Temporal Feature Alignment and Mutual Information Maximization for Video-Based Human Pose Estimation

摘要

多帧人体姿态估计一直是计算机视觉中一个引人注目的基本问题。由于视频中经常发生快速运动和姿势遮挡，因此该任务具有挑战性。现有技术的方法努力结合来自相邻帧（支持帧）的附加视觉证据以促进当前帧（关键帧）的姿态估计。到目前为止，已经消除的一个方面是当前方法直接跨帧聚合未对齐的上下文的事实。当前帧和相邻帧的姿态特征之间的空间未对准可能导致不令人满意的结果。更重要的是，现有的方法建立在简单的姿态估计损失的基础上，不幸的是，这不能限制网络充分利用来自相邻帧的有用信息。为了解决这些问题，我们提出了一种新的分层对齐框架，该框架利用粗到细的变形来逐步更新相邻帧，以在特征级别上与当前帧对齐。我们还建议明确监督从相邻帧中提取的知识，保证提取有用的互补线索。为了实现这一目标，我们从理论上分析了帧之间的互信息，并得出了最大化任务相关的互信息的损失。

关键词：人体姿态估计；视频姿态估计

1 引言

在仔细检查和实验现有方法的已发布实现 [3,7] 后，我们观察到它们在具有挑战性的情况下（如快速运动和姿势遮挡）会出现性能恶化。在快速运动场景中，现有方法由于运动模糊而难以识别左手腕。我们推测原因是双重的。(1) 常见的是，当前帧和相邻帧中的同一个人没有很好地对准，特别是对于涉及人类主体或相机的快速运动的情况。然而，现有的方法倾向于直接从相邻帧中聚集未对齐的上下文，这些空间未对齐的特征潜在地降低了模型的性能。(2) 现有技术的方法简单地采用传统的 MSE（关节的均方误差）损失来监督姿势热图的学习，同时缺乏对保证来自相邻帧的信息增益的有效约束以及在中间特征级别的监督。

在本文中，我们提出了一个新的框架，沿着理论分析，以应对上述挑战。所提出的方法，称为 FAMI-Pose（特征对齐和互信息最大化的姿态估计），由两个关键组成部分。(i)FAMI-Pose 进行由粗到细的变形，系统地更新相邻帧，以在特征级别与当前帧对齐。具体地，FAMI-Pose 首先执行全局变换，其整体地重新排列相邻帧特征以初步校正空间移位或抖动。随后，利用局部校准来自适应地移动和调制相邻帧特征的每个像素以增强特征对准。(ii)FAMIPose 进一步将信息理论目标作为特征级的额外中间监督。最大化这个互信息目标使我们的模型能够充

分挖掘相邻帧中的任务相关线索，提取有目的的互补知识来增强关键帧上的姿态估计。这项工作的贡献总结为：

- 我们提出从通过特征对齐有效利用时间上下文的角度来检查多帧人体姿势估计任务。
- 为了明确地监督从相邻帧中提取的知识，我们提出了一个信息论损失函数，它允许最大化从支持帧中挖掘的任务相关线索。

2 相关工作

在本节中，我们简要回顾了以下三个与我们的工作密切相关的主题，即基于图像的人体姿态估计，基于视频的人体姿态估计和特征对齐。

2.1 基于图像的人体姿态估计

基于图像的人体姿势估计的传统解决方案利用图像结构 [18, 21] 来对身体关节之间的空间关系进行建模。这些方法往往依赖于手工制作的功能，并具有有限的代表性能力。受深度学习 [8] 的爆炸式增长以及 PoseTrack [1, 11] 和 COCO [13] 等大规模姿态估计数据集的可用性的推动，已经提出了各种深度学习方法 [2, 6]。这些方法可以大致分为两种范式：自下而上和自上而下。自下而上的方法首先检测单个身体部位，然后将这些检测到的组成部分组装成整个人。[4] 提出了一种双卷积结构来模拟预测部分置信度图和部分亲和度字段（表示身体部分之间的关系）。另一方面，自上而下的方法首先检测人体边界框，然后估计每个边界框内的人体姿势。[20] 利用反卷积层来取代常用的双线性插值，用于特征图的空间上采样。[19] 中的一项最新工作提出了一种高分辨率网络（HRNet），它在整个推理过程中保留了高分辨率特征图，在多个基于图像的基准测试中获得了最先进的结果。

