软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(4):2039−2054 [doi: 10.13328/j.cnki.jos.006837]

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

基于人体和场景上下文的多人 3D 姿态估计*

何建航, 孙郡瑶, 刘 琼

(华南理工大学 软件学院, 广东 广州 510006)

通信作者: 刘琼, E-mail: liuqiong@scut.edu.cn

摘 要: 深度歧义是单帧图像多人 3D 姿态估计面临的重要挑战, 提取图像上下文对缓解深度歧义极具潜力. 自顶向下方法大多基于人体检测建模关键点关系, 人体包围框粒度粗背景噪声占比较大, 极易导致关键点偏移或误匹配, 还将影响基于人体尺度因子估计绝对深度的可靠性. 自底向上的方法直接检出图像中的人体关键点再逐一恢复 3D 人体姿态. 虽然能够显式获取场景上下文, 但在相对深度估计方面处于劣势. 提出新的双分支网络, 自顶向下分支基于关键点区域提议提取人体上下文, 自底向上分支基于三维空间提取场景上下文. 提出带噪声抑制的人体上下文提取方法, 通过建模“关键点区域提议”描述人体目标, 建模姿态关联的动态稀疏关键点关系剔除弱连接减少噪声传播. 提出从鸟瞰视角提取场景上下文的方法, 通过建模图像深度特征并映射鸟瞰平面获得三维空间人体位置布局; 设计人体和场景上下文融合网络预测人体绝对深度. 在公开数据集 MuPoTS-3D 和 Human3.6M 上的实验结果表明: 与同类先进模型相比, 所提模型 HSC-Pose 的相对和绝对 3D 关键点位置精度至少提高 2.2% 和 0.5%; 平均根关键点位置误差至少降低 4.2 mm.

关键词: 多人场景 3D 姿态估计; 关键点区域提议; 人体上下文; 场景上下文; 人体绝对深度

中图法分类号: TP391

中文引用格式: 何建航, 孙郡瑶, 刘琼. 基于人体和场景上下文的多人3D姿态估计. 软件学报, 2024, 35(4): 2039–2054. http://www.jos.org.cn/1000-9825/6837.htm

英文引用格式: He JH, Sun JY, Liu Q. Multi-person 3D Pose Estimation Using Human-and-scene Contexts. Ruan Jian Xue Bao/Journal of Software, 2024, 35(4): 2039–2054 (in Chinese). http://www.jos.org.cn/1000-9825/6837.htm

Multi-person 3D Pose Estimation Using Human-and-scene Contexts

HE Jian-Hang, SUN Jun-Yao, LIU Qiong

(School of Software Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract: Depth ambiguity is an important challenge for multi-person three-dimensional (3D) pose estimation of single-frame images, and extracting contexts from an image has great potential for alleviating depth ambiguity. Current top-down approaches usually model key point relationships based on human detection, which not only easily results in key point shifting or mismatching but also affects the reliability of absolute depth estimation using human scale factor because of a coarse-grained human bounding box with large background noise. Bottom-up approaches directly detect human key points from an image and then restore the 3D human pose one by one. However, the approaches are at a disadvantage in relative depth estimation although the scene context can be obtained explicitly. This study proposes a new two-branch network, in which human context based on key point region proposal and scene context based on 3D space are extracted by top-down and bottom-up branches, respectively. The human context extraction method with noise resistance is proposed to describe the human by modeling key point region proposal. The dynamic sparse key point relationship for pose association is modeled to eliminate weak connections and reduce noise propagation. A scene context extraction method from a bird's-eye-view is proposed. The human position layout in 3D space is obtained by modeling the image's depth features and mapping them to a bird's-eye-view plane. A

* 基金项目: 广东省自然科学基金 (2021A1515011349); 国家自然科学基金 (61976094)
收稿时间: 2022-05-31; 修改时间: 2022-08-16, 2022-09-26; 采用时间: 2022-11-22; jos 在线出版时间: 2023-07-28
CNKI 网络首发时间: 2023-08-01

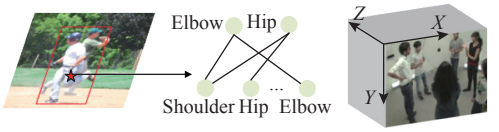
network fusing human and scene contexts is designed to predict absolute human depth. The experiments are carried out on public datasets, namely MuPoTS-3D and Human3.6M, and results show that compared with those by the state-of-the-art models, the relative and absolute position accuracies of 3D key points by the proposed HSC-Pose are improved by at least 2.2% and 0.5%, respectively, and the position error of mean roots of the key points is reduced by at least 4.2 mm.

Key words: multi-person 3D pose estimation; keypoint region proposal; human context; scene context; absolute human depth

人体姿态估计是人体行为识别的重要技术, 广泛应用于安全监控, 自动驾驶, 人机交互, 虚拟现实和运动分析等领域, 深受学术界和工业界青睐. 随着社会的进步和技术的发展, 识别理解人体动作和人体间位置关系变得日愈重要. 3D 人体姿态估计即从单张图像或视频估计 3D 人体关键点位置. 由于单视角人体深度和尺度模糊, 从单张图像恢复多人场景 3D 姿态是一个挑战性问题[1]. 多人场景绝对 3D 姿态估计不仅包含相对 3D 姿态估计 (相对深度), 还包含绝对人体位置及位置间的关系估计 (绝对深度)[2]. 相对深度指人体关键点间的距离, 关联肢体朝向, 关键点间的运动学约束等人体上下文信息. 绝对深度指人体根关键点与相机间的距离, 关联相机位置, 人体位置及人体间的位置关系等场景上下文信息[3,4].

当前主流的多人 3D 姿态估计分为自顶向下 (top-down) 和自底向上 (bottom-up) 两类方法[4,5]. 自顶向下方法通过逐一检出场景中的人体恢复 3D 姿态, 对获取人体上下文优势明显, 但无法获取场景上下文. 自底向上方法直接检测场景中的人体关键点再逐一恢复 3D 人体姿态. 虽然能够显式获取场景上下文, 但在相对深度估计方面处于劣势.

Moon 等人[3]采用自顶向下框架, 基于人体检测获得人体包围框 (bounding box) 再建模人体关键点关系 (如图 1 所示) 和提取人体尺度因子, 按照近大远小的透视原理预测人体绝对深度. 继之, Guo 等人[2]基于 2D 人体姿态提取尺度校正因子控制人体尺度免遭姿态变化的影响. 然而, Moon 等人和 Guo 等人基于 3D 统计尺度处理 2D 图像人体的方式使模型敏感于场景中变化的人体尺度. 为此, Dabral 等人[5]通过获取各种尺度的 3D 人体骨骼长度修正尺度因子, 缓解人体姿态或体态变化对绝对深度估计的影响. 但是, 缺失场景布局等场景上下文信息, 仅靠尺度因子很难估计遮挡目标的绝对深度.



(a) 人体包围框 (b) 关键点关系 (c) 三维空间模型

图 1 基于人体包围框建模关键点关系

Zhen 等人[4]提出自底向上的单步多人 3D 姿态估计方法, 利用沙漏网络获得场景 2D 关键点热图及其亲和场 (part affinity fields, PAFs), 人体相对及绝对深度映射, 再基于人体深度感知部件关联算法恢复多人场景 3D 姿态. 由于不涉及人体检测, 该方法的相对深度估计显薄弱. 为此, Wang 等人[6]提出单阶段自底向上计算框架, 并设计人体关键点分布感知模型并行完成人体相对和绝对深度估计. 由于敏感于场景中的多种人体尺度, 其相对深度估计尚不如自顶向下方法. Cheng 等人[7]交叉联合自顶向下和自底向上网络进行两分支 3D 姿态估计, 并通过合并获得最终结果. 尽管该方法能兼顾人体和场景上下文, 但是重复计算人体姿态成本高昂. Wang 等人[8]分别应用自顶向下和自底向上分支进行相对和绝对人体深度估计, 明显降低计算开销.

虽然联合自顶向下和自底向上框架进行多人场景 3D 姿态估计更具潜力, 但是, 目前方法普遍忽视了两个重要问题: (1) 基于人体包围框恢复相对人体姿态仍然遭受背景噪声的影响, 当背景复杂或人体位置关系紧凑时, 包围框通常还包含他人, 树木, 草地及路面等噪声, 如图 1(a) 红线框所示. (2) 显式的绝对深度估计局限于图像平面, 场景中的位置关系仅限于描述图像平面像素间的距离.

关于第 1 个问题, 2D 姿态估计使用关键点集取代人体包围框提取人体上下文明显降低了背景噪声对性能的影响[9]. 但是, 这种方法对人体尺度变化十分敏感, 而且未考虑诸如边缘, 肢体朝向等关键点邻域特征, 改进姿态估

计效果受限. Lin 等人[10]基于人体检测构建 2D 关键点热图, 聚焦关键点区域提取人体上下文, 能够明显减少背景或服饰等噪声干扰; 但是, 获取准确的关键点热图成本高昂. 近期, 连通性和对称性等人体先验广泛应用于姿态估计图卷积网络的设计, 以减少关键点间的噪声传播[11]. 但是, 仅适合建模静态关键点关系, 很难适应变化的人体姿态. 关于第 2 个问题, 近期 Reading 等人[12]设计人体深度分布分类网络进行单目 3D 目标检测, 他们从鸟瞰视角提取场景上下文获得人体位置布局的方法对多人场景 3D 姿态估计具有深远的启发意义. 但是, 鸟瞰视角场景缺失人体尺度信息, 而透视原理表明人体尺度是绝对深度估计的重要依据.

从全局图像挖掘人体和场景上下文是缓解深度歧义的关键[7]. 本文提出新的两分支网络, 自顶向下分支采用“关键点区域提议”替代人体包围框描述人体目标, 并兼及背景噪声、边缘、肢体朝向等信息优化关键点区域特征描述, 进而建模姿态关联的动态稀疏关键点关系提高模型的相对姿态恢复能力. 自底向上分支从鸟瞰平面而非图像平面提取场景上下文获得三维空间人体位置布局, 联合人体和场景上下文可靠预测人体绝对深度. 本文主要贡献如下.

