Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation*
Automatically Generated Evaluation Report

Team Name: maskrnn

Version: 9.0

Split Name: keypoints_testdev2015.json

Friday 4th August, 2017- 15:59

1 Human Pose and Skeleton Color Coding

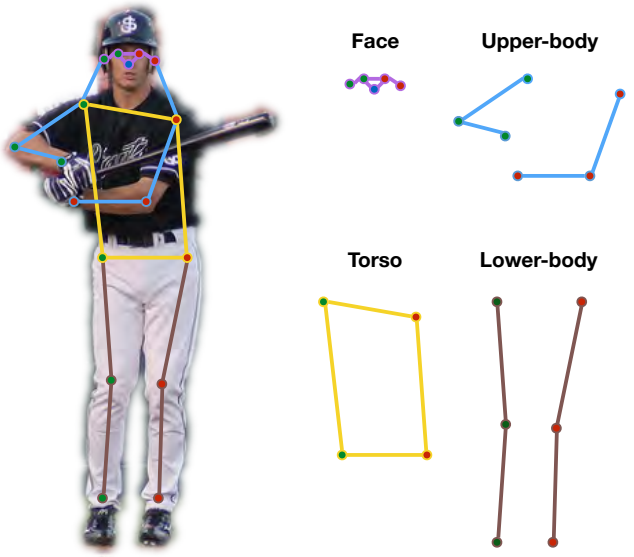

Figure 1: **Detection's Skeleton Color Coding.**

We adopt the following color coding when visualizing an algorithm's keypoint detections:

- The location of the left and right parts of the body is indicated respectively with red and green dots; the location of the nose is plotted in blue.
- Face keypoints (*nose*, *eyes*, *ears*) are connected by purple lines.
- Upper-body keypoints (*shoulders*, *elbows*, *wrists*) are connected by blue lines.
- Torso keypoints (*shoulders*, *hips*) are connected by yellow lines.
- Lower-body keypoints (*hips*, *knees*, *ankles*) are connected by brown lines.

*Code available at: `https://github.com/matteorr/coco-analyze`

2 Overall Detector Characteristics

- **Num. Detections:** 167517
- **Num. Images [with Detections]:** 20288 [18709]

