**Multi-source Deep Learning for Human Pose Estimation**

**The 7 W's**

Q1. Visual appearance score, appearance mixture type and deformation are three important information sources for human pose estimation. This paper proposes to build a multi-source deep model in order to extract non-linear representation from these different aspects of information sources.

Q2. It is relevant because the deep model out performs state-of-the-art by up to 8.6 percent on three public benchmark datasets.

Q3. Krizhevsky *et al.* [23] pro- posed a large-scale deep convolutional network [27] with breakthrough on the large-scale ImageNet object recognition dataset [8], attaining a significant gap compared with existing approaches that use shallow models, and bringing high impact to research in computer vision.

Q4. To the best of our knowledge, however, deep model for human pose estimation has not yet been explored. Our work is inspired by multi-modality models that learn from multiple modalities such as audio, visual, text data [35, 47, 17]. In contrast to these works, we investigate multi- source learning from single modality, which is image data in pose estimation.

Q5.

* We propose a deep architecture to construct the non- linear representation from different aspects of information sources. To the best of our knowledge, this paper is the first to use deep model for pose estimation.
* The body articulation patterns (global and more abstract representations) are captured by the deep model from the information sources (local and less abstract representations). For each information source, more abstract representation at the higher layer is composed by the less abstract representation of all body parts in the lower layer. Then representations of all information sources in the higher layer are fused for pose estimation.

Q6. Both the task for detecting human and the task for estimating body locations are jointly learned using a single deep model. Joint learning of these tasks with a shared representation improves pose estimation accuracy. Since this model is a post-processing of information sources, it is very flexible in terms of integrating with existing approaches that use different information sources, features, or articulation models.

Q7. Learning deep model from pixels for pose estimation and analyzing the influence of training data number will be the future work.