(1) 提出新的双分支网络, 自顶向下分支基于关键点区域提议提取人体上下文, 自底向上分支基于三维空间提取场景上下文.

(2) 提出带噪声抑制的人体上下文提取方法, 建模“关键点区域提议”描述人体目标, 建模姿态关联的动态稀疏关键点关系剔除弱连接减少噪声传播.

(3) 提出基于鸟瞰视角场景上下文预测场景布局及位置关系的方法, 设计人体和场景上下文融合网络预测人体绝对深度.

(4) 在公开数据集 MuPoTS-3D 和 Human3.6M 的实验结果表明: 与同类先进模型相比, 本文模型 HSC-Pose 的相对和绝对 3D 关键点位置精度至少提高 2.2% 和 0.5%; 根关键点位置误差至少降低 4.2 mm.

1 噪声抑制人体上下文提取

采用自顶向下的方法提取人体上下文, 涉及人体检测和人体关键点关系建模. 建模“关键点区域提议”描述人体目标能够显著降低背景噪声对人体上下文的影响, 建模姿态关联的动态稀疏关键点关系进一步抑制噪声传播, 以提高关键点位置估计精度. 相关网络设计涉及关键点区域提议模块 KRPM (keypoint region proposal module) 和姿态关联关键点关系模块 PKRM (pose-relative keypoint relationship module).

1.1 建模“关键点区域提议”

1.1.1 提取关键点区域

设人体关键点 k 的包围框真值为 $(\mu_k, w_k, h_k, \theta_k)$, 其中, $k \in \{0,\ldots,K-1\}$, K 表示除根关键点之外的人体关键点数目, $\mu_k = [x_k \in \mathbb{R}, y_k \in \mathbb{R}]^{\mathrm{T}}$, $w_k \in \mathbb{R}$, $h_k \in \mathbb{R}$ 和 $\theta_k \in \mathbb{R}$ 分别表示包围框中心坐标, 宽, 高和旋转角. 关键点 k 的预测包围框为 $(\hat{\mu}_k, \hat{w}_k, \hat{h}_k, \hat{\theta}_k)$. 得益于旋转角, 关键点包围框具有肢体朝向适应性, 利于避开背景噪声.

采用 BasicBlock[13]构建中心分类网络和包围框回归网络预测中心分类图 $\boldsymbol{C}$ ($\boldsymbol{C} \in \mathbb{R}^{1\times(H/4)\times(W/4)}$) 和包围框图谱 $\boldsymbol{M}$ ($\boldsymbol{M} \in \mathbb{R}^{5\times(H/4)\times(W/4)}$) 获得像素级人体中心 (人体根关键点) 及预测包围框. 其中, H 和 W 分别表示输入图像的高度和宽度. 分类网络和回归网络如后文图 2 所示, 分类图 $\boldsymbol{C}$ 反映图像像素 p 是否属于人体根关键点区域, $\boldsymbol{M}$ 反映图像像素 p 是否对应人体 K 个关键点的包围框 $(\hat{\mu}_k, \hat{w}_k, \hat{h}_k, \hat{\theta}_k)$, 偏移量估计 $\hat{\mu}_k$ 是通过度量像素 p 到包围框中心的距离获得[14], 即 $\hat{\mu}_k = \mu_p + \hat{o}_k$, $\hat{o}_k \in \mathbb{R}^2$ 表示偏移量, $\mu_p \in \mathbb{R}^2$ 表示像素 p 的坐标.

采用 3×3 窗口的非极大值抑制 NMS (non-maximum suppression) 计算从中心分类图 $\boldsymbol{C}$ 中提取候选人体根关键点, 进而从关键点包围框图谱 $\boldsymbol{M}$ 中提取候选人体 K 个感兴趣关键点区域提议.

1.1.2 损失函数设计

使用 L1 或 L2 损失函数优化 $\hat{\mu}_k$, $\hat{w}_k$, $\hat{h}_k$ 和 $\hat{\theta}_k$, 需要提供人工标注真值 w_k, h_k 和 θ_k, 而且各个参数的优化过程相互独立. 为了使关键点区域提议尽可能覆盖人体关键点并减少背景噪声占比, 需联合优化上述参数. 使得关键点包围框中心接近当前关键点时, 适当缩小包围框尺寸以滤除更多背景噪声. 为此, 我们采用高斯分布建模关键点包

围框, 进而求取真值包围框和预测包围框的高斯分布之间的 KL 散度. 基于 KL 散度计算设计损失函数 *KLD* (Kullback-Leibler divergence), 如公式 (1).

$$KLD=\frac{(\Delta x\cos\hat{\theta}_k+\Delta y\sin\hat{\theta}_k)^2}{2\hat{w}_k^2/(2\lambda)^2}+\frac{(\Delta y\cos\hat{\theta}_k-\Delta x\sin\hat{\theta}_k)^2}{2\hat{h}_k^2/(2\lambda)^2}+\ln(\hat{w}_k/(2\lambda))+\ln(\hat{h}_k/(2\lambda)) \tag{1}$$

其中, $\Delta x=x_k-\hat{x}_k$, $\Delta y=y_k-\hat{y}_k$. λ 表示包围框扩展系数. 根据高斯分布的 3-Sigma 法则, 设置 λ=3, 进而通过实验获得 λ 取值与关键点覆盖率的关系验证设置 λ=3 的合理性, 详见表 1. 损失函数 *KLD* 中选用 $\hat{w}_k$ 和 $\hat{h}_k$ 而不是 w_k 和 h_k, 训练过程中根据包围框预测误差调整损失, 例如, 预测误差 $(\Delta x\cos\hat{\theta}_k+\Delta y\sin\hat{\theta}_k)^2$ 越大, 则最小化 *KLD* 的过程中 $\hat{w}_k^2$ 越大, 将促使包围框尽可能覆盖当前关键点.

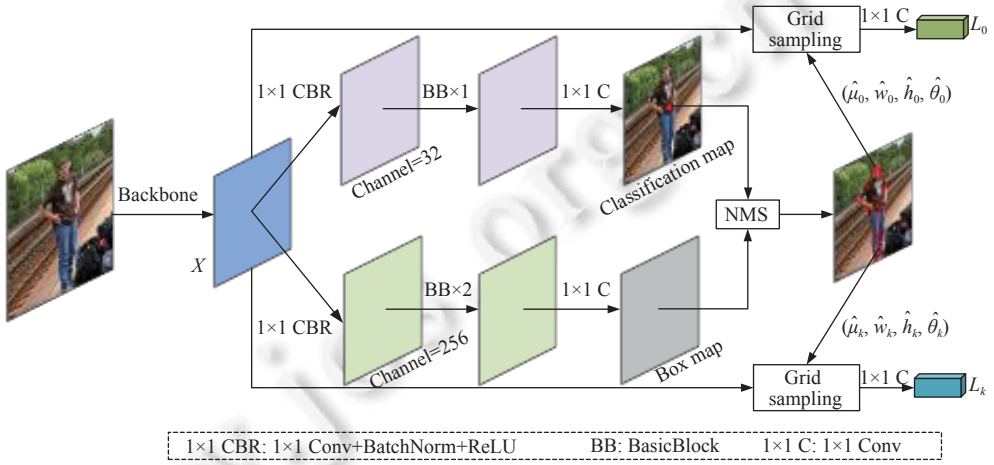


图 2 关键点区域提议 (KRPM 模块)

表 1 包围框扩展系数取值与关键点覆盖率

包围框扩展系数	关键点覆盖率 (%)
1.0	42.5
2.0	81.3
3.0	92.1

通过联合优化 Δx, Δy, $\hat{w}_k$ 和 $\hat{h}_k$ 预测值 $\hat{\theta}_k$, 损失函数 *KLD* 无需设定包围框旋转角真值 θ_k. *KLD* 关于 $\hat{\theta}_k$ 的偏导数如公式 (2), 可以看出: (1) 包围框面积不变时, $\hat{w}_k$ 和 $\hat{h}_k$ 的差值越大, *KLD* 关于 $\hat{\theta}_k$ 偏导的绝对值越大, 即 $\hat{\theta}_k$ 优化越显著; (2) 例如, $[\cos\hat{\theta}_k,\sin\hat{\theta}_k]=\left[\Delta x/\sqrt{(\Delta x)^2+(\Delta y)^2},\Delta y/\sqrt{(\Delta x)^2+(\Delta y)^2}\right]$ 时, 损失函数 *KLD* 关于 $\hat{\theta}_k$ 的偏导数为 0, 说明 $\hat{\theta}_k$ 最终促使包围框朝向 $[\Delta x,\Delta y]^{\mathrm{T}}$.

$$\frac{\partial KLD}{\partial\hat{\theta}_k}=(2\lambda)^2\left(\frac{1}{\hat{w}_k^2}-\frac{1}{\hat{h}_k^2}\right)(\Delta x\cos\hat{\theta}_k+\Delta y\sin\hat{\theta}_k)(\Delta y\cos\hat{\theta}_k-\Delta x\sin\hat{\theta}_k) \tag{2}$$

1.1.3 优化关键点区域

为了根据下游任务调整关键点包围框参数, 采用网格采样 GS (grid sampling)[15]从特征图 $\boldsymbol{X}$ (源自骨干网络输出) 提取关键点区域特征. 不同于 RoI Align[16], 网格采样利用下游任务优化关键点包围框 $\hat{\mu}_k$, $\hat{w}_k$, $\hat{h}_k$ 和 $\hat{\theta}_k$. 假设候选人体关键点 k 的区域特征为 $\boldsymbol{L}_k\in\mathbb{R}^{C_L\times H_L\times W_L}$, H_L 和 W_L 分别表示 $\boldsymbol{L}_k$ 的高和宽. 由 $\boldsymbol{L}_k$ 的像素位置 α_L 计算特征图 $\boldsymbol{X}$ 对应的采样位置 α_X, 如公式 (3), $[(W_L-1)/2,(H_L-1)/2]^{\mathrm{T}}$ 表示 $\boldsymbol{L}_k$ 中心坐标.

$$\alpha_X=\hat{\mu}_k+\begin{bmatrix}\cos\hat{\theta}_k & -\sin\hat{\theta}_k\\ \sin\hat{\theta}_k & \cos\hat{\theta}_k\end{bmatrix}\begin{bmatrix}\hat{w}_k/W_L & 0\\ 0 & \hat{h}_k/H_L\end{bmatrix}\left(\alpha_L-\begin{bmatrix}(W_L-1)/2\\ (H_L-1)/2\end{bmatrix}\right) \tag{3}$$

1.2 建模姿态关联的关键点关系

通过模型训练获得的静态关键点关系, 姿态恢复能力明显不足[11]. 近期基于关键点间的语义相似性获得的动态关键点关系, 具有姿态变化适应性能够增强模型的姿态恢复能力[17]. 然而, 当前动态关键点关系建模方法多半基于包围框描述人体目标, 不涉及关键点区域噪声特征和姿态几何特征等优化因素, 模型的鲁棒性仍然遭受背景噪声的影响.