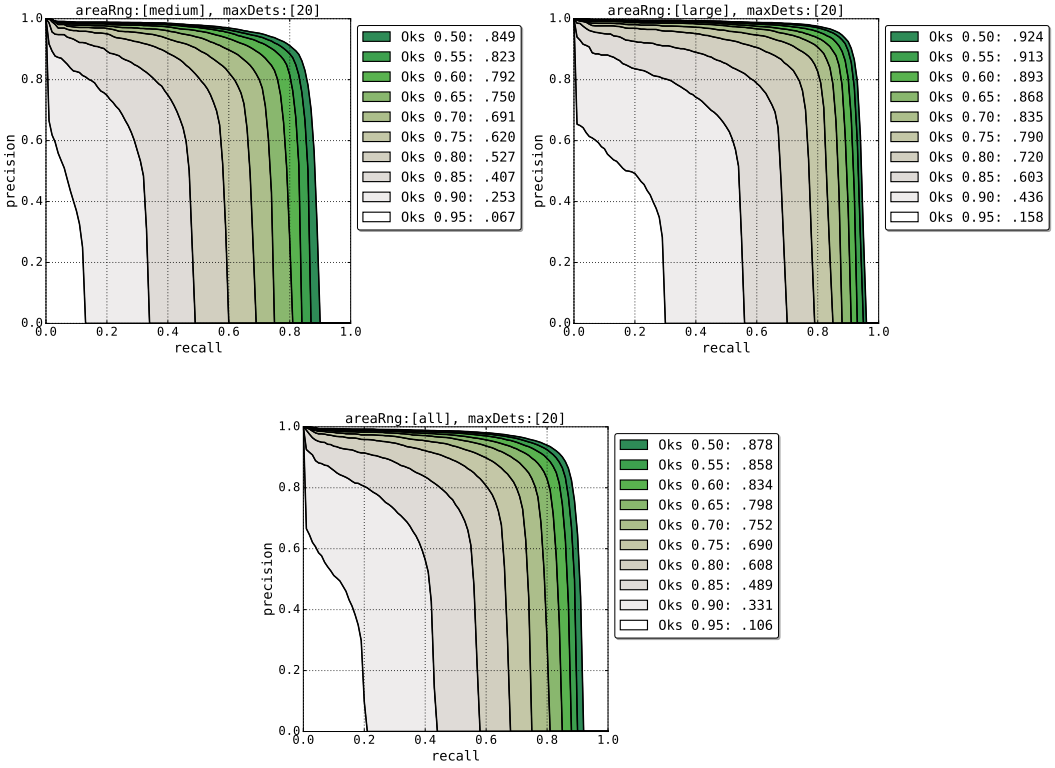

Figure 2: **Precision Recall Curves at all OKS thresholds and area ranges.**

3 Error Impact on AP

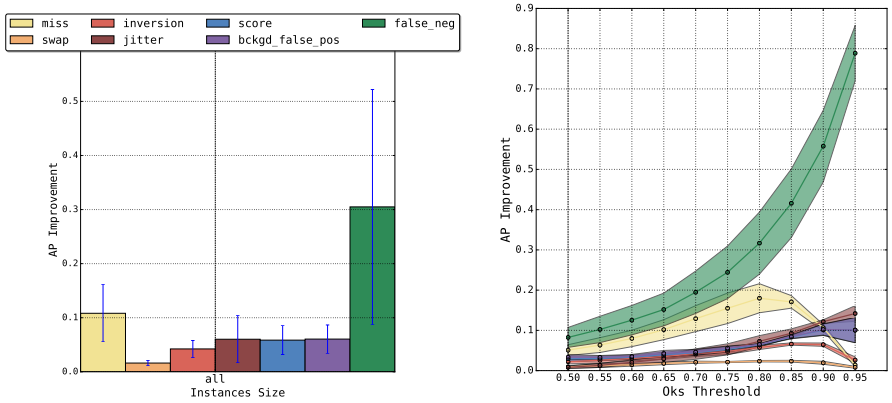

Figure 3: **AP Improvement.** The AP improvement after errors of each type are completely removed, (Left) averaged over all OKS evaluation thresholds at the area range including all detections; (Right) averaged across area ranges at all OKS evaluation thresholds. The value of .85 OKS represents the threshold above which also human annotators have a significant disagreement (around 30%) in estimating the correct position of a keypoint.

4 Localization Errors

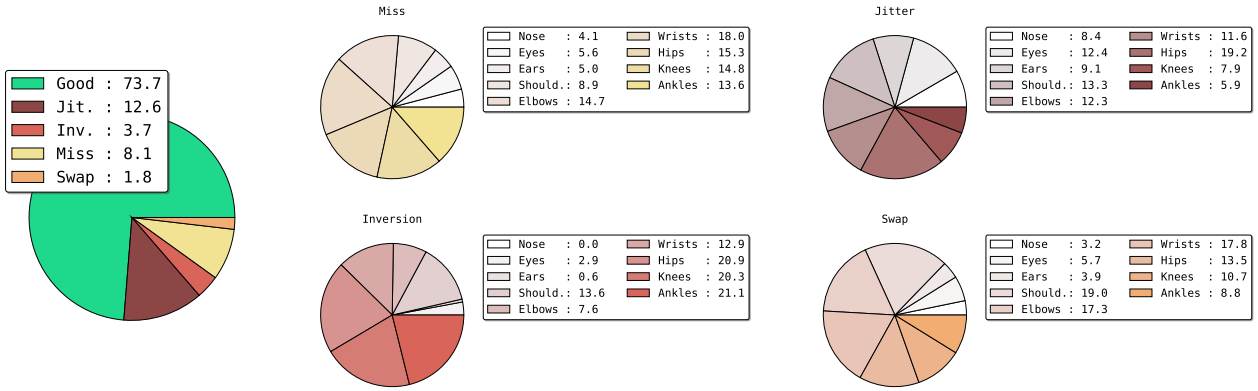

Figure 4: **Predicted Keypoint Analysis.** (Left) The overall percentage of the algorithm's predicted keypoints that are good or have a localization error. (Right) Breakdown of the localization errors over keypoint types.

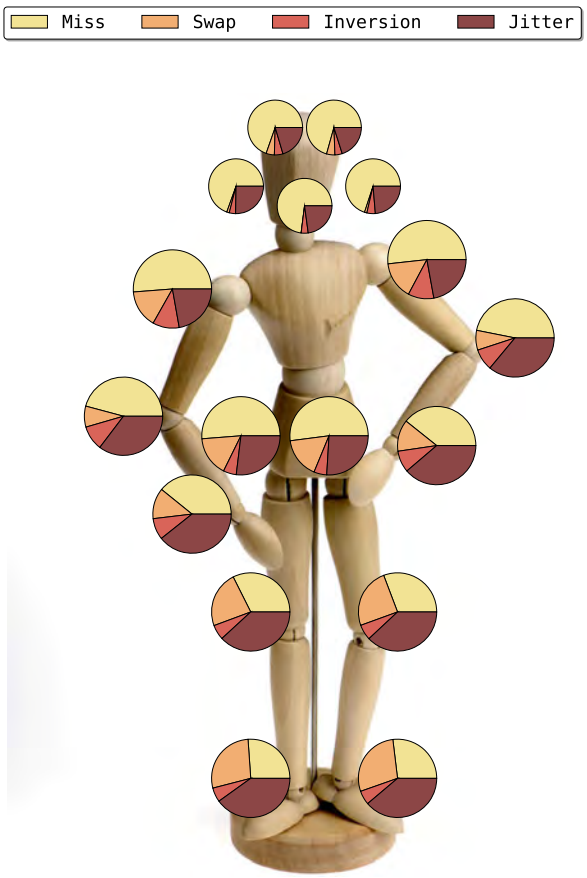

Figure 5: **Human Keypoint Breakdown.** The frequency of each localization error for every keypoint of the human body.

