位置; 另一方面, 在堆叠沙漏网络前加上了预处理模块, 有效地提取关节结构作为先验, 在可见的基础上为推断困难关键点提供了指导; 最后将不同的沙漏模块进行分层连接, 对估计结果进行特征融合, 以进一步改善人体姿态估计的精度. 本方法在两个数据集上进行了测试, 在取得了较好的结果的同时, 也在实验中也观察到了一些失败, 包括多人或罕见的姿势在图片中, 在一个端到端的体系结构中处理多个人也是一个具有挑战性的问题, 今后的工作将继续对这两方面的情况探索.

参考文献

1 Wang CY, Wang YZ, Yuille AL. An approach to pose-based action recognition. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 915–922.

2 Liang ZJ, Wang XL, Huang R, *et al*. An expressive deep model for human action parsing from a single image. Proceedings of 2014 IEEE International Conference on Multimedia and Expo (ICME). Chengdu: IEEE, 2014. 1–6.

3 Cho NG, Yuille AL, Lee SW. Adaptive occlusion state estimation for human pose tracking under self-occlusions. Pattern Recognition, 2013, 46(3): 649–661. [doi: 10.1016/j.patcog.2012.09.006]

4 Nie BX, Xiong CM, Zhu SC. Joint action recognition and pose estimation from video. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1293–1301.

5 Shotton J, Sharp T, Kipman A, *et al*. Real-time human pose recognition in parts from single depth images. Communications of the ACM, 2013, 56(1): 116–124. [doi: 10.1145/2398356.2398381]

6 Sarafianos N, Boteanu B, Ionescu B, *et al*. 3D human pose estimation: A review of the literature and analysis of covariates. Computer Vision and Image Understanding, 2016, 152: 1–20. [doi: 10.1016/j.cviu.2016.09.002]

7 Dantone M, Gall J, Leistner C, *et al*. Human pose estimation using body parts dependent joint regressors. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 3041–3048.

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

9 Tang W, Yu P, Wu Y. Deeply learned compositional models for human pose estimation. Proceedings of the 15th European Conference on Computer Vision (ECCV). Cham: Springer, 2018. 197–214.

10 Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016. 483–499.

11 Ke LP, Chang MC, Qi HG, *et al*. Multi-scale structure-aware network for human pose estimation. Proceedings of the 15th European Conference on Computer Vision (ECCV). Cham: Springer, 2018. 731–746.

12 Chen YL, Wang ZC, Peng YX, *et al*. Cascaded pyramid network for multi-person pose estimation. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7103–7112.

13 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.

14 Cao Z, Hidalgo G, Simon T, *et al*. OpenPose: Realtime multi-person 2D Pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172–186. [doi: 10.1109/TPAMI.2019.2929257]

15 He KM, Zhang XY, Ren Q, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

16 Wei SE, Ramakrishna V, Sheikh Y, *et al*. Convolutional pose machines. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4724–4732.

17 Johnson S, Everingham M. Learning effective human pose estimation from inaccurate annotation. Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011. 1465–1472.

18 Andriluka M, Pishchulin L, Gehler P, *et al*. 2D human pose estimation: New benchmark and state of the art analysis. Proceedings of 2014 IEEE Conference on computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 3686–3693.

19 Yang Y, Ramanan D. Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2878–2890. [doi: 10.1109/TPAMI.2012.261]

20 Pishchulin L, Insafutdinov E, Tang SY, *et al*. DeepCut: Joint

subset partition and labeling for multi person pose estimation. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4929–4937.

21 Bulat A, Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression. Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016. 717–732.

22 Sun K, Lan CL, Xing JL, *et al*. Human pose estimation using global and local normalization. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 5600–5608.

23 Chu X, Yang W, Ouyang WL, *et al*. Multicontext attention for human pose estimation. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5669–5678.

24 Peng X, Tang ZQ, Yang F, *et al*. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2226–2234.

25 Chen Y, Shen CH, Wei XS, *et al*. Adversarial posenet: A structure-aware convolutional network for human pose estimation. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1221–1230.

26 Yang W, Li S, Ouyang WL, *et al*. Learning feature pyramids for human pose estimation. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1290–1299.