本文在建模关键点区域提议的基础上, 综合关键点区域特征, 关键点区域噪声特征和姿态几何特征等因素建模姿态关联的关键点关系, 能够显式抑制噪声对模型性能的影响. 实现过程分两个步骤: (1) 建模动态关键点关系, 所定义的人体关键点间的关联权重各不相等, 当前关联权重随目标关键点区域特征, 关键点区域噪声特征和姿态几何特征等的变化而变化. (2) 在动态关键点关系的基础上建模动态稀疏关键点关系, 剔除关联权重低于设定阈值的关键点关系.

1.2.1 建模动态关键点关系

关键点区域噪声将造成关键点包围框中心偏离关键点, 还将沿关联路径传播, 降低关键点区域特征和关键点关系的可靠性. 因此, 防止噪声传播是建模动态关键点关系的另一目标. 损失函数 *KLD* 设计表明: 在同等人体尺度下, 关键点包围框尺寸越小质量越高, 参见公式 (2). 逐人体规范化关键点包围框获得关键点区域噪声特征 $\boldsymbol{V}_{\text{noise}}$, 基于 $\boldsymbol{V}_{\text{noise}}$ 和 $\boldsymbol{L}_k$ 进行关键点关系建模, 降低噪声响应.

但是, 关键点包围框关联关键点的邻近区域, 所提取的关键点区域特征不包含关键点间的位置约束. 为了弥补这个不足, 采用人体姿态几何特征 (以三维向量 $\boldsymbol{V}_{\text{geo}}$ 表示) 描述关键点间的位置约束 (含距离和相对偏移). 逐人体计算关键点间的距离和相对偏移, 获得距离集合 $\{\|\hat{\mu}_i-\hat{\mu}_k\|\,|(i,k)\in\varepsilon\}$ 和偏移集合 $\{\hat{\mu}_i-\hat{\mu}_k|(i,k)\in\varepsilon\}$. 然后将距离集合和偏移集合的内容分别拼接成通道数为 $K(K-1)/2$ 的距离特征和通道数为 $K(K-1)$ 的相对偏移特征. 对距离特征和相对偏移特征进行通道拼接获得 $\boldsymbol{V}_{\text{geo}}$. ε 表示当前关键点的骨骼连接集合. $\boldsymbol{V}_{\text{geo}}$ 的引入使建模的关键点关系包含关键点间的空间位置约束.

基于 $\boldsymbol{L}_k$, $\boldsymbol{V}_{\text{noise}}$, $\boldsymbol{V}_{\text{geo}}$ 评估关键点的质量及其对相关关键点关系的贡献, 逐人体执行以下计算: 沿通道拼接当前关键点的 $\boldsymbol{L}_k$, $\boldsymbol{V}_{\text{noise}}$, $\boldsymbol{V}_{\text{geo}}$ 获得相应向量表达, 再采用压缩-激活计算 (squeeze-and-excitation)[18]获得表达动态关键点关系的矩阵 $\boldsymbol{R}_{dy}\in\mathbb{R}^{K\times K}$, 其原理过程如图 3 所示.

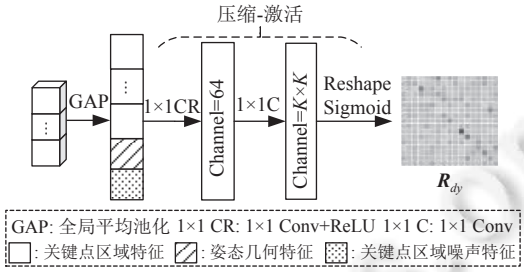
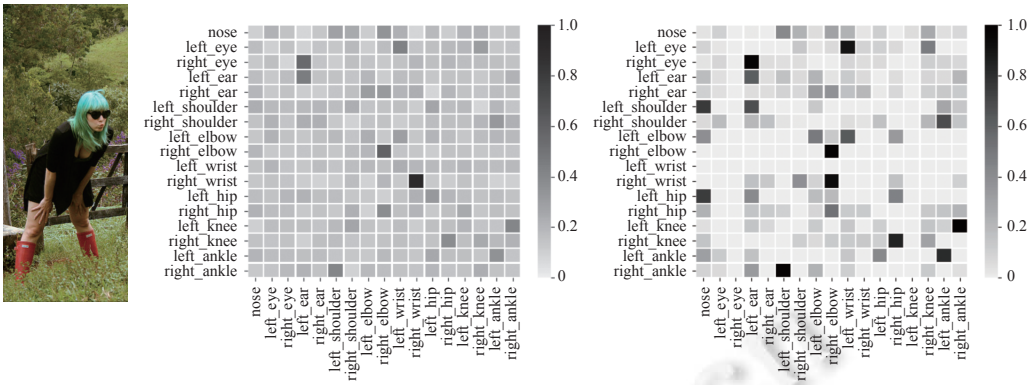


图 3 建模动态关键点关系

1.2.2 动态稀疏关键点关系

采用矩阵 $\boldsymbol{R}_{dy}$ 表达人体动态关键点关系, 当前 $\boldsymbol{R}_{dy}$ 的元素值均处于 (0, 1) 区间. 元素值趋于“1”表示渐强的关键点关系, 趋于“0”表示渐弱的关键点关系. 针对实际人体观察到 $\boldsymbol{R}_{dy}$ 中存在为数不少的弱连接关键点关系, 如图 4(b) 中的浅灰色关键点关系. 弱连接对恢复当前姿态的正面贡献不大, 却提供了噪声传播途径.

构建 $\boldsymbol{R}_{sp}^{dy}=\sum_{\tau}\omega_{\tau}f(\boldsymbol{R}_{dy},\tau)$ 表达姿态关联的动态稀疏关键点关系, 其中, $f(\boldsymbol{R}_{dy},\tau)$ 表示小于 τ 的 $\boldsymbol{R}_{dy}$ 元素被置为 0; ω_{τ} 表示加权系数, 满足 $\sum\omega_{\tau}=1$; 预设 $\tau\in\{0.35,0.5,0.75\}$. 对于不同姿态, $\boldsymbol{R}_{dy}$ 中元素值趋于“1”的关键点关系分布各异. 如果将 $\boldsymbol{R}_{sp}^{dy}$ 退化为固定阈值的关键点关系矩阵, 容易因关键点关系过稀疏或过密集而导致人体姿态估计欠约束或负迁移. 为此, 利用图 5 所示网络根据当前人体的关键点区域特征, 关键点区域噪声特征和姿态几何特征动态生成 ω_{τ} 并获得相应的 $\boldsymbol{R}_{sp}^{dy}$.



(a) 目标人体 (b) 动态关键点关系 (c) 动态稀疏关键点关系

图 4 目标人体 (姿态) 的关键点关系

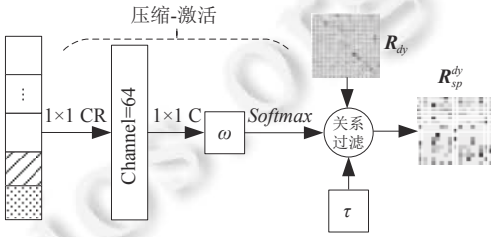
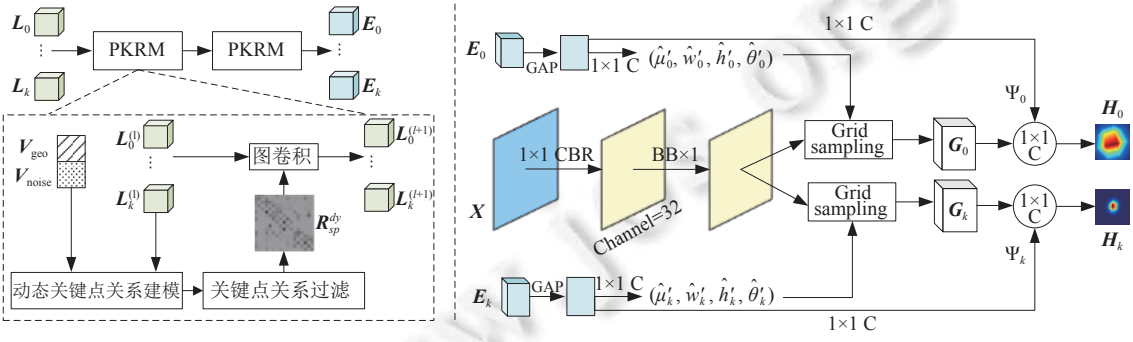


图 5 剔除弱连接的关键点关系

1.2.3 相对 3D 姿态估计

对存在紧凑位置关系的多人场景, 可能出现人体根关键点重叠, 基于根关键点估计其他 K 个关键点区域不够可靠. 基于姿态关联的动态稀疏关键点关系矩阵 $\boldsymbol{R}_{sp}^{dy}$ 和图卷积[11]提取人体上下文, 借助人体上下文获得增强关键点区域特征 $\boldsymbol{E}_k$, $k\in\{0,\ldots,K-1\}$, 如图 6(a) 所示. 基于 $\boldsymbol{E}_k$ 优化关键点区域 $(\hat{\mu}_k',\hat{w}_k',\hat{h}_k',\hat{\theta}_k')$ 获得新的关键点区域特征 $\boldsymbol{G}_k\in\mathbb{R}^{C_G\times H_G\times W_G}$, 再利用动态生成的卷积权重 Ψ_k[19]卷积 $\boldsymbol{G}_k\in\mathbb{R}^{C_G\times H_G\times W_G}$, 获得关键点热图估计 $\boldsymbol{H}_k\in\mathbb{R}^{32\times H_G\times W_G}$, 如图 6(b) 所示.