Localization Errors Taxonomy:

- **Jitter:** small error around the correct keypoint location.
- **Miss:** large localization error, the detected keypoint is not within the proximity of any body part.
- **Inversion:** confusion between semantically similar parts belonging to the same instance. The detection is in the proximity of the true keypoint location of the wrong body part.
- **Swap:** confusion between semantically similar parts of different instances. The detection is within the proximity of a body part belonging to a different person.

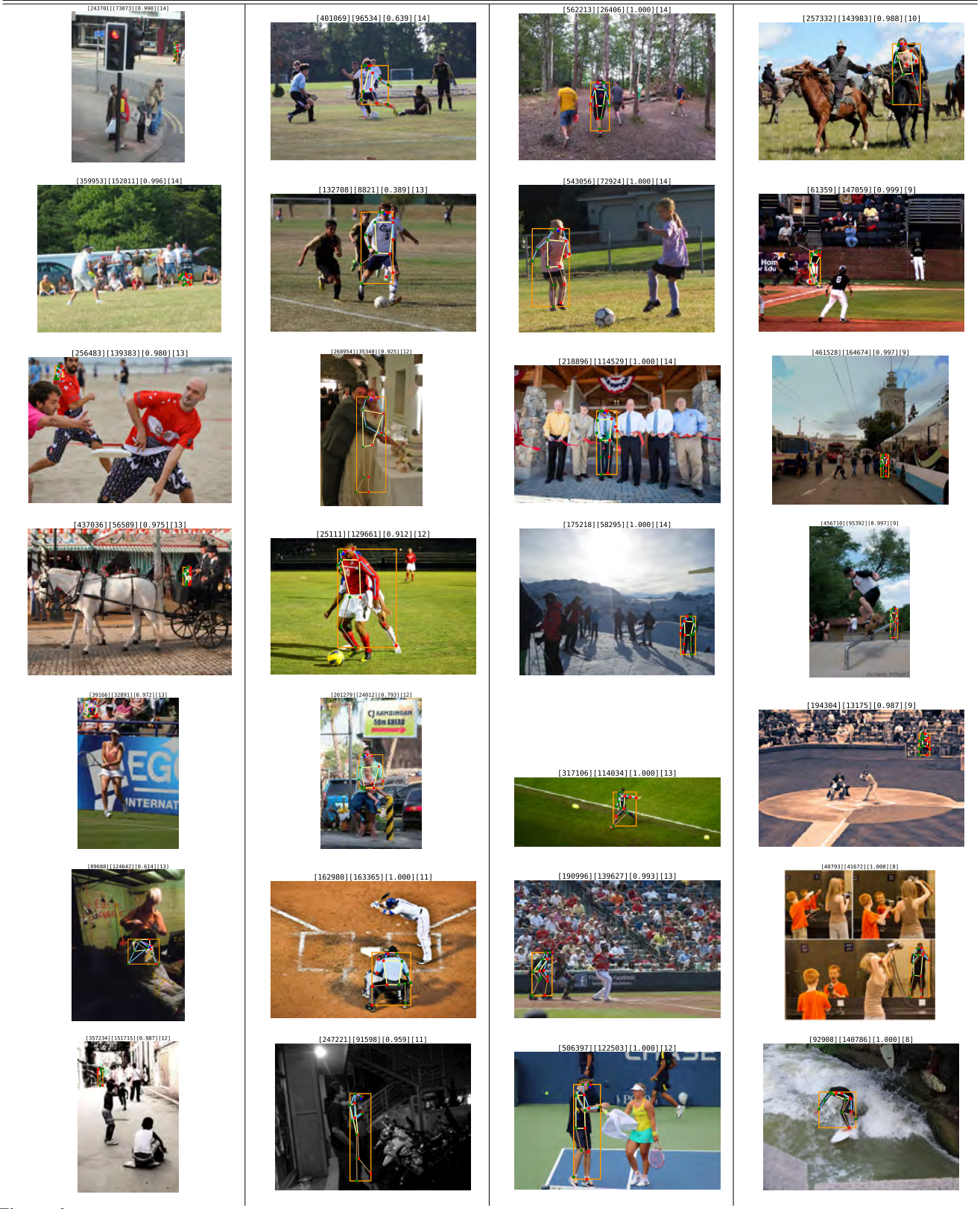

Figure 6: **Top Localization Errors.** The detections with the highest number of localization errors of each type. The color coding of the detection skeleton is described in Sec. 1. Each image title contains the following information: [image_id, detection_id, detection_score, number_of_errors]. Errors in the COCO annotations might cause good detections to appear in the above examples.

