Final Project Report

Title: High Performance Body Pose Estimation

Course: Programming for Visual Computing II (CMPT 733), Simon Fraser University

Lecturer: Ali Mahdavi Amiri

Sponsor: Huawei Technologies Canada

Group Members:

- Mohammadreza Dorkhah
- Anupam Jose
- Pranav Sharma

1. Introduction:

Human Body Pose Estimation is one of the essential tasks in computer vision, this involves the representation of the orientation of a person in a graphical format. Highlighting its body parts and joint positions. This problem was solved earlier using the part-based models. The basic idea of part-based model was attributed to the human skeleton. Any object having the property of articulation can be broken down into smaller parts wherein each part can take different orientations, resulting in different articulations of the same object. Different scales and orientations of the main object can be articulated to scales and orientations of the corresponding parts. Another way tried by the computer vision community was using an articulated model with quaternion. The kinematic skeleton was constructed by a tree-structured chain where each rigid body segment has its local coordinate system that was transformed into the world coordinate system.

In recent years, the use of deep learning has become quite prevalent in the field where rather than building an explicit model for the parts as above, the appearances of the joints and relationships between the joints of the body are learned from large training sets. These models focus on extracting the 2D positions of joints (keypoints), the 3D positions of joints, or the 3D shape of the body from either a single or multiple images.

2. TimeLine:

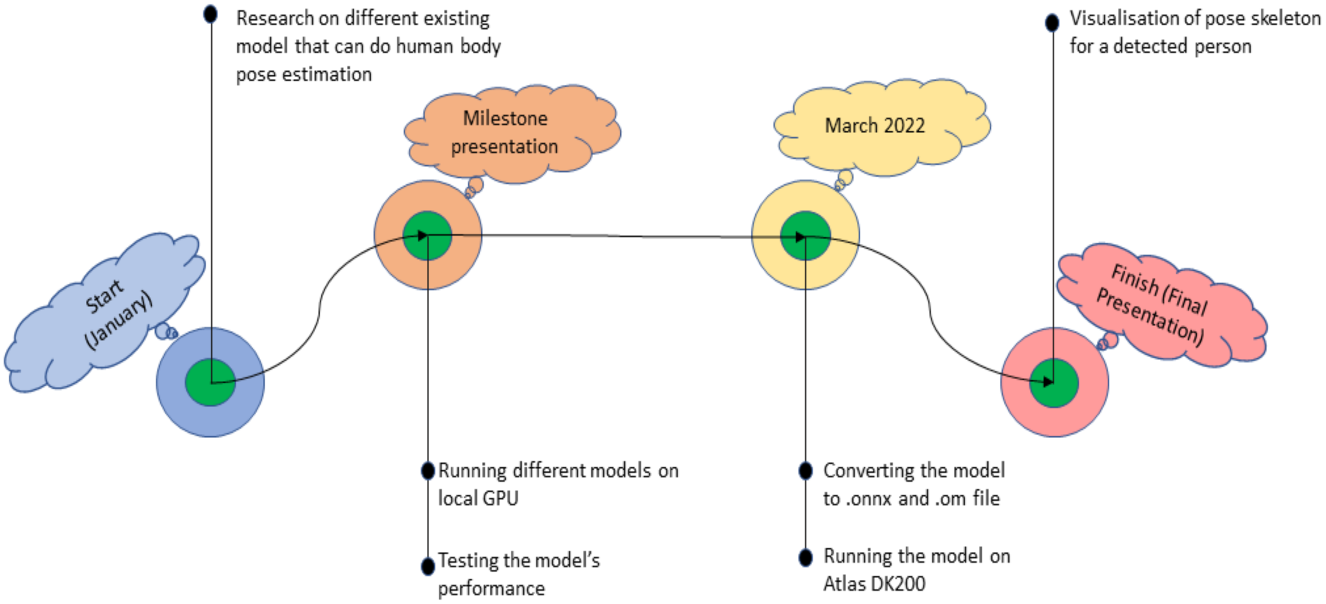
Research on different existing model that can do human body pose estimation
Visualisation of pose skeleton for a detected person
Milestone presentation
March 2022
Start (January)
Finish (Final Presentation)
Running different models on local GPU
Testing the model's performance
Converting the model to .onnx and .om file
Running the model on Atlas DK200

