**Residual Pose: A Decoupled Approach for Depth-based 3D Human Pose Estimation**

**Link-** [**https://arxiv.org/pdf/2011.05010.pdf**](https://arxiv.org/pdf/2011.05010.pdf)

**The 7 W's**

Q1. We propose to leverage recent advances in reliable 2D pose estimation with Convolutional Neural Networks (CNN) to estimate the 3D pose of people from depth images in multiperson Human-Robot Interaction (HRI) scenarios.

Q2. Although 3D pose estimation has been a very important topic of research, factors like person self-occlusions, pose variations, sensing conditions and low computational budget increase the challenge of deploying accurate, reliable and efficient 3D pose estimation systems.

Q3. Other approaches model the 3D location distributions of 3D points with respect to their parents in a kinematic tree [9], [19]. As a typical example, the seminal work of Shotton [19] employed random forests to classify depth pixels into different human joints and used weighted voting to estimate their 3D locations. In [15] depth images are transformed into a voxelized representation and 3D Gaussian likelihoods are predicted for each body joint per voxel using a costly 3D-CNN.

Q4. However, these methods usually work on image crops centered around the person. As a consequence, to handle the multi-person case, a person detector is still needed as a first step, followed by multiple forward passes of a relatively heavy image processing network to estimate the 3D pose of each detected person, leading to an increased computational cost.

Q5. Our main idea is to better exploit the depth information and decouple the task in two main steps: 2D multi-person pose estimation and 3D pose regression. A simple but effective scheme which lifts the 2D estimates to 3D using the depth information and pose priors (to handle partial occlusion); and a novel efficient residual pose 3D regression method that works on this set of points. This makes our approach computationally lighter for multi-person HRI settings since compared to CNNs applied to image crops for 3D pose prediction, the cost of our 3D regression scheme is much smaller, and the cost saving is proportional to the number of people in the scene.

Q6.

* we investigate an innovative method decoupling the 3D pose estimation task into an accurate and efficient CNN-based 2D bottom-up multi-person pose estimation method and 3D pose regression;
* we propose a simple 2D-to-3D lifting scheme which handles 2D body joint miss detections;
* we introduce a novel method for 3D pose regression from lifted 2D estimates by relying on a residual-pose deep-learning architecture;
* we demonstrate that despite its simplicity, our approach achieves very competitive results on different public datasets and is suitable for multi-party HRI scenarios.

Q7. One limitation of our model is that it does not consider the skeleton kinematics in the learning process. Additionally, body motion modelling can be introduced to introduce temporal consistency in our 3D predictions.