2.2 基于视频的人体姿态估计

针对基于图像的数据训练的姿态估计模型不能很好地推广到视频序列，因为它们不能从相邻帧中包含丰富的线索。为了跨帧建模和利用时间上下文，一种直接的方法是采用卷积 LSTM，如 [2] 中所提出的。这种模型的一个关键缺点可能是它们倾向于在不同的框架之间不对齐特征，这不利地降低了支持框架的效力。[17] 通过计算连续帧之间的光流来显式地估计运动场，并且这些运动线索随后用于对准姿态热图。[14] 估计关键帧和支持帧之间的运动偏移，并且这些偏移提供了在连续帧上执行姿态热图的重新缩放的基础。在这两种情况下，姿态估计精度将严重依赖于光流或运动偏移估计的性能。此外，这些方法在中间特征水平上缺乏有效的监督可能导致不准确的姿态估计。

2.3 特征对齐

特征对准是许多计算机视觉任务的重要主题（例如，语义分割 [12, 16]、对象检测 [5]），并且最近已经做出了许多努力来解决这个问题。[15] 提出了一个 indexguided 框架，采用索引来指导池和上采样。[9] 建议学习像素的变换偏移以对齐上采样的特征图。[10] 提出了一个对齐特征聚合模块，用于对齐多个不同分辨率的特征，以实现更好的聚合。虽然以前的方法主要解决网络输入和输出之间的空间不对准，但我们关注的是时间（即，跨帧）特征对齐。

3 本文方法

3.1 本文方法概述

方法概述 框架的概述如图 1所示。对于每个支持帧 $I^i_{t+\delta}$，FAMI-pose 执行两阶段分层转换以在特征级别将 $I^i_{t+\delta}$ 与关键帧 I^i_t 对齐。具体来说，FAMI-pose 由两个主要模块组成，即全局变换模块和局部校准模块。首先对 I^i_t 和 $I^i_{t+\delta}$ 进行特征提取，分别得到 Z^i_t 和 $Z^i_{t+\delta}$。然后，这些特征被送入我们的全局变换模块，该模块学习仿射变换的参数，以获得粗对齐的支持框架特征 $\bar{Z}^i_{t+\delta}$。然后，Z^i_t 和 $\bar{Z}^i_{t+\delta}$ 被传递给本地校准模块，该模块执行像素方向的变形以产生精确对齐的特征 $\bar{\bar{Z}}^i_{t+\delta}$。最后，我们聚合所有对齐的支持帧特征 $\{\bar{\bar{Z}}^i_{t+\delta} \mid \delta \in N\}$ 和关键帧特征 Z^i_t 以获得我们的增强特征 $\tilde{Z}^i_{t+\delta}$。$\tilde{Z}^i_{t+\delta}$ 被传递到检测头，该检测头输出姿势估计 $\hat{H}^i_t$。任务目标是最小化热图估计损失 $\mathcal{L}_h$，该损失衡量 $\hat{H}^i_t$ 与地面实况 H^i_t 之间的差异。在此基础上，我们还设计了一个互信息目标线性矩阵不等式 $\mathcal{L}_{MI}$，它实现了特征级别的监督，以最大化 $\tilde{Z}^i_t$ 中编码的互补任务相关信息量。