5 Scoring Errors

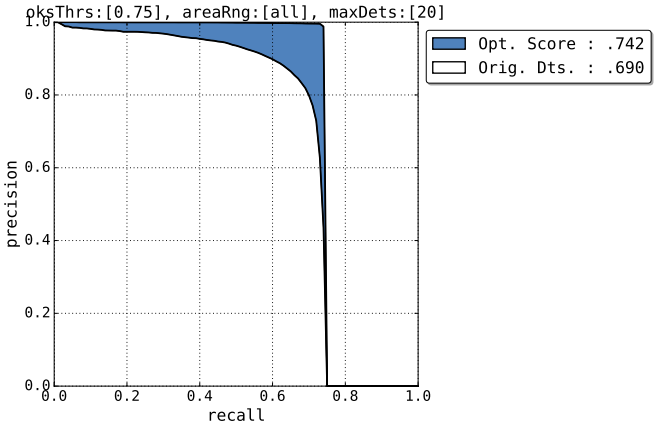

Figure 7: **Optimal Score Precision Recall Curve.** The PRC using the original detections, and the improvement obtained (blue area) when using the optimal score.

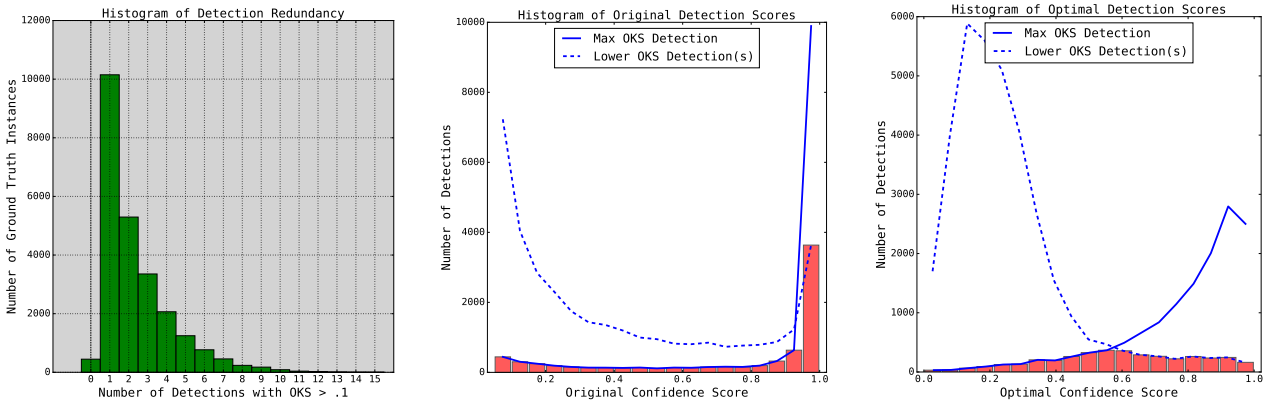

Figure 8: **Detection Scores Analysis.** We compute the following quantities over all ground-truth instances: (Left) The histogram of the number of detections having an OKS $> .1$ with a given ground-truth. The histogram of detections' (Center) original and (Right) optimal confidence scores. We plot separately the detections achieving the maximum OKS with a given ground-truth instance (continuous line) and the other detections achieving OKS of at least .1 (dashed line); in red we highlight how many detections have high OKS and low score and vice versa. A bimodal distribution of confidence scores for detections obtaining high OKS versus low OKS, large separation between the means and a small count of red detections are indication of an overall better score.