Figure 1: Timeline of the project

We started this project in January by reviewing different papers related to Human Body Pose Estimation. We found different relevant papers [1, 2, 3, 4] that take an Image or video as input and give 2D pose estimates and some model even give 3D pose estimation. Then we also looked into different models for 2D key point generations such as OpenPose [1], Lightweight OpenPose [2], Unipose [3], and MHFormer [4]. We then converted these models into offline models and tested the performance on local GPU. We successfully converted the pre-trained models and got similar performance results on the benchmark 4 datasets as proposed in the paper. In the month of February, we worked on getting the 2D key points from the above-mentioned models. Then, we converted these models to ".onnx" and ".om" file format so as to run these models on Atlas 200DK board and got the desired outputs. Finally, we were able to achieve the required visualizations of the human pose skeleton for a detected person in the test image and video using the Atlas 200DK board.

3. Overview:

3.1. OpenPose:

The first model we made the pipeline for was the OpenPose model, the OpenPose model is a widely used 3D pose estimation model from 2018. It involves a neural network implementation consisting of VGG-19 layers as its backbone. The model is capable of detecting 134 keypoints for explicitly detecting the hands and feet joints. However, we are using 18 basic keypoints here. We tried two implementations for OpenPose: Pytorch implementation as well as the Tensorflow implementation. [1]

3.2. LightWeight OpenPose:

Another model that had a very similar architecture with OpenPose was the Lighweight Openpose, this model takes in pointers from its parent while making the architecture lightweight using works of mobilenet v2, a dilated Mobilenet v2 is used along with 2 extra convolution layers. The model is trained using MS COCO and CMU Panoptic datasets. This results in even less position error than its parent architecture while also being lightweight. The model was also optimized by the creators to work with Intel OpenVINO and tensorflow [2].

3.3. UniPose:

Unipose was one of the state-of-the-art models for single person pose detection from the year 2021. UniPose incorporated contextual segmentation and joint localization for estimation of human pose in a single stage, maintaining with high accuracy and without relying on statistical postprocessing methods. This was achieved using a novel waterfall module that leveraged the efficiency of progressive filtering in the cascade architecture, while maintaining multi-scale fields-of-view comparable to spatial pyramid configurations. An LSTM implementation is used for multi-frame processing in UniPose [3].

3.4. MHFormer:

MH Former or Multi-Hypothesis Transformer is another state-of-the-art 3D Pose detection implementation from 2021. This model learns spatio-temporal representations of multiple plausible pose hypotheses. In order to effectively model multi-hypothesis dependencies and build strong relationships across hypothesis features. It has a detailed three step process to reach its precision. The final representation is enhanced and the synthesized pose is very accurate. [4]

4. 3D Pose Estimation Pipeline:

The pipeline for 3D poses estimation includes three parts:

- Object Detection
- 2D Pose Generation
- 3D Pose Generation

The first part involves object detection of the person in the image, this part is usually done using the existing YOLOv3 weights for the object detection part of the respective model, in our case all of the converted models OpenPose, Unipose, MHFormer have a well-defined object detection model for the human body detection.

Once the bounding box containing a human is extracted from the image, this becomes the baseline input for our 2D key point generation for our 2D pose estimation. This model predicts a pose skeleton having 17 keypoints (joints) corresponding to ankles, knees, hips, elbow, head structure etc.

These 2D keypoints are then passed to a 3D Pose estimating model of MHFormer which generates an accurate representation of fully convolution model based on dilated temporal convolutions over 2D Key-points. This temporal convolutional model takes 2D Key-point sequences as input and generates 3D Pose estimates as output.