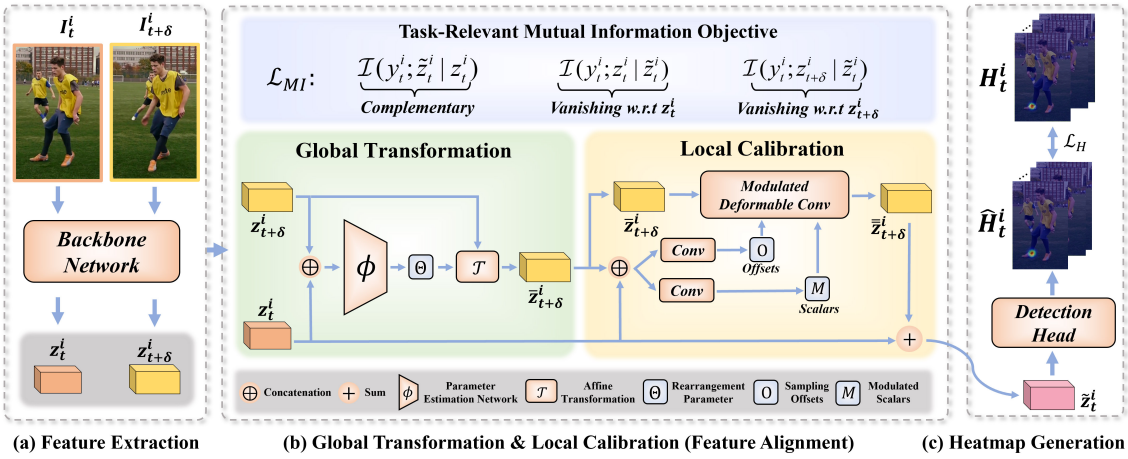

图 1. 我们的 FAMI-Pose 框架的整体管道。目标是在支持帧的帮助下检测关键帧 I^i_t 中的人 i 的姿势。为了说明清楚，我们在该图中仅示出了单个支撑框架 $I^i_{t+\delta}$。我们首先提取它们各自的特征 Z^i_t 和 $Z^i_{t+\delta}$。然后将这些特征交给我们的全局变换模块和局部校准模块进行时间对齐。所有支持帧的关键帧特征 Z^i_t 和对齐特征 $\bar{\bar{Z}}^i_{t+\delta}$ 被聚合为 $\tilde{Z}^i_{t+\delta}$，其被传递到输出姿态估计 $\hat{H}^i_t$ 的检测头。除了测量 $\hat{H}^i_t$ 和地面真实值 H^i_t 之间差异的热图估计损失 $\mathcal{L}_h$ 之外，我们通过我们的互信息目标 $\mathcal{L}_{MI}$ 引入了额外的特征级监督，以从支持框架中提取最大的任务相关补充信息。

3.2 特征对齐

特征对齐从特征提取开始，这是用 HRNet-W48 网络 [19]（基于图像的人体姿势估计的最先进方法）作为骨干完成的。然后，使提取的特征 Z^i_t 和 $Z^i_{t+\delta}$ 通过全局变换模块和局部校准模块，以逐渐将 $Z^i_{t+\delta}$ 与 Z^i_t 对准。我们想强调的是，我们不追求图像级对齐，而是驱动网络学习支持框架和关键框架之间的特征级对齐。

全局变换 我们观察到，大多数失败的情况下，在视频中的姿态估计发生由于快速移动的人或相机，这不可避免地导致大的空间偏移或相邻帧之间的抖动。为了将支撑框架与关键框架对齐，我们设计了一个全局转换模块（GTM）。GTM 计算全局仿射变换的空间重排参数，以获得支持帧特征 $Z^i_{t+\delta}$ 与关键帧特征 Z^i_t 的粗略初步对准。

局部校准　全局变换模块产生粗略的对齐。然后，我们设计本地校准模块（LCM），以在像素级执行细致的微调，从而产生精细对齐的特征 $\bar{\bar{Z}}^i_{t+\delta}$。

具体来说，给定 $\bar{Z}^i_{t+\delta}$ 和 Z^i_t，我们独立地估计特征 $\bar{Z}^i_{t+\delta}$ 的卷积核采样偏移 $\mathcal{O}$ 和调制标量 $\mathcal{M}$。

自适应学习的内核偏移 $\mathcal{O}$ 和调制标量 $\mathcal{M}$ 分别对应于相对于关键帧特征 $\bar{Z}^i_{t+\delta}$ 的每个像素 Z^i_t 的位置偏移和强度波动。

随后，我们通过调制可变形卷积实现局部校准操作 [22]。给定初步对齐的特征 $\bar{Z}^i_{t+\delta}$ 内核偏移 $\mathcal{O}$ 和调制标量 $\mathcal{M}$ 作为输入，调制可变形卷积输出微调特征 $\bar{\bar{Z}}^i_{t+\delta}$。

为了预测对互信息损失的讨论，我们想要指出的是，关键帧特征 Z^i_t 仅用于计算 GTM 中的全局变换参数和 LCM 中的卷积参数。其信息不会传播到最终对齐的支撑框架特征 $\bar{\bar{Z}}^i_{t+\delta}$ 中。

3.3　互信息目标

我们当然可以直接以端到端的方式训练 FAMI-Pose，并丢失姿势热图。考虑到我们对提取用于姿态估计的时间特征的系统检查，调查在特征级引入监督是否会促进任务将是富有成效的。简单地说，我们可以将特征级目标公式化为支持帧特征 $Z^i_{t+\delta}$ 和关键帧特征 Z^i_t 之间的 $L1$ 或 $L2$ 差。然而，这样的刚性对齐很可能会导致从支持框架中侵蚀补充的特定任务信息。因此，如此优化的时间特征将不足以提供相关的支持信息以促进姿态估计。因此，我们必须强调来自支持框架的有目的的补充信息。为此，受 [21，72] 的启发，我们提出了一个互信息目标，该目标旨在最大限度地提高增强特征 $\tilde{Z}^i_{t+\delta}$ 中与任务相关的互补信息量。

