**Human Pose Estimation via Convolutional Part Heatmap Regression**

**The 7 W's**

Q1. We proposed a CNN cascaded architecture for human pose estimation particularly suitable for learning part relationships and spatial context, and robustly inferring pose even for the case of severe part occlusions. Key feature of our network is the joint regression of part detection heatmaps.

Q2. It is relevant because the proposed architecture is very simple and can be trained end-to-end, achieving top performance on the MPII and LSP data sets.

Q3. Recently proposed methods for articulated human pose estimation using CNNs can be classified as detection-based or regression-based. Detection-based methods are relying on powerful CNN-based part detectors which are then combined using a graphical model or refined using regression. Regression-based methods try to learn a mapping from image and CNN features to part locations.

Q4. As a mapping from CNN features to part locations might be difficult to learn in one shot, regression-based methods can be also applied sequentially (i.e. in a cascaded manner). Our CNN cascade is based on a two-step detection-followed-by-regression approach (see Fig. 1) and as such is related to both detection-based and regression-based methods.

Q5. This paper is on human pose estimation using Convolutional Neural Networks. Our main contribution is a CNN cascaded architecture specifically designed for learning part relationships and spatial context, and robustly inferring pose even for the case of severe part occlusions. To this end, we propose a detection-followed-by-regression CNN cascade. The first part of our cascade outputs part detection heatmaps and the second part performs regression on these heatmaps.

Q6. The benefits of the proposed architecture are multi-fold: It guides the network where to focus in the image and effectively encodes part constraints and context. More importantly, it can effectively cope with occlusions because part detection heatmaps for occluded parts provide low confidence scores which subsequently guide the regression part of our network to rely on contextual information in order to predict the location of these parts.