**A Top-down Approach to Articulated Human Pose Estimation and Tracking**

**Link-** [**https://arxiv.org/pdf/1901.07680.pdf**](https://arxiv.org/pdf/1901.07680.pdf)

**The 7 W's**

Q1. In this paper, following the top-down approach, we aim to build a strong baseline system with three modules: human candidate detector, single-person pose estimator and human pose tracker.

Q2. It is relevant because it consists of three modules, which conduct human candidate detection, pose estimation, and pose tracking respectively. It examines each module in the system with ablation studies on various models and configurations while discussing their pros and cons.

Q3.

* PoseTrack[12] and ArtTrack[11] primarily introduce multi-person pose tracking challenge and propose a graph partitioning formulation, which transforms the pose tracking problem into a minimum cost multi-cut problem.
* Another line of research explores top-down approach [6,20,19] by operating multi-person human pose estimation on each frame and linking them based on appearance similarities and temporal adjacencies.
* A naive solution is to apply multi-target object tracking on human detection candidates across frames and then estimate human poses for each human tubelet.

Q4.

* However, hand-crafted graphical models are not scalable for long and unseen clips.
* While these are feasible methods, it neglects unique attributes of keypoints. Compared to the tracked bounding boxes, keypoints can potentially be helpful cues for both the bounding boxes and the keypoints tracking.

Q5.

* Firstly, we choose a generic object detector among state-of-the-art methods to detect human candidates.
* Then, cascaded pyramid network is used to estimate the corresponding human pose. For the single-person human pose estimator, we adopt Cascaded Pyramid Networks (CPN) [3] with slight modifications. We first train the CPN network with the merged dataset of PoseTrack 2018 and COCO for 260 epochs.
* Finally, we use a flow-based pose tracker to render keypoint-association across frames, i.e., assigning each human candidate a unique and temporally-consistent id, for the multi-target pose tracking purpose.

Q6. The main advantage of top-down approaches is their capability in disassembling the task into multiple comparatively easier tasks, i.e., object detection and single-person pose estimation. The object detector is expert in detecting hard (usually small) candidates, so that the pose estimator will perform better with a focused regression space.