互信息　互信息（MI）是随机变量之间共享的信息量的度量。形式上，MI 量化了两个随机变量 v_1 和 v_2 的统计依赖性：

$$\mathcal{I}(v_1; v_2) = \mathbb{E}_{p(v_1,v_2)}[log\frac{p(v_1, v_2)}{p(v_1)(v_2)}],$$

其中 $p(v_1, v_2)$ 是 v_1 和 v_2 之间的联合概率分布，而 $p(v_1)$ 和 $p(v_2)$ 是它们的边缘。

互信息损失　在这个框架内，我们学习有效时间特征对齐的主要目标可以表述为：

$$max\ \mathcal{I}(y^i_t; \tilde{Z}^i_t \mid Z^i_t),$$

其中,y^i_t 表示标签，并且 $\mathcal{I}(y^i_t; \tilde{Z}^i_t \mid Z^i_t)$ 表示增强特征 $\tilde{Z}^i_t$ 中的任务相关信息的量，补充于（即，排除）来自关键帧特征 Z^i_t 的信息。直观地说，优化这一目标将最大限度地提高我们试图从相邻帧中提取的额外相关和补充信息，以支持姿态估计任务。

此外，我们引入了两个正则化项来减轻信息丢失：

$$min\ [\mathcal{I}(y^i_t; Z^i_{t+\delta} \mid \tilde{Z}^i_t)\ +\ \mathcal{I}(y^i_t; Z^i_t \mid \tilde{Z}^i_t)].$$

在特征对齐期间，$\mathcal{I}(y^i_t; Z^i_{t+\delta} \mid \tilde{Z}^i_t)$ 和 $\mathcal{I}(y^i_t; Z^i_t \mid \tilde{Z}^i_t)$ 分别测量 $Z^i_{t+\delta}$ 和 Z^i_t 中消失的任务相关信息。它们有助于信息的非破坏性传播。同时最小化这两个项将防止在 $Z^i_{t+\delta}$ 和 Z^i_t 中的过度信息损失，同时最大化主要互补任务相关互信息目标。

4 复现细节

4.1 与已有开源代码对比

我在论文源代码基础上，添加以下 SimCC 模块代码。SimCC 的核心思想是将人体姿态估计视为垂直和水平坐标的两个分类任务，并通过将每个像素划分为多个 bin 来减少量化误差。如图 2所示出了由骨干网络和两个分类器头组成的 SimCC 的示意图。

```
        # head检测头
        if self.coord_representation == 'simdr' or
                self.coord_representation == 'sa-simdr':
        self.mlp_head_x = nn.Linear(cfg.MODEL.HEAD_INPUT,
         int(cfg.MODEL.IMAGE_SIZE[0]*cfg.MODEL.SIMDR_SPLIT_RATIO))
        self.mlp_head_y = nn.Linear(cfg.MODEL.HEAD_INPUT,
         int(cfg.MODEL.IMAGE_SIZE[1]*cfg.MODEL.SIMDR_SPLIT_RATIO))

        # 水平和垂直处理
        elif self.coord_representation == 'simdr' or
                self.coord_representation == 'sa-simdr':
        x = rearrange(x_, 'b c h w -> b c (h w)')
        pred_x = self.mlp_head_x(x)
        pred_y = self.mlp_head_y(x)
        return pred_x, pred_y
```

4.2 实验环境搭建

Ubuntu 20.04
python 3.8.2
pytorch- 1.11.0+cu113
CUDA 11.4
1 个 NVIDIA RTX 3090 GPU

4.3 创新点

水平和垂直分类器（即，每个分类器只有一个线性层）分别附加在主干之后以执行坐标分类。对于基于 **CNN** 的主干，我们简单地将输出的关键点表示从 (n, H, W) 扁平化为 $(n, H \times W)$ 以进行分类。与使用多个昂贵的去卷积层作为头的基于热图的方法 **[20]** 相比，**SimCC** 头更轻巧和简单。坐标分类。