(a) 建模动态稀疏关键点关系 (PKRM 模块) (b) 关键点区域特征优化

图 6 动态稀疏关键点关系 (PKRM) 及关键点区域优化

热图 $\boldsymbol{H}_k$ 提供了关键点 k 处于 $32\times H_G\times W_G$ 三维空间各位置的置信度, 加权求和这些位置获得关键点位置向量 $J_k\in\mathbb{R}^3$[20], 如公式 (4), 其中 $Softmax(\boldsymbol{H}_k)(z',y',x')$ 表示经 $Softmax$ 归一化后 (z',y',x') 处的置信度. 再采用级联网络[21]优化候选 3D 姿态估计 $\boldsymbol{J}=\{J_0,\ldots,J_{K-1}\}$.

$$J_k=(z,y,x)=\sum_{z'=0}^{32-1}\sum_{y'=0}^{H_G-1}\sum_{x'=0}^{W_G-1}(z',y',x')\times Softmax(\boldsymbol{H}_k)(z',y',x') \tag{4}$$

2 鸟瞰视角场景上下文提取及人体绝对深度预测

本节介绍图像虚拟深度创建, 鸟瞰视角场景特征映射, 场景上下文提取和人体绝对深度预测等, 如图 7 所示.

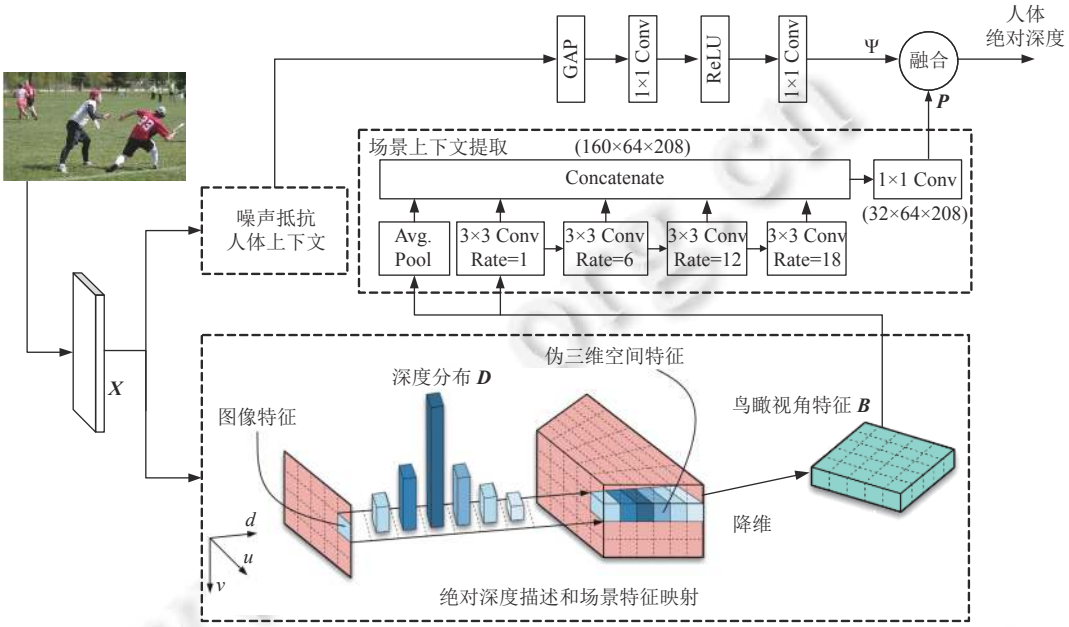


图 7 场景上下文提取

2.1 图像深度特征和场景上下文

2.1.1 创建图像虚拟深度

设在相机坐标系 X-Y-Z 下, 图像处于 X-Y 平面, 场景处于 X-Z 平面, 参见图 1(c). 采用三维矩阵 $\boldsymbol{D}\in\mathbb{R}^{N_D\times(H/4)\times(W/4)}$ 描述具有深度分布的图像空间. 其中 H 表示图像高度, W 表示图像宽度. 设图像平面像素 (u,v) 在其深度维度有 N_D 个预设绝对深度, 以 $\boldsymbol{D}(u,v)\in\mathbb{R}^{N_D\times 1}$ 表示 (u,v) 点的 N_D 个绝对深度的概率. 采用自然对数求取 (u,v) 点处的 N_D 个预设绝对深度 z_i, $i\in\{0,\ldots,N_D-1\}$, 如公式 (5).

$$\ln z_i=\ln\beta+i\left(\frac{\ln\eta-\ln\beta}{(N_D-1)}\right) \tag{5}$$

其中, β 和 η 表示针对特定数据集统计获得的绝对深度最小值和最大值, 如对于 MuPoTS-3D 数据集 β 和 η 分别为 0.5 m 和 10 m, 符合当前连续波单目深度相机的有效范围[22].

级联 BasicBlock[13]和 *Softmax* 计算图像特征 $\boldsymbol{X}$ (骨干网络输出) 提取图像的深度分布 $\boldsymbol{D}$, 为使 (u,v) 点的 N_D 个预设绝对深度的期望接近真值, 选用 DFL (distribution focal loss)[23]作为监督学习损失函数, 如公式 (6).

$$\mathrm{DFL}=-((z_{i+1}-z^*(u,v))\ln(S_i(u,v))+(z^*(u,v)-z_i)\ln(S_{i+1}(u,v))) \tag{6}$$

其中, z_i 和 z_{i+1} 分别表示第 i 和 $i+1$ 个预设绝对深度, $z_i<z^*(u,v)<z_{i+1}$, $z^*(u,v)$ 表示 (u,v) 点的绝对深度真值, $S_i(u,v)$ 和 $S_{i+1}(u,v)$ 分别表示 (u,v) 点的第 i 和 $i+1$ 个绝对深度的概率值.

2.1.2 鸟瞰视角场景特征映射

在 X-Y-Z 下, X-Y 平面的一列图像像素对应 X-Z 平面的一个像素, X-Y 平面的相同列像素可能涉及多个人

体, 这种多对一的映射关系存在歧义性. 降维图像特征 $\boldsymbol{X}$ 获得通道数目为 C_F 的特征表达 $\boldsymbol{F}\in\mathbb{R}^{C_F\times(H/4)\times(W/4)}$, 按图像深度分布 $\boldsymbol{D}$ 映射 $\boldsymbol{F}$ 获得 $\boldsymbol{E}\in\mathbb{R}^{C_F\times N_D\times(H/4)\times(W/4)}$, 如公式 (7).

$$\boldsymbol{E}(i,u,v)=\boldsymbol{D}(i,u,v)\odot\boldsymbol{F}(u,v) \tag{7}$$

其中, $\boldsymbol{F}(u,v)\in\mathbb{R}^{C_F\times1}$ 表示图像像素 (u,v) 的特征向量, $\boldsymbol{E}(i,u,v)\in\mathbb{R}^{C_F\times1}$ 表示三维空间 (i,u,v) 处的特征向量, $\boldsymbol{D}(i,u,v)\in\mathbb{R}$ 表示 (u,v) 处于第 i 个 ($i\in[0,N_D-1]$) 预设绝对深度的概率.

在 X-Y-Z 空间提取场景上下文的计算复杂度为 $\mathrm{O}(N_D\times C_F\times H\times W)$, 为降低计算复杂度, 纵向压缩 $\boldsymbol{E}$ 并将特征通道由 C_F 降维至 C_B, 获取形状为 $C_B\times N_D\times(W/4)$ 的鸟瞰视角场景映射的初始特征. 为了优化像素绝对深度, 上采样图像深度维度特征获得鸟瞰视角场景映射特征 $\boldsymbol{B}\in\mathbb{R}^{C_B\times2N_D\times(W/4)}$, 提取场景上下文的计算复杂度降为 $\mathrm{O}(N_D\times C_B\times W)$. 场景深度 z_d 与 $\boldsymbol{B}$ 的深度 $d\in[0,2N_d-1]$ 的存在如公式 (8) 所示的关系.

$$\ln z_d=\ln\beta+d\left(\frac{\ln\eta-\ln\beta}{(2N_D-1)}\right) \tag{8}$$

2.1.3 场景上下文提取

利用瀑布带孔空间金字塔 WASP (waterfall atrous spatial pyramid)[24], 在不同感受野下建模实体间的位置关系, 获得人体间的场景位置, 即提取场景上下文. 瀑布带孔空间金字塔计算如公式 (9).

$$\boldsymbol{P}=Conv_{1\times1}\left(\left(\bigcup_{l=1}^{4}Conv_{3\times3}(\boldsymbol{O}_{l-1};r_l)\right)\right)\cup Gap(\boldsymbol{O}_0) \tag{9}$$

其中, $Conv_{1\times1}(\cdot)$ 表示 1×1 卷积, $Conv_{3\times3}(\cdot;r_l)$ 表示第 l 个 3×3 卷积, r_l 表示带孔率 (r_1, r_2, r_3, r_4 分别等于 1, 6, 12, 18). $\boldsymbol{O}_{l-1}\in\mathbb{R}^{C_B\times2N_D\times(W/4)}$ 表示第 $l-1$ 个带孔卷积的输出特征; $\boldsymbol{O}_0$ 等于场景映射特征 $\boldsymbol{B}$; $\boldsymbol{P}\in\mathbb{R}^{C_B\times2N_D\times(W/4)}$ 表示场景上下文. $Gap(\cdot)$ 表示全局平均池化. $\cup$ 表示通道拼接.