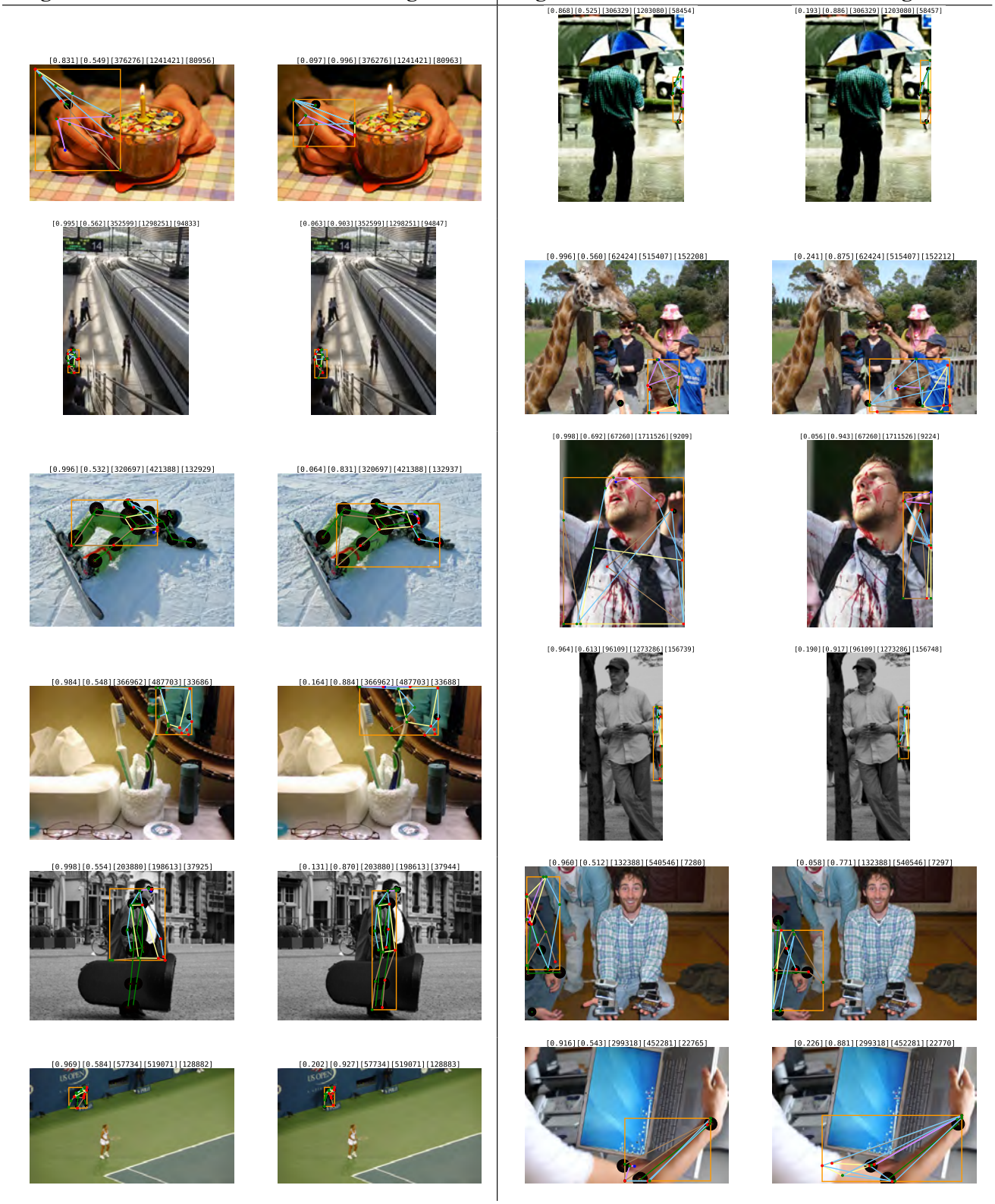

Figure 9: **Top Scoring Errors.** Scoring errors ordered by relevance top to bottom and left to right. Each scoring error consists of a ground-truth annotation and a pair of detections shown side by side, one with high score and low OKS (left), and one with low score and high OKS (right). The relevance is computed as the geometric mean between the difference of the OKS obtained by the two detections and the difference of their confidence score. The ground truth skeleton is in green, and the color coding of the detection skeleton is described in Sec. 1. Each image title contains the following information: [detection_score, OKS, image_id, ground_truth_id, detection_id].

6 Background False Positives

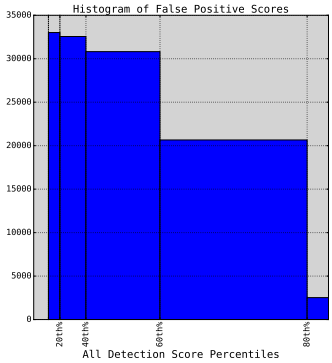

Figure 10: **Histogram of Scores.** Histogram of the confidence scores of all Background False Positives errors. The problematic cases are those with a score falling in the highest percentile of the overall detection scores (rightmost bin).

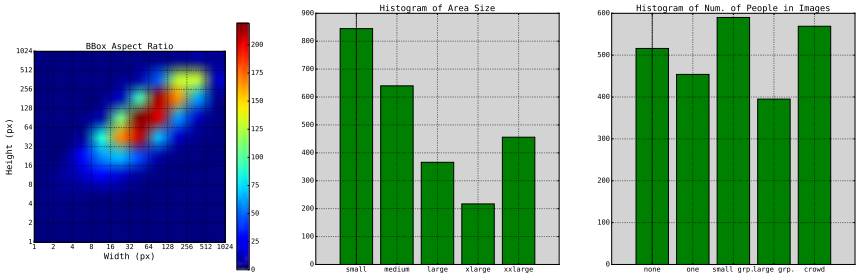

Figure 11: **High Score Background False Positives Analysis.** (Left) Heatmap showing the most frequent Bounding Box Aspect Ratios; (Center) Histogram of the area sizes; (Right) Histogram of the number of people in an image with False Positives. The above plots are computed for the subset of Background False Positives having confidence score in the top-20th percentile of overall scores (rightmost bin in the previous Figure).