为了实现分类，我们建议将每个连续坐标值统一离散为一个整数作为模型训练的类标签：$C_x \in [1, N_x], C_y \in [1, N_y]$，其中 $N_x = W \cdot k$ 和 $N_y = H \cdot k$ 分别表示水平和垂直轴的 **bin** 数量。k 为分裂因子，设置为 **≥ 1** 以减小量化误差，从而获得亚像素定位精度。为了产生最终的预测，**SimCC** 根据主干学习的 **n** 个关键点表示独立地执行垂直和水平坐标分类。具

体地，给定第 **i** 个关键点表示作为输入，第 **i** 个关键点预测 o_x^i 和 o_y^i 分别由水平和垂直坐标分类器生成。此外，**Kullback-Leibler** 散度被用作训练的损失函数。

并将 **SimCC** 与原论文中的互信息损失形成残差模块，起到对模型的约束作用

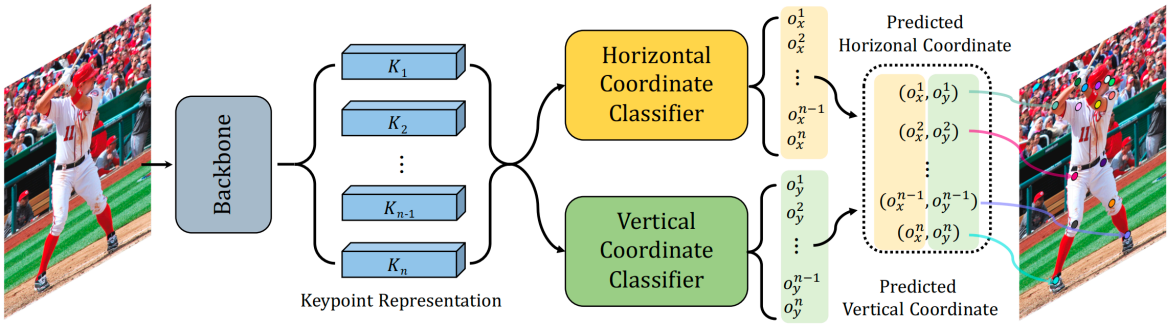

图 **2. SimCC** 框架图

5 实验结果分析

通过表1所示，我们可以看出，采用 **SimCC** 和互信息损失构成的残差模块，与论文中原模型的效果稍差些。

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
FAMI-Pose	87.6	88.6	84.1	77.8	80.0	81.2	73.8	82.2
SmiCC-Pose	87.8	88.0	83.6	76.6	80.6	79.7	71.3	81.6

表 **1. PoseTrack2017** 验证集的定量结果

6 总结与展望

本篇论文中，研究了多帧人体姿态估计任务的角度，有效地利用时间背景，通过特征对齐和互补信息挖掘。提出了一个分层的粗到细的网络，逐步对齐支持框架功能的关键帧功能。在理论上，我们进一步引入了一个互信息目标的有效监督的中间功能。

参考文献

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018.

[2] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 7035–7044, 2020.

[3] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely-labeled videos. *Advances in neural information processing systems*, 32, 2019.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[5] Yuntao Chen, Chenxia Han, Naiyan Wang, and Zhaoxiang Zhang. Revisiting feature alignment for one-stage object detection. *arXiv preprint arXiv:1908.01570*, 2019.

[6] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020.

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[8] Yanbin Hao, Zi-Niu Liu, Hao Zhang, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. Person-level action recognition in complex events via tsd-tsm networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4699–4702, 2020.

[9] Shihua Huang, Zhichao Lu, Ran Cheng, and Cheng He. Fapn: Feature-aligned pyramid network for dense image prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 864–873, 2021.

[10] Zilong Huang, Yunchao Wei, Xinggang Wang, Wenyu Liu, Thomas S Huang, and Humphrey Shi. Alignseg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):550–557, 2021.

[11] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2020, 2017.

[12] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 775–793. Springer, 2020.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[14] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 525–534, 2021.

[15] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3266–3275, 2019.

[16] Davide Mazzini. Guided upsampling network for real-time semantic segmentation. *arXiv preprint arXiv:1807.07466*, 2018.

[17] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 1913–1921, 2015.

[18] Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded models for articulated pose estimation. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*, pages 406–420. Springer, 2010.

[19] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[20] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[21] Xiaoqin Zhang, Changcheng Li, Xiaofeng Tong, Weiming Hu, Steve Maybank, and Yimin Zhang. Efficient human pose estimation via parsing a tree structure based human model. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1349–1356. IEEE, 2009.

[22] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019.