2.2 多人场景人体绝对深度预测

场景上下文缺失人体高度信息, 根据透视原理, 人体尺度因子与人体绝对深度正相关. 融合人体和场景上下文预测人体绝对深度具有可行性. 根据第 2.1.2 节的讨论, 如果将 X-Z 平面的场景上下文反映射至 X-Y 平面的人体上下文进行融合计算, 再次涉及多对一歧义性. 故采用以人体上下文卷积核解码场景上下文的融合算法. 从关键点区域特征 $\boldsymbol{E}_k$ 中逐实例动态生成人体卷积权重, 如公式 (10).

$$\Psi=\psi\left(\varphi\left(\pi\left(\bigcup_{k=0}^{K-1}Gap(\boldsymbol{E}_k)\right)\right)\right) \tag{10}$$

其中, $\pi(\cdot)$ 表示 1×1 卷积将特征通道数降维至 64, $\varphi(\cdot)$ 表示 ReLU 激活函数, $\psi(\cdot)$ 表示通过 1×1 卷积从人体上下文提取动态卷积权重 Ψ. 进而利用 Ψ 将场景上下文解码为对应人体的场景热图. 为了降低优化难度, 加快网络收敛, 引入相对位置图谱 $\boldsymbol{Q}$ 引导模型聚焦根关键点初始深度的近邻区域. 具体过程如公式 (11).

$$\boldsymbol{H}_b=\zeta(\boldsymbol{P}\cup\boldsymbol{Q};\Psi) \tag{11}$$

其中, $\boldsymbol{H}_b$ 表示鸟瞰热图. $\zeta(\cdot)$ 利用 Ψ 进行 1×1 卷积, 将人体上下文融入场景上下文. 再采用 soft-argmax 转换场景热图为深度坐标 $\hat{d}\in\mathbb{R}$ [20], 并借助公式 (8) 计算人体绝对深度 $\hat{z}$.

相对位置图谱 $\boldsymbol{Q}$ 指场景中逐像素相对人体位置 (根关键点) 的偏移量分布, 原理过程如图 8 所示. 假设图像像素 (u,v) 代表一个人体, 从深度分布 $\boldsymbol{D}$ 获取像素 (u,v) 的 N_D 个预设绝对深度概率分布 $\boldsymbol{D}(u,v)\in\mathbb{R}^{N_D}$, 则像素 (u,v) 在场景中的位置 (u,d) 计算如公式 (12).

$$(u,d)=\left(u,\ \frac{2N_D-1}{(N_D-1)}\sum_{i}^{N_D}S_i(u,v)\times i\right) \tag{12}$$

其中, $S_i(u,v)$ 表示该像素处于第 i 个预设绝对深度的概率.

本文设计融合网络融合人体和场景上下文[3,4]预测多人场景人体绝对深度, 从原理上兼顾发挥两种上下文在预测多人场景人体绝对深度中的作用, 提高人体绝对深度估计的精准性和可靠性. 从技术上回避了反映射产生的歧义.

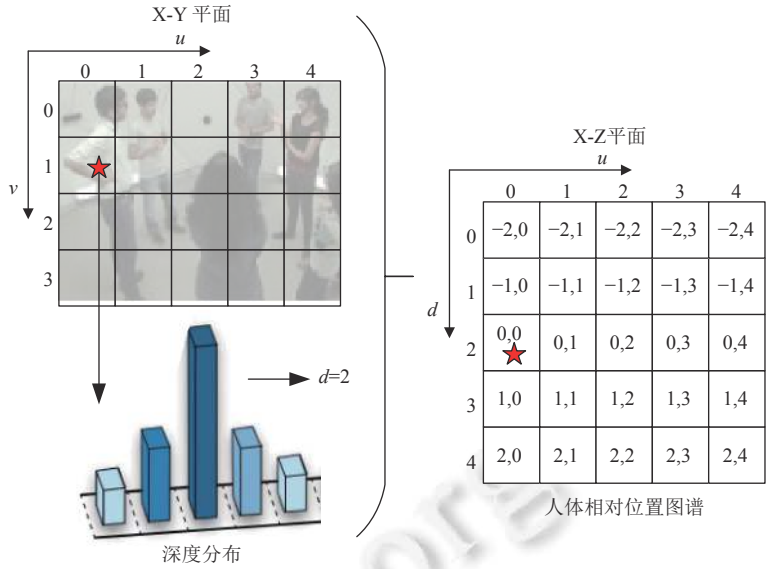


图 8　相对位置图谱计算

3 实　验

本文在提出关键点区域提议描述人体目标的基础上建模姿态关联关键点关系提取人体上下文, 提出基于图像深度建模及鸟瞰特征映射提取场景上下文, 设计融合网络融合人体和场景上下文估计人体绝对深度, 获得多人 3D 姿态估计模型 HSC-Pose (human-and-scene context based multi-person 3D pose estimation), 如图 9 所示.

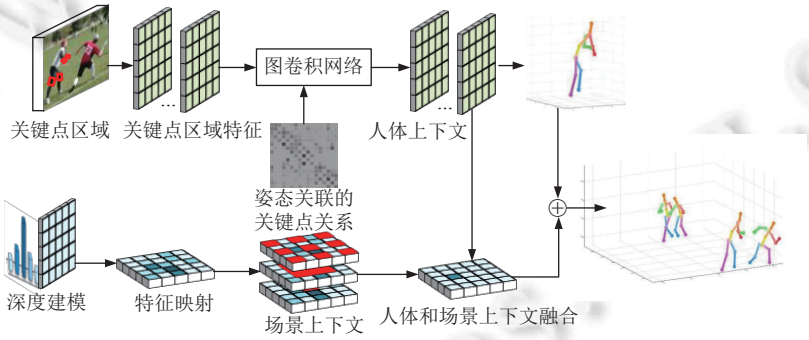


图 9　基于人体和场景上下文的多人 3D 姿态估计

3.1 创建 HSC-Pose

鉴于 Mask RCNN 的网络结构, 选择 HrNet32 作为骨干网络[25], 设计人体上下文提取网络和图像深度建模及特征映射场景上下文提取网络. 人体上下文提取网络包括人体描述模块 (human description module, HDM) 和姿态关联关键点关系模块 PKRM. HDM 涉及关键点区域提议 KRPM, 网格采样 GS 和损失函数 KLD. KRPM 的超参数: 关键点区域特征 $\boldsymbol{L}_k$ 和 $\boldsymbol{G}_k$ 的通道数 C_L 和 C_G, 高 H_L 和 H_G, 宽 W_L 和 W_G, 这些超参数对模型精度和计算复杂度的影响如图 10(a)–(d) 所示, 平衡模型性能和计算复杂度选取较优的参数集合, C_L, H_L, W_L 和 C_G, H_G, W_G 分别设置为 16, 9, 9 和 32, 25, 25. 利用 PKRM 进行多轮人体上下文提取, 如图 6(a) 所示, 提取次数对模型精度和计算复杂度的影响如图 10(e) 所示, 提取次数选为 2 效果较优.

场景上下文提取网络涉及图像深度建模及特征映射模块, 它们的超参数包括: 绝对深度个数 N_D, 图像特征 $\boldsymbol{F}$ 的通道数 C_F, 场景特征 $\boldsymbol{B}$ 的通道数目 C_B, 它们对模型精度和计算复杂度的影响如图 11 所示. 为平衡模型性能和计算复杂度, N_D, C_F 和 C_B 分别设置为 32, 16, 32.

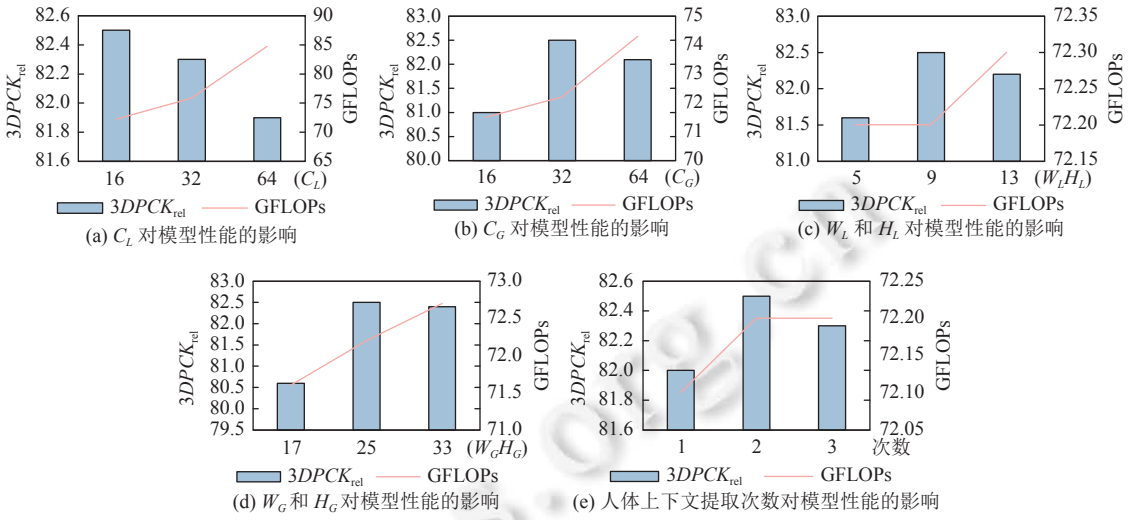


图 10 人体上下文提取网络的超参数与 HSC-Pose 性能

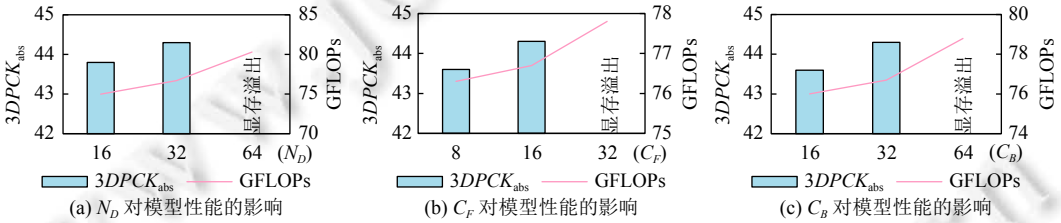


图 11 场景上下文提取网络的超参数与 HSC-Pose 性能

3.2 实验数据和评价指标

本文选择 COCO[26], MuCo-3DHP/MuPoTS-3D[27]和 Human3.6M[28]数据集评估模型性能. COCO 为 2D 人体姿态估计数据集, 训练集约 57 000 张图像. MuCo-3DHP/MuPoTS-3D 是多人 3D 姿态估计数据集, MuCo-3DHP 为训练集, MuPoTS-3D 为测试集. MuPoTS-3D 包括 20 个视频序列 (约 8 000 视频帧). Human3.6M 是室内场景 3D 人体姿态估计数据集, 包括 11 名受试者 (S1–S11) 的姿态数据. Human3.6M 训练集由 S1, S5, S6, S7 和 S8 这 5 个受试者数据组成, 测试集由 S9 和 S11 两个受试者数据组成.