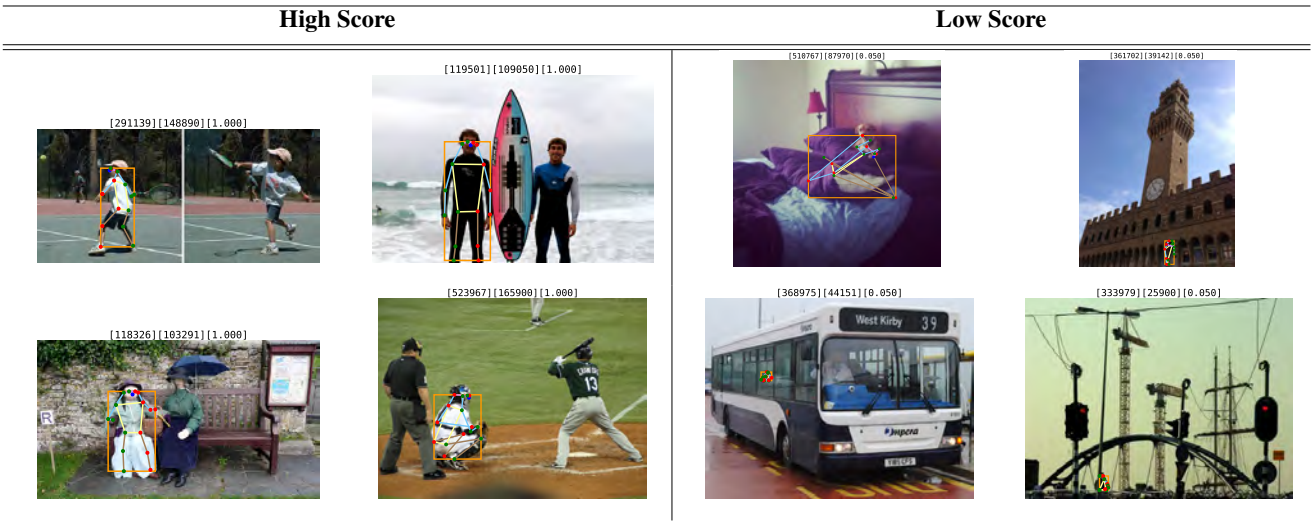

Figure 12: **False Positive Errors.** Errors in the COCO annotations might cause good detections to appear in the above examples.

7 False Negatives

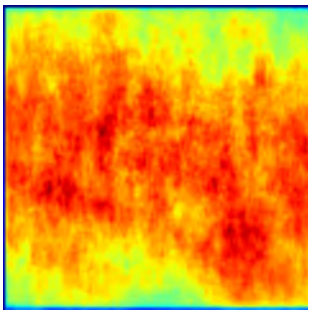

Figure 13: **Background False Negatives Heatmap.** Heatmap of the segmentation masks of all Background False Negatives.

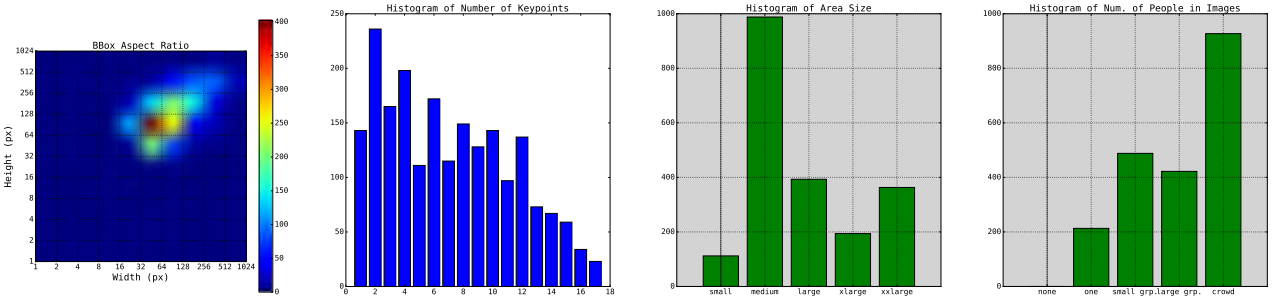

Figure 14: **Background False Negatives Analysis.** (Left) Heatmap showing the most frequent Bounding Box Aspect Ratios; (Center-Left) Histogram of the number of visible keypoints; (Center-Right) Histogram of the area sizes; (Right) Histogram of the number of people in an image with False Negatives.

