**A simple yet effective baseline for 3d human pose estimation**

**Link-** [**Baseline for 3D HPE**](https://openaccess.thecvf.com/content_ICCV_2017/papers/Martinez_A_Simple_yet_ICCV_2017_paper.pdf)

**The 7 W's**

Q1. HPE models despite their excellent performance, it is often not easy to understand whether their remaining error stems from a limited 2d pose (visual) understanding, or from a failure to map 2d poses into 3- dimensional positions. With the goal of understanding these sources of error, we set out to build a system that given 2d joint locations predicts 3d positions.

Q2. It is relevant because it considerably improves upon the previous best 2d-to-3d pose estimation result using noise-free 2d detections in Human3.6M, while also using a simpler architecture.

Q3. Recent work that learns the mapping between 2d and 3d with deep neural networks. Pavlakos et al. [33] introduced a deep convolutional neural network based on the stacked hourglass architecture [30] that, instead of regressing 2d joint probability heatmaps, maps to probability distributions in 3d space.

Q4. Our work considerably improves upon the previous best 2d-to-3d pose estimation result using noise-free 2d detections in Human3.6M, while also using a simpler architecture. However, since we are dealing with low-dimensional points as inputs and outputs, we can use simpler and less computationally expensive linear layers. RELUs [29] are a standard choice to add non-linearities in deep neural networks.

Q5. Our approach is based on a simple, deep, multilayer neural network with batch normalization [17], dropout [44] and Rectified Linear Units (RELUs) [29], as well as residual connections [16]. There are two extra linear layers: one applied directly to the input, which increases its dimensionality to 1024, and one applied before the final prediction, that produces outputs of size 3n.

Q6.

* Our first design choice is to use 2d and 3d points as inputs and outputs, in contrast to recent work that has used raw images or 2d probability distributions [33, 56] as inputs, and 3d probabilities [33], 3d motion parameters [54] or basis pose coefficients and camera parameter estimation as outputs.
* Our architecture benefits from multiple relatively recent improvements on the optimization of deep neural networks, which have mostly appeared in the context of very deep convolutional neural networks and have been the key ingredient of state-of-the-art systems submitted to the ILSVRC (Imagenet [10]) benchmark.
* Our work considerably improves upon the previous best 2d-to-3d pose estimation result using noise-free 2d detections in Human3.6M, while also using a simpler architecture.

Q7.

* Our network currently does not have access to visual evidence; we believe that adding this information to our pipeline, either via fine-tuning of the 2d detections or through multi-sensor fusion will lead to further gains in performance.
* We believe that a further exploration of the network architectures will result in improved performance. These are all interesting areas of future work.