MuCo-3DHP/MuPoTS-3D 的评价指标为 $3DPCK_{\text{abs}}$ 和 $3DPCK_{\text{rel}}$, 分别用于评估绝对和相对 3D 人体姿态估计的精度. 3*DPCK* (percentage of correct 3D keypoint) 计算如公式 (13).

$$3DPCK=\left(\sum_i\frac{\sum_p\phi(e_{pi}>150)}{\sum_p 1}\right)\Big/\sum_i 1 \tag{13}$$

其中, e_{pi} 表示姿态 p 中关键点 i 与对应真值间的欧氏距离 (mm), $\phi(\cdot)$ 表示克罗内克函数.

Human3.6M 的评价指标为 MPJPE (mean per joint position error), P-MPJPE (procrustes analysis MPJPE) 和 MRPE (mean root position error), MPJPE 评估相对深度下的关键点定位误差; P-MPJPE 评估相对深度下的姿态对齐关键点定位误差; MRPE 评估绝对深度下的人体根关键点定位误差. 误差指关键点真值与估值的欧氏距离, 以毫

米 (mm) 为单位.

3.3 模型训练

输入图像高度和宽度分别缩放为 512 和 832. 训练数据按 1:1 混合 COCO 中的 2D 姿态数据和 MuCo-3DHP 或 Human3.6M 中的 3D 姿态数据. 批量大小 (batch size) 设置为 24, 训练 30 个 epoch, 学习率初始值为 1E–3, 在第 10 和 20 个 epoch 时分别降至 1E–4 和 1E–5. HSC-Pose 网络整体损失函数如公式 (14), 场景上下文提取网络的损失函数 $loss_{\text{Scene}}$ 如公式 (15).

$$loss = loss_{\text{Scene}} + loss_{\text{NR}} \tag{14}$$

$$loss_{\text{Scene}} = loss_{\text{DFL}} + loss_{\text{BEV}} + 0.03loss_z \tag{15}$$

其中, 深度分布损失 $loss_{\text{DFL}}$, 如公式 (6); 鸟瞰热图损失 $loss_{\text{BEV}}$ 使用 L2 损失函数计算场景热图真值和预测值间的误差; 根关键点绝对深度损失 $loss_z$ 使用 Smooth-L1 损失函数计算根关键点绝对深度的预测误差.

人体上下文提取网络的损失函数 $loss_{\text{NR}}$, 如公式 (16).

$$loss_{\text{NR}} = loss_C + loss_{\text{heatmap}} + 0.03(loss_{\text{box}} + loss_{\text{box\_refine}} + loss_J) \tag{16}$$

其中, 中心分类图 $loss_C$ 和关键点热图 $loss_{\text{heatmap}}$ 用 L2 损失函数; 关键点包围框图谱 $loss_{\text{box}}$ 和优化关键点包围框 $loss_{\text{box\_refine}}$ 用损失函数 *KLD*, 如公式 (1); 相对深度下 3D 姿态估计 $loss_J$ 用 Smooth-L1 损失函数.

3.4 消融实验

在 MuPoTS-3D 测试集上进行消融实验, 分别评估人体上下文提取网络组件和场景上下文提取网络组件对模型 HSC-Pose 性能的影响.

3.4.1 人体上下文消融实验

消融实验分两组: HDM 的组件与 HSC-Pose 性能; PKRM 的组件与 HSC-Pose 性能. HDM 包括 KRPM, GS 和 *KLD*. 不同于使用人体包围框和 RoI Align 的基准模型, KRPM 不涉及关键点包围框真值, 无法选用常规损失函数进行消融实验; 网格 GS 自带监督学习能力; *KLD* 是本文设计的损失函数. 所以关于 HDM 的消融实验分 5 组, 见表 2. GS 实验: 较基准性能, HSC-Pose 的 $3DPCK_{\text{rel}}$ 基本一致, 因为此时仍然使用人体包围框, 包含大量背景噪声. KRPM+GS 和 KRPM+*KLD* 实验: 较基准性能, 模型 HSC-Pose 的 $3DPCK_{\text{rel}}$ 分别提高 0.4% 和 0.7%, KRPM+*KLD*+GS 实验: 较基准性能, 提高 1.0%. 实验表明 HDM 提高 HSC-Pose 的性能达 1.0%, 说明 HDM 噪声抑制效果明显.

表 2 HDM 的组件与 HSC-Pose 性能

编号	KRPM	GS	*KLD*	$3DPCK_{\text{rel}}$ (%)↑
1	—	—	—	80.2
2	—	√	—	80.3
3	√	√	—	80.6
4	√	—	√	80.9
5	√	√	√	81.2

表 3 PKRM 的关系解析组件与 HSC-Pose 性能

编号	$\boldsymbol{V}_{\text{noise}}$	$\boldsymbol{V}_{\text{geo}}$	$\boldsymbol{V}_{\text{region}}$	τ = 0.35	τ = 0.50	τ = 0.75	Linear	$\boldsymbol{R}_{sp}^{dy}$	$3DPCK_{\text{rel}}$ (%)↑
1	—	—	—	—	—	—	—	—	79.1
2	√	—	—	—	—	—	—	—	79.9
3	√	√	—	—	—	—	—	—	80.2
4	√	√	√	—	—	—	—	—	81.1
5	√	√	√	√	—	—	—	—	81.7
6	√	√	√	—	√	—	—	—	82.1
7	√	√	√	—	—	√	—	—	81.5
8	√	√	√	—	—	—	√	—	82.2
9	√	√	√	—	—	—	—	√	82.5

PKRM 的建模组件含建模姿态关联的关键点关系和剔除弱连接关键点关系; 建模姿态关联的关键点关系的动态关系矩阵涉及噪声特征 $\boldsymbol{V}_{\text{noise}}$, 人体姿态几何特征 $\boldsymbol{V}_{\text{geo}}$ 和关键点区域特征 $\boldsymbol{V}_{\text{region}}$, 讨论这些成分对 HSC-Pose 性能的影响, 实验设计依次加入 $\boldsymbol{V}_{\text{noise}}$, $\boldsymbol{V}_{\text{geo}}$ 和 $\boldsymbol{V}_{\text{region}}$, 见表 3. 较之基准性能 (使用静态关系矩阵[11], 实验 1), 实验 2–4 的结果表明 HSC-Pose 的 $3DPCK_{\text{rel}}$ 分别提高 0.8% (实验 2), 1.1% (实验 3) 和 2.0% (实验 4), 说明 $\boldsymbol{V}_{\text{noise}}$, $\boldsymbol{V}_{\text{geo}}$ 和

$\boldsymbol{V}_{\text{region}}$ 对解析关键点关系效果明显.

采用不同阈值 τ 过滤关键点关系, 线性组合 3 种稀疏度 (τ=0.35, 0.50 和 0.75) 生成动态稀疏关系矩阵. 分析 τ 取不同阈值 (τ=0.35, 0.50 和 0.75) 及动态稀疏关系矩阵 $\boldsymbol{R}_{sp}^{dy}$ 对 HSC-Pose 性能的影响, 设计实验 5–实验 8, 如表 3. 较之实验 4, 实验 5–实验 7 的实验结果表明 HSC-Pose 的 $3DPCK_{\text{rel}}$ 分别提高 0.6% (实验 5), 1.0% (实验 6) 和 0.4% (实验 7), 说明过滤关键点关系减少噪声传播效果明显. 较之实验 5, 实验 8 (采用线性层组合 3 种阈值的稀疏关系矩阵) 并未展现出性能的显著提升, 因为这种组合权重等价于采用单一固定阈值, 相当于姿态共享. 较之实验 8, 实验 9 的结果表明, HSC-Pose 的 $3DPCK_{\text{rel}}$ 至少提高 0.3%, 说明 $\boldsymbol{R}_{sp}^{dy}$ 通过动态组合权重优化关键点关系稀疏度的方式效果明显.

姿态关联关键点关系描述模块 PKRM 的姿态估计组件涉及关键点包围框优化, 动态卷积和级联网络, 为此设计 5 个实验, 见表 4. 关于关键点包围框优化和动态卷积, 较之基准模型, HSC-Pose 的 $3DPCK_{\text{rel}}$ 分别提高 0.8% (实验 2) 和 0.4% (实验 3). 两者组合提高 1.3% (实验 4). 这说明利用人体上下文优化关键点包围框有利于精确描述关键点区域提议; 动态卷积利用人体上下文抑制包围框内残存噪声. 经级联优化的实验 (见实验 5), 较之实验 4, HSC-Pose 的 $3DPCK_{\text{rel}}$ 提高 2.6%.