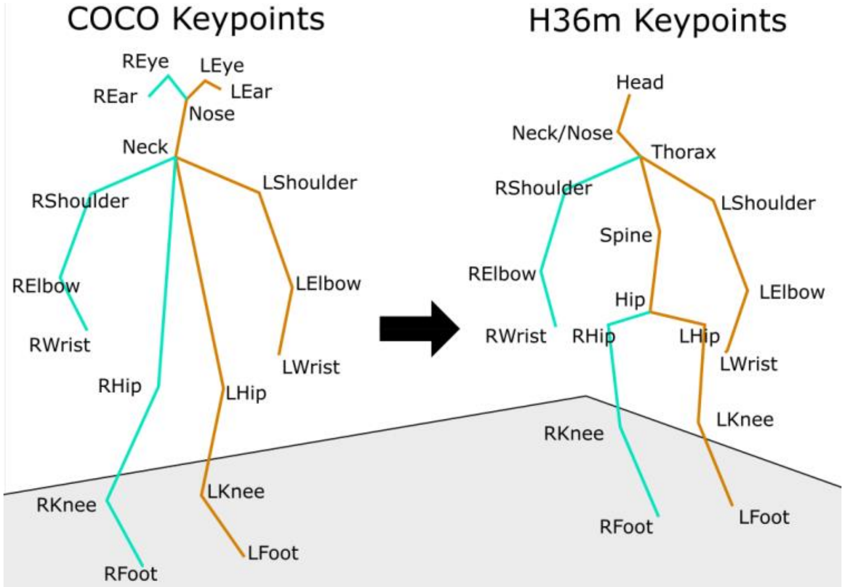

Figure 2: Comparison of COCO keypoint format and H36m keypoint format

5. Results:

Shown below are the results of running the above stated pipeline end to end on Atlas 200DK and on local GPU. We used different images consisting of single person, multi person, single person with background and Complex pose images. We have taken these four cases so as to see how different models perform on these different images.

