

# PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model

George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris,  
Jonathan Tompson, and Kevin Murphy

Google Research

fgpapan, tylerzhu, lcchen, spyros, tompson, kpmurphy@google.com

**Abstract.** We present a box-free bottom-up approach for the tasks of pose estimation and instance segmentation of people in multi-person images using an efficient single-shot model. The proposed PersonLab model tackles both semantic-level reasoning and object-part associations using part-based modeling. Our model employs a convolutional network which learns to detect individual keypoints and predict their relative displacements, allowing us to group keypoints into person pose instances. Further, we propose a part-induced geometric embedding descriptor which allows us to associate semantic person pixels with their corresponding person instance, delivering instance-level person segmentations. Our system is based on a fully-convolutional architecture and allows for efficient inference, with runtime essentially independent of the number of people present in the scene. Trained on COCO data alone, our system achieves COCO test-dev keypoint average precision of 0.665 using single-scale inference and 0.687 using multi-scale inference, significantly outperforming all previous bottom-up pose estimation systems. We are also the first bottom-up method to report competitive results for the person class in the COCO instance segmentation task, achieving a person category average precision of 0.417.

**Keywords:** Person detection and pose estimation, segmentation and grouping.

## 1 Introduction

The rapid recent progress in computer vision has allowed the community to move beyond classic tasks such as bounding box-level face and body detection towards more detailed visual understanding of people in unconstrained environments. In this work we tackle in a unified manner the tasks of multi-person detection, 2-D pose estimation, and instance segmentation. Given a potentially cluttered and crowded ‘in-the-wild’ image, our goal is to identify every person instance, localize its facial and body keypoints, and estimate its instance segmentation mask. A host of computer vision applications such as smart photo editing, person and

