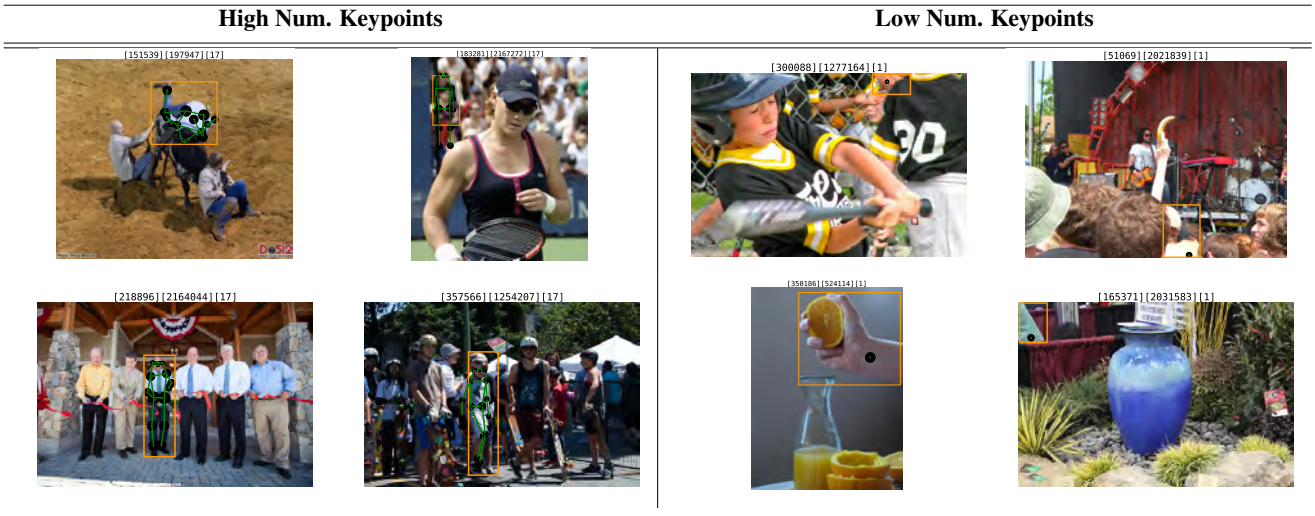

Figure 15: **False Negative Errors.**

8 Performance and Error Sensitivity to Occlusion and Crowding

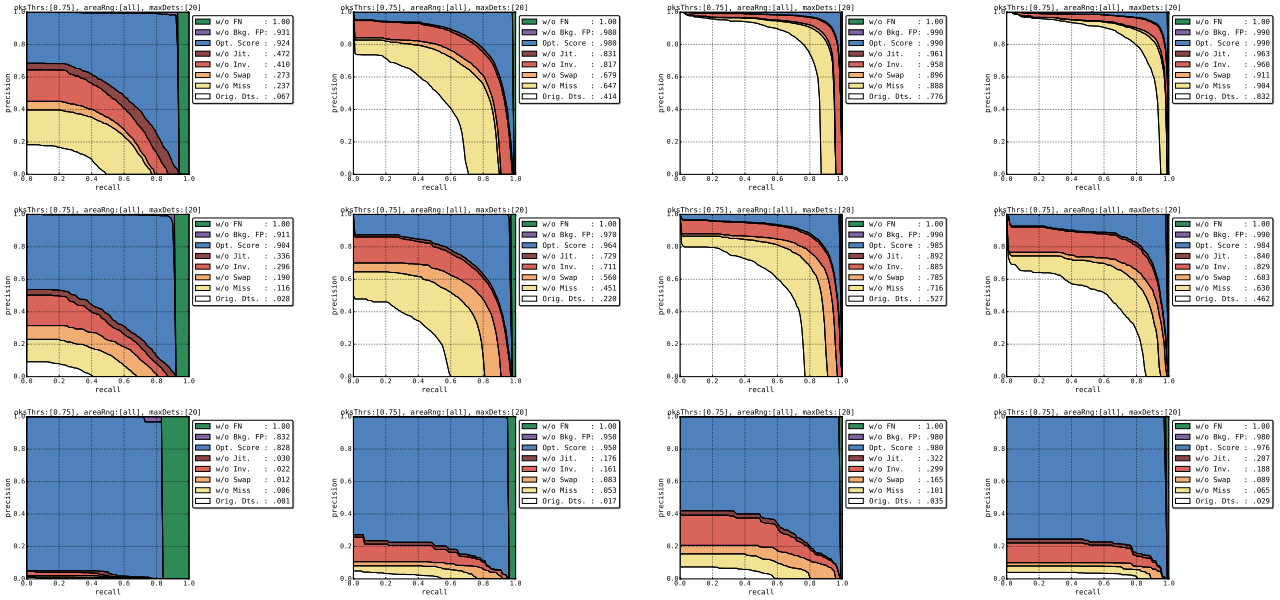

Figure 16: **Performance Sensitivity.** We separate the ground-truth instances in COCO into twelve benchmarks, based on number of visible keypoints (occlusion) and overlap between annotations (crowding), more details are discussed in the Main Paper. We show the Precision Recall Curves with individual errors breakdown obtained by evaluating performance separately on each benchmark. The last row is computed on few instances (since these hard examples are under-represented in COCO), therefore results may have high variance.