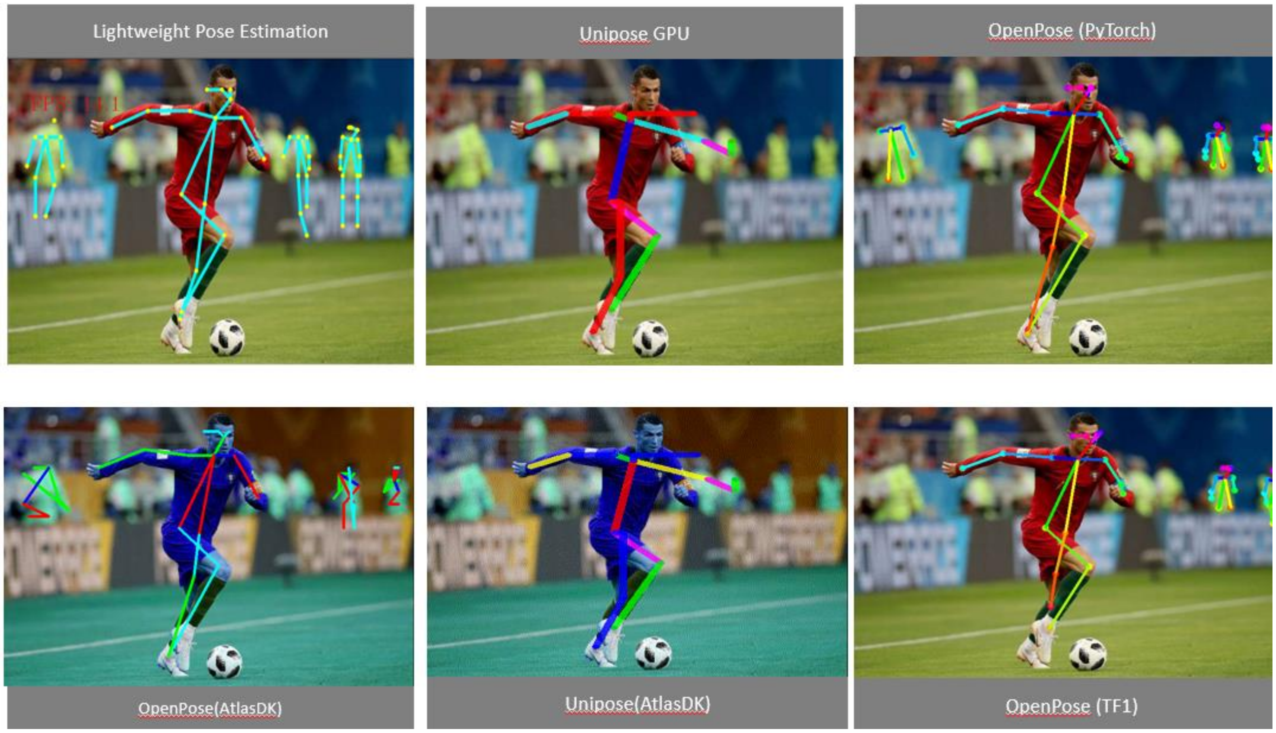

Figure 3: Single Person with Background People

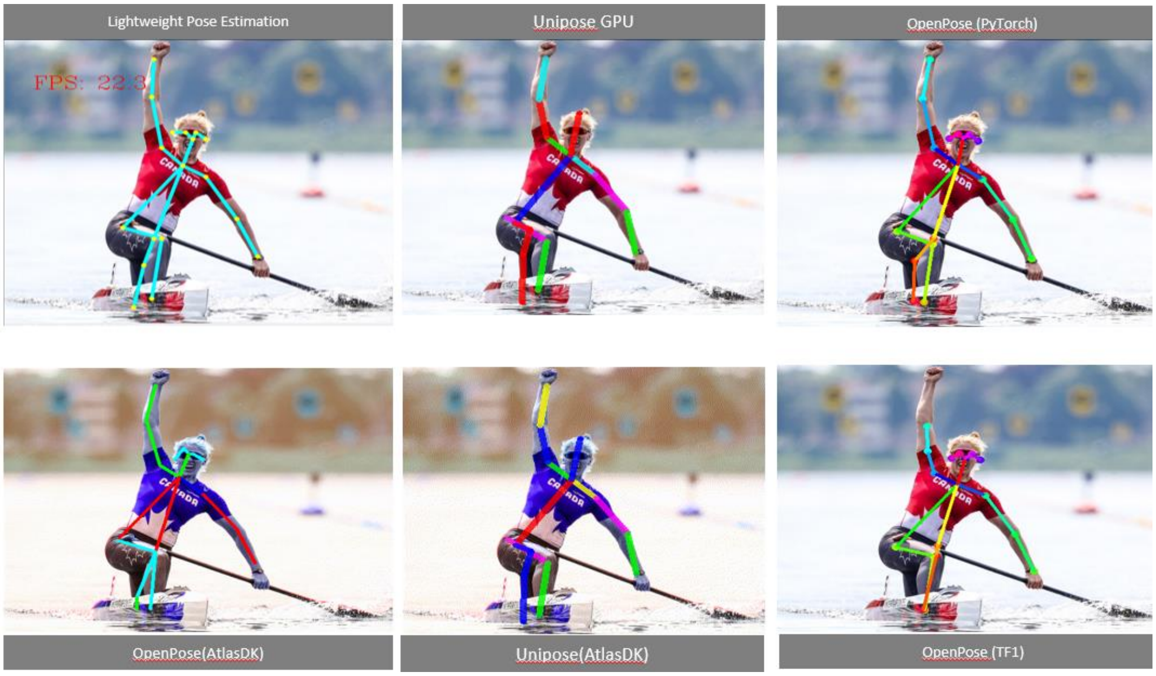

Figure 4: Singe Person with Occluded Parts

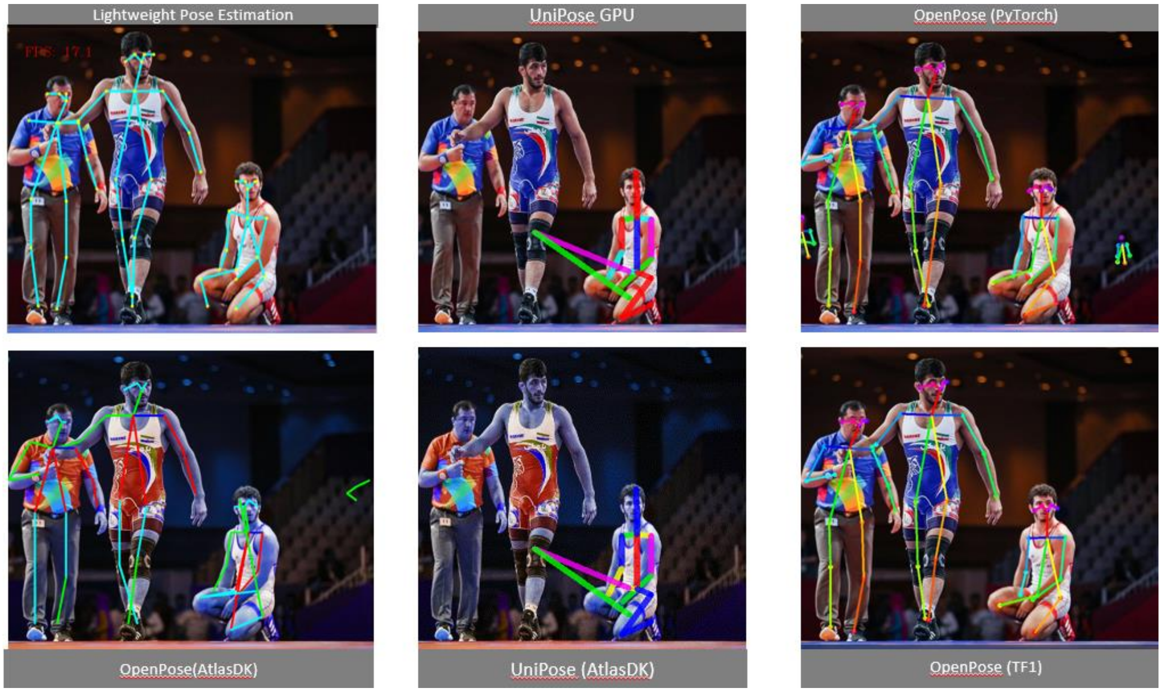

Figure 5: Multiple Persons

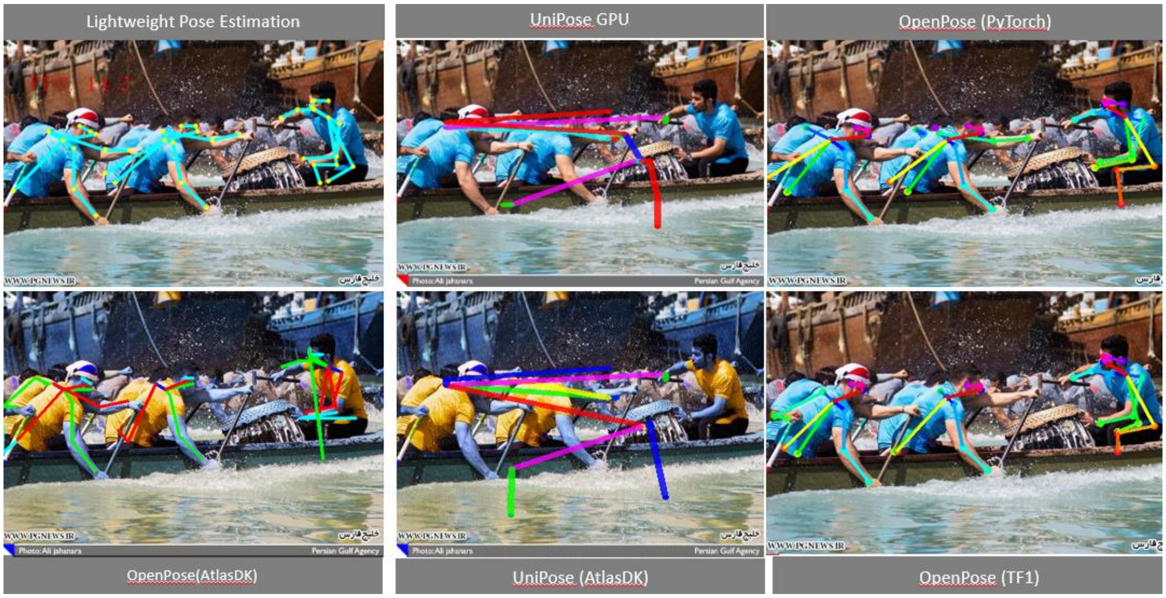

Figure 6: Multiple Persons with Background People and Occluded Parts

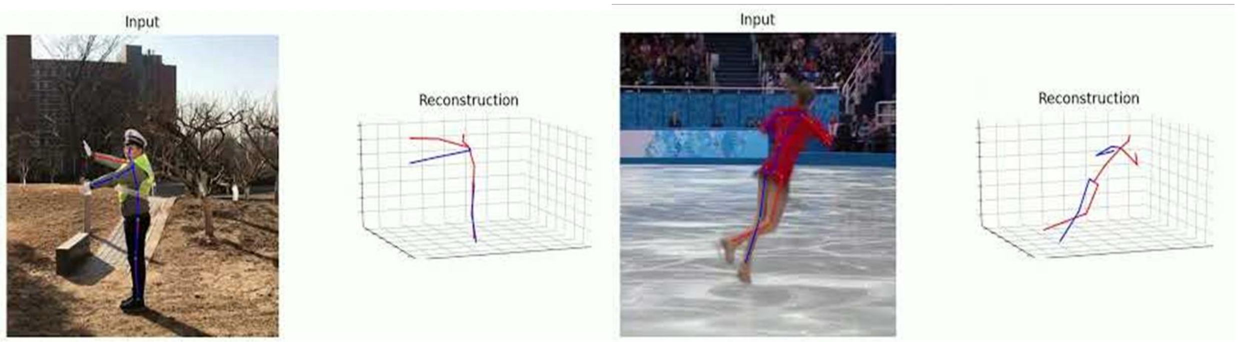

Figure 7: MHFormer Results

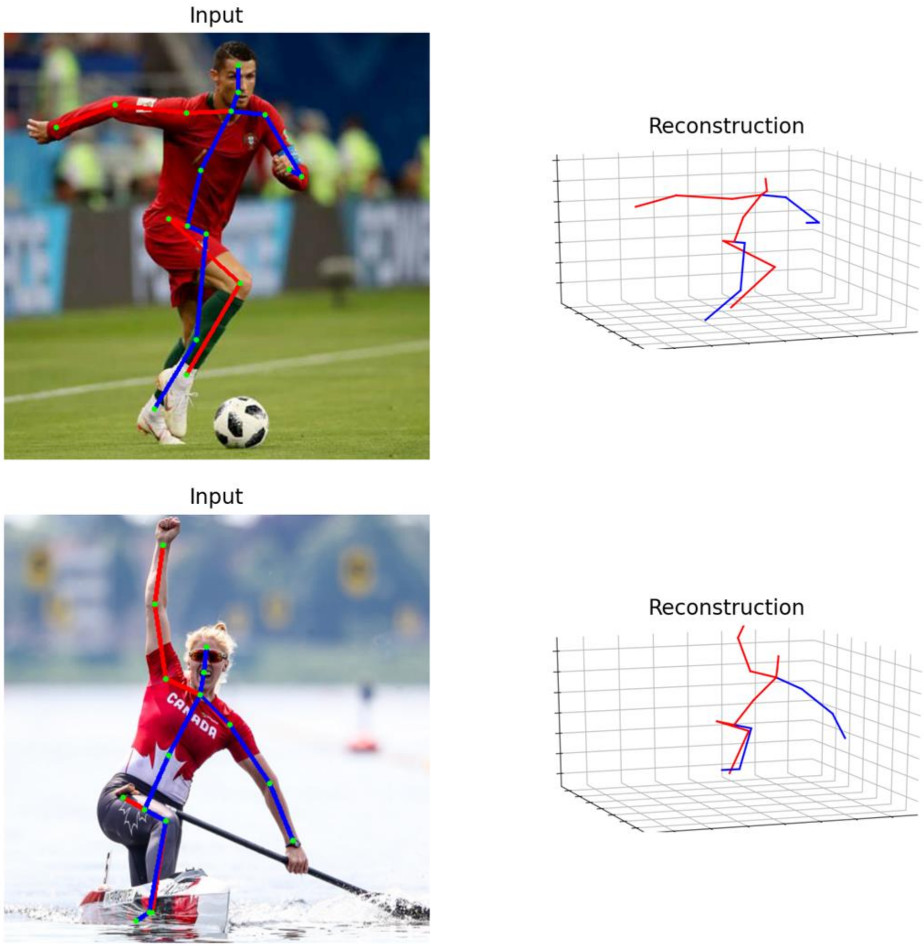

Figure 8: MHFormer 3D Pose Estimation

6. What We've Learned:

During this project we had the chance to review and implement different Human Pose Estimation models, therefore we could notice their differences and the effect of these differences in their output. However, the most time-consuming part of the work was the conversion of each model to ".onnx" and then to ".om" in order to be used on Atlas 200DK device. We faced many failures in this conversion due to technical difficulties, so we learned all steps, variables and functions should be convertible. In some cases, we tried other replacements for some of the models specifically in the Object Detection part. Furthermore, building the pipeline for deployment on Atlas 200DK was another hard task that required us to get familiar with the ACL functionalities and its associated restrictions.

References:

[1] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7291-7299

[2] Daniil Osokin. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. arXiv preprint arXiv:1811.12004, 2018.

[3] Bruno Artacho, Andreas Savakis Unipose; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7035-7044

[4] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, Luc Van Gool MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation: Computer Vision and Pattern Recognition 2022

[5] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7753–7762, 2019

[6] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017

[7] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https: //github.com/facebookresearch/detectron2, 2019