表 4 PKRM 的姿态估计组件与 HSC-Pose 性能

编号	关键点包围框优化	动态卷积	级联优化	$3DPCK_{\text{rel}}$ (%)↑
1	—	—	—	81.2
2	√	—	—	82.0
3	—	√	—	81.6
4	√	√	—	82.5
5	√	√	√	85.1

3.4.2 场景上下文消融实验

场景上下文提取网络包括 5 个组件: 深度分布 DD, 瀑布带孔金字塔 WASP, 动态卷积 DC, 相对位置图谱 RP 和损失函数 DFL. 其中, DD, DFL 和 WASP 用以提取场景上下文. DC 和 RP 用以融合人体和场景上下文.

验证 DD, DFL 和 WASP 的有效性, 设计 DD, DD+DFL 和 DD+DFL+WASP 等实验, 如表 5 的实验 2–实验 4. 较基准模型, 实验 2–实验 4 的结果表明 HSC-Pose 的 $3DPCK_{\text{abs}}$ 分别提高 1.6% (实验 2), 2.2% (实验 3) 和 2.9% (实验 4). 这说明将 DD 引入深度维度表达离散绝对深度的方法, 将深度提取归为分类问题, 降低模型的优化难度. DFL 能有效监督绝对深度期望值接近真值. WASP 扩大特征感受野提取场景上下文对缓解绝对深度歧义有积极作用.

表 5 场景上下文提取网络组件与 HSC-Pose 性能

编号	视角	DD	DFL	WASP	DC	RP	$3DPCK_{\text{abs}}$ (%)↑
1	X-Y	—	—	—	—	—	39.7
2	X-Y	√	—	—	—	—	41.3
3	X-Y	√	√	—	—	—	41.9
4	X-Y	√	√	√	—	—	42.6
5	X-Z	√	√	√	—	—	43.0
6	X-Z	√	√	√	√	—	43.4
7	X-Z	√	√	√	√	√	44.3
8	X-Z	—	—	—	√	√	38.4
9	X-Z	√	—	—	√	√	41.6
10	X-Z	√	√	—	√	√	42.8

验证 DD, DFL 和 WASP 对提取 X-Z 平面的场景上下文的作用, 以及 DC 和 RP 在融合人体和场景上下文中的作用, 设计实验 5–实验 7, 见表 5. 较之实验 4, 实验 5 (DD+DFL+WASP) 的结果表明, HSC-Pose 的 $3DPCK_{\text{abs}}$ 提高 0.4%, 说明在 X-Z 平面提取场景上下文对提高模型性能效果更明显. 较之实验 5, 实验 6 (仅 DC 进行人体和场

景上下文融合) 的结果表明 HSC-Pose 的 $3DPCK_{abs}$ 提高 0.4%, 而实验 7 (利用 DC+RP 进行上下文融合), 比实验 6 提高 0.9%. DC 融合人体和场景上下文, 利于缓解绝对深度歧义. 而 RP 在上下文融合过程中更关注人体区域, 不仅降低了优化难度, 还能增强效果.

图像特征映射至场景依赖 DD 和 DFL, 讨论 DD 和 DFL 在特征映射过程中的作用, 设计实验 8–实验 10, 见表 5. 较之实验 1, 实验 8 结果表明 HSC-Pose 的 $3DPCK_{abs}$ 降低 1.3%. 说明不采用 DD 和 DFL 很难应对“图像至场景”特征映射存在的多对一歧义. 较实验 8, 实验 9 和实验 10 (DD 和 DD+DFL) 的结果表明: HSC-Pose 的 $3DPCK_{abs}$ 分别提高 3.2% (实验 9) 和 4.4% (实验 10), DD 和 DFL 对缓解特征映射歧义效果明显, 并能促进绝对深度期望值接近真值.

3.5 测试集 MuPoTS-3D 上的对比实验

本文特色在于: (1) 采用关键点区域提议替代人体包围框, 提取高信噪比人体上下文; (2) 从鸟瞰视角提取场景上下文, 获得三维空间下的人体位置布局. 通过融合人体和场景上下文可靠预测人体绝对深度. 参与实验的 3 类先进方法 (state-of-the-arts) 包括: 自顶向下, 自底向上和联合自顶向下和自底向上的方法. 自顶向下方法所选择的代表工作包括文献 [2,3,10]. 自底向上方法所选代表工作包括文献 [4,6]. 综合自顶向下和自底向上方法所选择的代表性工作包括文献 [7,8,29].

在 MuPoTS-3D 上进行同类方法比较, 实验结果见表 6. 配置“matched people”评估与真值匹配的结果. 较文献 [3,4], HSC-Pose 的 $3DPCK_{rel}$ 和 $3DPCK_{abs}$ 至少提高 3.4% 和 6.3%. 较文献 [2], HSC-Pose 的 $3DPCK_{abs}$ 提高 2.7%. 各模型性能上的差异归咎于不同的上下文提取方法, 文献 [2,3] 都基于人体上下文, 但是, 文献 [2] 在文献 [3] 基础上补充了人体姿态信息, HSC-Pose 包括场景上下文和带噪声抑制的人体上下文. 较文献 [10], HSC-Pose 的 $3DPCK_{rel}$ 和 $3DPCK_{abs}$ 分别提高 2.2% 和 9.8%. 文献 [3,10] 都基于包围框提取人体上下文, 且未提取场景上下文. 较文献 [29], HSC-Pose 的 $3DPCK_{abs}$ 低 2%, 但文献 [29] 因合成数据计算开销更大.

表 6 MuPoTS-3D 上的对比实验结果

文献	GFLOPs	Matched people		All people	
		$3DPCK_{rel}$ (%)↑	$3DPCK_{abs}$ (%)↑	$3DPCK_{rel}$ (%)↑	$3DPCK_{abs}$ (%)↑
[5]	—	74.2	—	71.3	—
[3]	603.5	82.5	31.8	81.8	31.5
[10]	—	83.7	35.2	—	—
[4]	197.7	80.5	38.7	73.5	35.4
[8]	320.2	—	—	82.0	43.8
[2]	—	83.5	42.3	82.5	39.2
[6]	—	—	—	82.7	39.2
[7]	—	—	—	89.6	48.0
[29]	220	—	47.0	—	44.0
HSC-Pose	76.8	85.9	45.0	85.1	44.3

配置“all people”: 较之文献 [3,4], HSC-Pose 的 $3DPCK_{rel}$ 和 $3DPCK_{abs}$ 提高 3.3% 和 8.9%; 较之文献 [8], HSC-Pose 的 $3DPCK_{rel}$ 和 $3DPCK_{abs}$ 提高 3.1% 和 0.5%; 较之文献 [7], HSC-Pose 的 $3DPCK_{rel}$ 和 $3DPCK_{abs}$ 均明显降低, Cheng 等人[7]利用时序提取上下文的方法值得探究. 关于计算复杂度 (GFLOPs), HSC-Pose 优势显著.

3.6 Human3.6M 测试集上的对比实验

在 Human3.6M 测试集上评估模型的 MPJPE, P-MPJPE 和 MRPE, 见表 7. 较之单人 3D 姿态估计方法, HSC-Pose 的 P-MPJPE 至少降低 1.8 mm. HSC-Pose 的 MPJPE 高于文献 [30] 0.1 mm, 低于多数同类方法. 较之多人 3D 姿态估计方法, HSC-Pose 的 MPJPE 和 P-MPJPE 分别至少降低 1.2 mm 和 0.4 mm, 绝对深度指标 MRPE 至少降低 4.2 mm, 说明 HSC-Pose 对缓解深度歧义效果显著.

表 7 在 Human3.6M 测试集的对比实验结果

方法 (mm)↓	单人姿态估计						多人姿态估计				
	[20]	[31]	[30]	[32]	[33]	[1]	[3]	[2]	[10]	[8]	HSC-Pose
MPJPE	49.6	48.6	47.3	49.5	48.6	57.0	54.4	52.7	—	48.6	47.4
P-MPJPE	35.7	—	31.9	33.4	42.1	—	35.2	33.8	—	30.5	30.1
MRPE	—	—	—	—	—	—	120	95.7	77.6	—	73.4

3.7 定性分析

以下可视化评估 SMAP[4], CAD[3], DAS[6]和 HSC-Pose 等模型的多人 3D 姿态估计性能, 如图 12 所示.

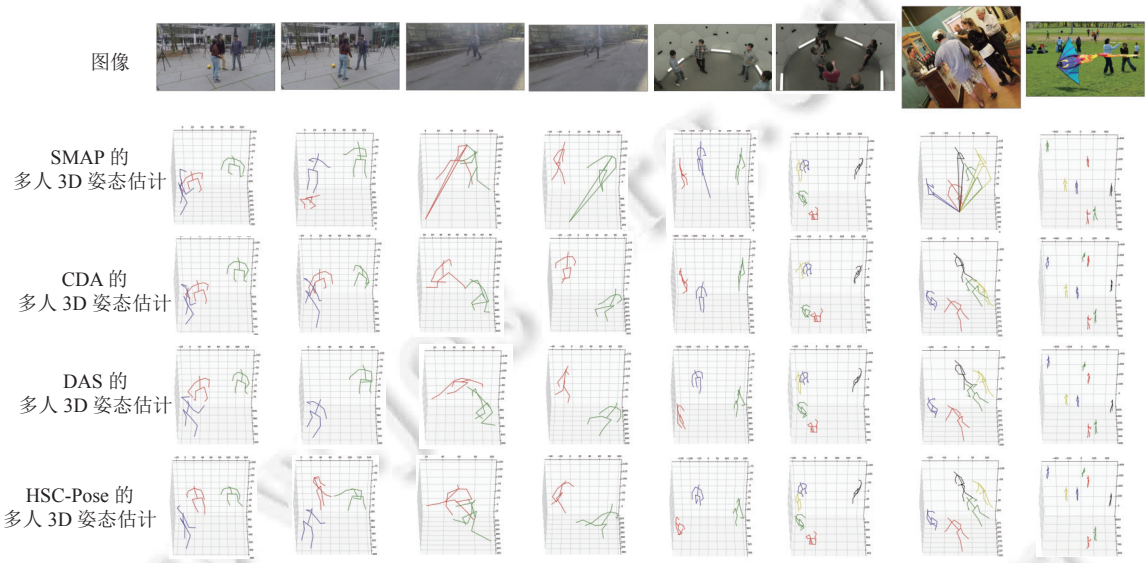


图 12 可视化评价多种场景下的 HSC-Pose 性能

为了分析模型的鲁棒性, 从 MuPoTS-3D 测试集中选取人体遮挡和背景噪声显著的图片作为测试数据, 可视化实验结果, 如图 12 第 1–4 列. 第 1 行自左至右第 1 张图像: 背景噪声较少, 所有方法的结果相似. 第 2 张图像: 红色人体目标的人体中心被严重遮挡, SMAP 的绝对深度估计误差明显, DAS 出现漏检. HSC-Pose 的绝对深度估计显优, 应对人体遮挡的效果更鲁棒.

