**Multi-person Pose Estimation**

Zuoyue Li

**1 Introduction**

**1.1 Problem Definition**

The task of multi-person pose estimation in an image is to identify every person instance and to localize its facial and body keypoints. Specifically, given an RGB image *I* = {0,1,…,255}*H*×*W*×3, our goal is to find a set of person keypoints *K* = {*Kij*}*i*=1,…*N*, *j*=1,…,*M*, where *N* is the number of people in the image, *M* is the number of keypoints to be detected for a single person, *Kij* refers to the pixel coordinates of *i*-th person’s *j*-th keypoint. Usually, the value of *M* should be well set in advance, i.e. each index should correspond to a joint of a person.

**1.2 Frameworks**

There are two types of framework for tackling the problem of multi-person pose estimation: the top-down approach and the bottom-up approach, both of which are illustrated as follows:  
(1) The top-down approach starts by identifying and localizing individual person instances roughly by means of a bounding box object detector, followed by single-person pose estimation.  
(2) The bottom-up approach starts by localizing identity-free semantic entities, i.e. individual keypoint proposals, followed by grouping them into person instances.

Both of the two pipelines are exactly two-step frameworks. In the top-down approaches, the second step, single-person pose estimation, relies heavily on the first step, human detection. If the bounding box of a person found is determined as false positive, the second step would tend to make false prediction. As for top-down approaches, although they are box-free, it usually requires additional information to group keypoints into a person instance, because the only keypoints proposal itself is not enough. If the additional joints connection information gives false guidance, the pose decoding scheme would further break down. This often appears in people crowd.

As for the runtime comparison, the bottom-up approaches are usually more efficient than the top-down approaches, because the time spent by the second step, single-person pose estimation of top-down approaches is proportional to the number of people in the image. By contrast, bottom-up approaches do not have the problem of computing efficiency in general.

**1.3 Motivation**

It is mentioned above that in the top-down approach, the decision of person bounding box is very important. It determines whether the single-person pose estimator can accurately predict the human pose or not. Although the existing networks for object detection are already very mature, we still hope that it can be further improved.

We believe that the heatmaps generated by bottom-up approaches can help the human detection scheme. This is because if the heatmaps do not show any peak (or only have some smoothed peaks) within the area decided by a positive human bounding box, then this bounding box is likely to be false positive. Similarly, if there is no bounding box wrapping an area with some obvious peaks, then there may be one or more missing bounding boxes (false negative) which are not retrieved. By contrast, the human bounding boxes predicted by object detector can also provide further information to the bottom-up approaches in the grouping scheme.

Our goal is to combine top-down and bottom-up methods, maximizing the use of all the known information, i.e. the global keypoints information (heatmaps) provided by bottom-up approaches, and the localization information (bounding boxes) provided by top-down methods. We hope this kind of combination can finally improve the he final person pose estimation results.

**2 Related Work**

**2.1 Top-down Approaches**

(1) Mask R-CNN

(2) AlphaPose

**2.2 Bottom-up Approaches**

(1)