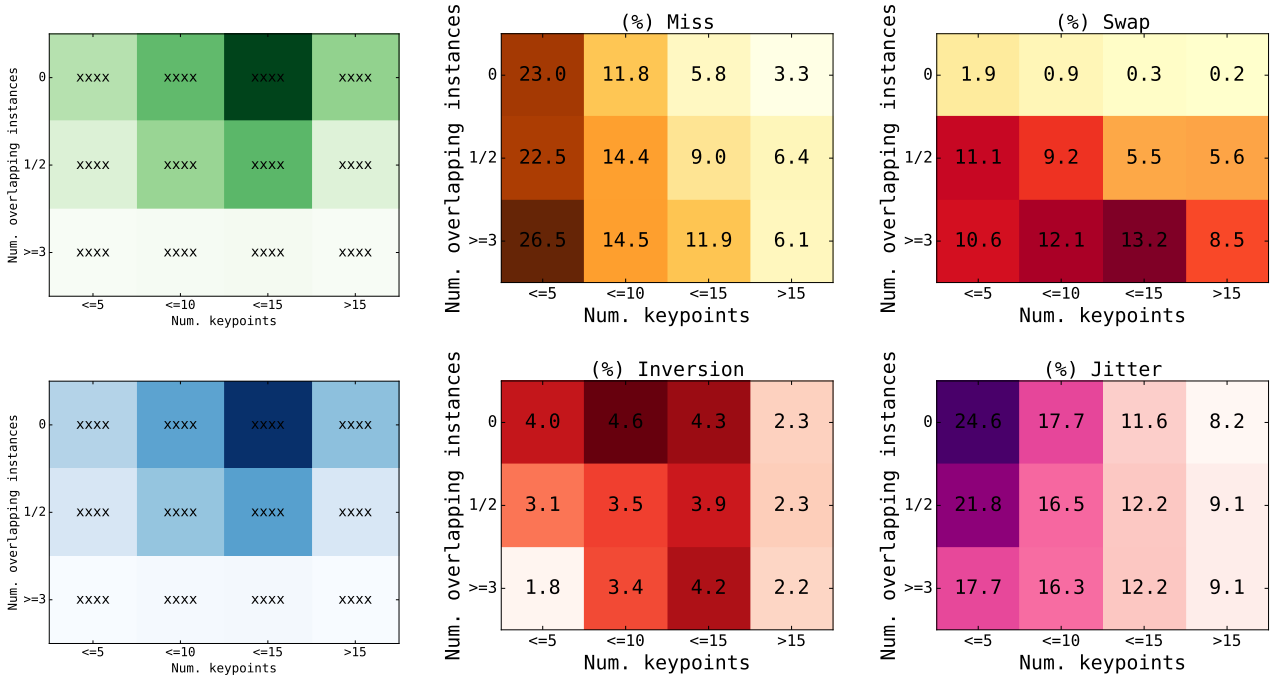

Figure 17: **Localization Error Sensitivity**. (Left Column) The total number of ground-truth instances (top) and keypoints (bottom) present in each Occlusion and Crowding benchmark; (Center and Right Columns) The percentage of localization errors present in the algorithm's detections for each Occlusion and Crowding benchmark.

9 Performance and Error Sensitivity to Instance Size

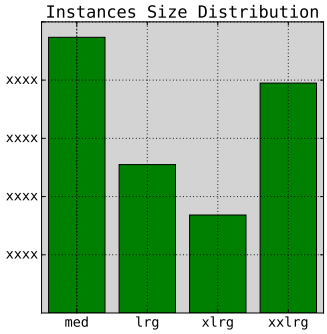

Figure 18: **Instance Size Benchmarks.** We separate the ground-truth instances in COCO into four benchmarks, based on the area size (measured in pixels), more details are discussed in the Main Paper. We show the total number of ground-truth instances in each benchmark.

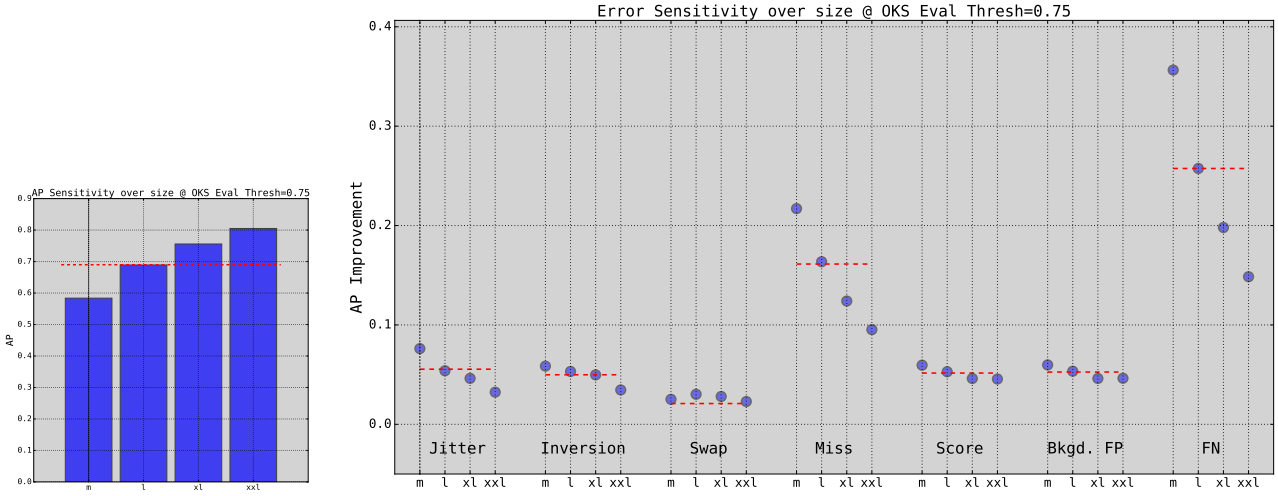

Figure 19: **Sensitivity to Instance Size**. (Left) The AP of an algorithm when evaluating performance separately on each Size benchmark; (Right) The AP improvement when correcting each error type after separately evaluating on each of the Size benchmarks; a higher AP improvement means that an error is present in higher quantities (correcting it causes a greater AP improvement). The red dashed line show the performance when evaluating jointly on the instances of all Size benchmarks.