为了分析模型应对视角变化的鲁棒性, 挑选与训练图像视角差异明显的数据作为测试数据, 可视化实验结果如图 12 第 5–6 列. 对应第 1 行第 5–6 张图像, CAD 估计的绝对深度估计误差显著, 因其仅基于人体包围框尺度计算绝对深度. 较其他方法, HSC-Pose 仍获得了较合理的人体位置, 但是估计精度有所下降. 虽然未限定图像视角, 但是受制于数据多样性不足 HSC-Pose 的视角泛化能力有限.

为了分析模型应对复杂场景的鲁棒性, 从 COCO 验证集挑选复杂场景图像作为测试数据, 如第 7–8 列所示. 对应第 1 行第 7–8 张图像, HSC-Pose 和 CAD 的精度优于 SMAP 和 CAD. HSC-Pose 在人体姿态细节上优于 DAS, 例如第 7 张图像的红色人体: DAS 仅关注图像全局上下文, HSC-Pose 兼及人体和场景上下文, 而且 HSC-Pose 恢复人体姿态细节的效果明显. 但是, 包括 HSC-Pose 在内的所有参比方法, 对较远人体的姿态估计精度都显著下降, 一方面人体过小信息量不足; 另一方面, 目前训练数据中缺乏远距离 3D 人体姿态标注.

4 总 结

本文设计的自顶向下网络分支采用“关键点区域提议”替代人体包围框描述人体目标, 并兼及背景噪声、边缘、肢体朝向等信息优化关键点区域特征描述, 进而建模姿态关联的动态稀疏关键点关系, 剔除弱连接关键点关系并阻断

噪声传播, 提高模型的相对姿态恢复能力. 自底向上分支从鸟瞰平面而并非图像平面提取场景上下文获得仿三维空间人体位置布局, 通过融合人体和场景上下文可靠预测人体绝对深度. 在 MuPoTS-3D 和 Human3.6M 数据集上进行了广泛的对比实验, 结果表明: 较当前先进模型, 本文多人场景 3D 姿态估计模型 HSC-Pose 性能更优, 而且计算复杂度明显降低. 但是本文模型 HSC-Pose 的可预测深度目前限于 100 m, 下一步工作将增强预测更远目标姿态的模型泛化能力. 此外, 本文工作目前限于单帧图像, 今后将着力基于多视角和多帧图像的多人场景 3D 人体姿态估计研究.

References:

[1] Yang B, Li HP, Zeng H. Three-dimensional human pose estimation based on video. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2463–2469 (in Chinese with English abstract). [doi: 10.13700/j.bh.1001-5965.2019.0384]

[2] Guo Y, Ma LC, Li Z, Wang X, Wang F. Monocular 3D multi-person pose estimation via predicting factorized correction factors. Computer Vision and Image Understanding, 2021, 213: 103278. [doi: 10.1016/j.cviu.2021.103278]

[3] Moon G, Chang JY, Lee KM. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 10132–10141. [doi: 10.1109/iccv.2019.01023]

[4] Zhen JN, Fang Q, Sun JM, Liu WT, Jiang W, Bao HJ, Zhou XW. SMAP: Single-shot multi-person absolute 3D pose estimation. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 550–566. [doi: 10.1007/978-3-030-58555-6_33]

[5] Dabral R, Gundavarapu NB, Mitra R, Sharma A, Ramakrishnan G, Jain A. Multi-person 3D human pose estimation from monocular images. In: Proc. of the 2019 Int'l Conf. on 3D Vision. Los Alamitos: IEEE Computer Society, 2019. 405–414. [doi: 10.1109/3dv.2019.00052]

[6] Wang ZT, Nie XC, Qu XC, Chen YP, Liu S. Distribution-aware single-stage models for multi-person 3D pose estimation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 13086–13195. [doi: 10.1109/CVPR52688.2022.01275]

[7] Cheng Y, Wang B, Yang B, Tan RT. Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 7645–7655. [doi: 10.1109/cvpr46437.2021.00756]

[8] Wang C, Li JF, Liu WT, Qian C, Lu C. HMOR: Hierarchical multi-person ordinal relations for monocular multi-person 3D pose estimation. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 242–259. [doi: 10.1007/978-3-030-58580-8_15]

[9] Tian Z, Chen H, Shen CH. DirectPose: Direct end-to-end multi-person pose estimation. arXiv:1911.07451, 2019.

[10] Lin JH, Lee GH. HDNet: Human depth estimation for multi-person camera-space localization. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 633–648. [doi: 10.1007/978-3-030-58523-5_37]

[11] Zou ZM, Tang W. Modulated graph convolutional network for 3D human pose estimation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 11457–11467. [doi: 10.1109/ICCV48922.2021.01128]

[12] Reading C, Harakeh A, Chae J, Waslander SL. Categorical depth distribution network for monocular 3D object detection. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8551–8560. [doi: 10.1109/cvpr46437.2021.00845]

[13] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/cvpr.2016.90]

[14] Geng ZG, Sun K, Xiao B, Zhang ZX, Wang JD. Bottom-up human pose estimation via disentangled keypoint regression. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14671–14681. [doi: 10.1109/cvpr46437.2021.01444]

[15] Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT, 2015. 2017–2025.

[16] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/iccv.2017.322]

[17] Xu XX, Zou Q, Lin X. Adaptive hypergraph neural network for multi-person pose estimation. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI, 2022. 2955–2963. [doi: 10.1609/aaai.v36i3.20201]

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/cvpr.2018.00745]

[19] Mao WA, Tian Z, Wang XL, Shen CH. FCPose: Fully convolutional multi-person pose estimation with dynamic instance-aware

convolutions. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9030–9093. [doi: 10.1109/cvpr46437.2021.00892]

[20] Sun X, Xiao B, Wei FY, Liang S, Wei YC. Integral human pose regression. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 536–553. [doi: 10.1007/978-3-030-01231-1_33]

[21] Li SC, Ke L, Pratama K, Tai YW, Tang CK, Cheng KT. Cascaded deep monocular 3D human pose estimation with evolutionary training data. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6172–6188. [doi: 10.1109/cvpr42600.2020.00621]

[22] Horaud R, Hansard M, Evangelidis G, Menier C. An overview of depth cameras and range scanners based on time-of-flight technologies. arXiv:2012.06772, 2020.

[23] Li X, Wang WH, Wu LJ, Chen S, Hu XL, Li J, Tang JH, Yang J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1763.

[24] Artacho B, Savakis A. UniPose: Unified human pose estimation in single images and videos. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7033–7042. [doi: 10.1109/cvpr42600.2020.00706]

[25] Sun K, Xiao B, Liu D, Wang JD. Deep high-resolution representation learning for human pose estimation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5686–5796. [doi: 10.1109/cvpr.2019.00584]

[26] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[27] Mehta D, Sotnychenko O, Mueller F, Xu WP, Sridhar S, Pons-Moll G, Theobalt C. Single-shot multi-person 3D pose estimation from monocular RGB. In: Proc. of the 2018 Int'l Conf. on 3D Vision. Verona: IEEE, 2018. 120–130. [doi: 10.1109/3dv.2018.00024]

[28] Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325–1339. [doi: 10.1109/tpami.2013.248]

[29] Su JJ, Wang CY, Ma XX, Zeng WJ, Wang YZ. VirtualPose: Learning generalizable 3D human pose models from virtual data. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 55–71. [doi: 10.1007/978-3-031-20068-7_4]

[30] Chen ZR, Huang Y, Yu HY, Xue B, Han K, Guo YR, Wang L. Towards part-aware monocular 3D human pose estimation: An architecture search approach. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 715–732. [doi: 10.1007/978-3-030-58580-8_42]

[31] Li JF, Bian SY, Zeng AL, Wang C, Pang B, Liu WT, Li C. Human pose regression with residual log-likelihood estimation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 11005–11014. [doi: 10.1109/iccv48922.2021.01084]

[32] Zheng XT, Chen XM, Lu XQ. A joint relationship aware neural network for single-image 3D human pose estimation. IEEE Trans. on Image Processing, 2020, 29: 4747–4758. [doi: 10.1109/tip.2020.2972104]

[33] Xia HL, Zhang TT. Self-attention network for human pose estimation. Applied Sciences, 2021, 11(4): 1826. [doi: 10.3390/app11041826]

附中文参考文献:

[1] 杨彬, 李和平, 曾慧. 基于视频的三维人体姿态估计. 北京航空航天大学学报, 2019, 45(12): 2463–2469. [doi: 10.13700/j.bh.1001-5965.2019.0384]

何建航(1996－), 男, 硕士生, 主要研究领域为 2D 和 3D 人体姿态估计.

刘琼(1959－), 女, 博士, 教授, 博士生导师, 主要研究领域为机器学习, 深度学习视觉应用技术.

孙郡瑶(1997－), 女, 博士生, 主要研究领域为密集人体姿态估计.