Supplementary Materials for the ICCV 2017 Paper:
Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation

Matteo Ruggero Ronchi
www.vision.caltech.edu/~mronchi
California Institute of Technology, Pasadena, CA, USA

Pietro Perona
perona@caltech.edu

Abstract

We propose a new method to analyze the impact of errors in algorithms for multi-instance pose estimation and a principled benchmark that can be used to compare them. We define and characterize three main classes of errors - localization, scoring, and background - study how they are influenced by instance attributes and their impact on an algorithm’s performance. Our technique is applied to compare the two leading methods for human pose estimation on the COCO Dataset, measure the sensitivity of pose estimation with respect to instance size, type and number of visible keypoints, clutter due to multiple instances, and the relative score of instances. The performance of algorithms, and the types of error they make, are highly dependent on all these variables, but mostly on the number of keypoints and the clutter. The analysis and software tools we propose offer a novel and insightful approach for understanding the behavior of pose estimation algorithms and an effective method for measuring their strengths and weaknesses.

1. Supplementary Materials

This document accompanies the paper “Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation”.

We provide clarification on how to interpret some of the presented content and illustrate the results of our evaluation analysis on other datasets. Finally, we include the result of our analysis for other methods in addition to those contained in the Main Paper.

- Human Pose and Skeleton Color Coding (Sec. 1.1): Visualization of the color-coding of the human skeleton obtained from a pose estimation algorithm.
- Fine-Grained Precision Recall Plots (Sec. 1.2): In-depth explanation on how to interpret the performance plots computed by our analysis tools.
- Correction of Localization Errors (Sec. 1.3): Visualization of a predicted human skeleton as the localization errors it contains are progressively corrected.
- Multi-Instance Mouse Pose Evaluation (Sec. 1.4): Analysis of the performance of a multi-instance pose estimation algorithm on the Caltech Resident Intruder Mouse dataset CRIM13 [2].
- Performance Analysis Reports (Sec. 1.5): The performance reports obtained by running our analysis code on several algorithms.

1.1. Human Pose and Skeleton Color Coding

We adopt the following color coding to visualize 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.

Figure 1. Human Pose and Skeleton Color Coding.
• Torso keypoints (shoulders, hips) are connected by yellow lines.

• Lower-body keypoints (hips, knees, ankles) are connected by brown lines.

1.2. Fine-Grained Precision Recall Plots

Figure 2. Fine-Grained Error Analysis. We study the errors occurring in multi-instance pose estimation, and provide the tools for a fine-grained description of performance, which allows to quantify the impact of each type of error at a single glance.

Fig. 2 summarizes the impact of all types of error on the performance of a multi-instance pose estimation algorithm. It is composed of a series of Precision Recall (PR) curves where each curve is guaranteed to be strictly higher than the previous as the evaluation setting becomes more permissive. The legend shows the Area Under the Curve (AUC) obtained for each of the following evaluation settings:

- Oks .95, .85: PR curves obtained at the OKS thresholds of .95 and .85 respectively.

The remaining evaluations are performed with the lowest OKS threshold considered in the legend (.85 in this case).

- Miss, Swap, Inv., Jit.: PR curves after the algorithm’s keypoint detections are progressively corrected to remove each type of localization error, as shown in Sec. 1.3. As keypoint localization is corrected, the OKS between a detection and ground-truth match improves, possibly exceeding the current OKS evaluation threshold and becoming an additional True Positive. We show with different colors, the AUC improvements obtained by fixing each type of localization error.

- Score: PR curves after the algorithm’s keypoint detections have been rescored with the optimal confidence score described in the main paper.

- Bkg.: PR curves after all of the algorithm’s background False Positive detections are removed.

- FN: PR curves after all the False Negative errors are ignored.

In the case of the Cmu [3] algorithm, the AUC evaluated at Oks=.95 is only .096, but improves to .488 when lowering the threshold to .85. At this threshold, correcting all the miss errors results in a large improvement of the AUC to .628. Smaller AUC gains are obtained when correcting swaps, .681, and inversions, .775. Another large improvement is obtained when jitter errors are removed, resulting in an AUC of .900. This shows what would the performance of [3] be if it had a perfect localization of keypoints. When localization is very good, the impact of scoring is not as significant, but still results in an AUC improvement of about 2%. Optimally scoring detections greatly diminishes the impact of Background False Positives, as detections rarely remain unmatched. Finally, removing background False Negatives provides the remaining AUC to obtain perfect performance. In summary, Cmu’s errors are dominated by imperfect localization, mostly miss and jitter errors, and missed detections.

1.3. Sequential Correction of Localization Errors

The fine-grained PR curves shown in Fig. 2 are obtained by fixing an OKS threshold and evaluating the performance of an algorithm after progressively correcting its detections. To do so, we compute for every predicted keypoint what is the Keypoint Similarity (KS) with its corresponding ground-truth body part, and with different ground-truth body parts of the same person, and of other people in the image. This allows us to define the types of localization error, as done in Sec. 3.1 of the main paper, and correct them.

- Miss errors are corrected by repositioning a keypoint prediction on the .5 KS circle centered on the true location; left-elbow and wrists in Fig. 3.

- Swap and Inversion errors are corrected by repositioning a keypoint prediction at a distance from the correct ground-truth location so that the new value of KS is the same that the prediction had with the wrong body part it mistakenly detected (belonging to a different/same person for swap/inversion); in Fig. 3 the right-elbow is a swap, right-knee is an inversion.

- Jitter errors are corrected by repositioning a keypoint prediction on the .85 KS circle centered on the true location; left-ankle in Fig. 3.

Miss and jitter errors are corrected by bringing a prediction to a fixed distance from its true position. The new location of swaps and inversions depends instead on how good was the prediction of the wrong joint: after correction, a good/bad prediction of the wrong body part becomes a good/bad prediction (high/low KS) of the true body part.
1.4. Multi-Instance Mouse Pose Evaluation

The study of multi-instance pose estimation errors and performance conducted in the main paper extends beyond humans, to any object category where the location of parts is estimated along with a detection, and to situations where cluttered scenes may contain multiple object instances. This is common in fine-grained categorization, i.e. birds [1], or animal behavior analysis, i.e. mice [2] and flies [4], where part alignment is often crucial. To show the versatility of our software tools, we evaluated the performance of a top-down pose estimation algorithm on the CRIM13 [2] dataset, which consists of images of pairs of mice (a black resident and a white intruder) engaging in social behavior. For our experiment, we used 10000 images, separated into a Training, Validation and Test sets of 8500, 500 and 1000 images, for which human annotations of 7 keypoint locations (nose, ears, neck, hips, tail) were available. During evaluation, the Keypoint Similarity metric and the OKS between a detection and an annotation are computed in the same way described in Sec. 2.2 of the main paper. Fig. 4.(Left), shows the performance of a top-down method, composed of a Multi-Box object detector [6] to find each mouse, followed by a stacked hourglass network [7] for predicting the keypoint locations. Results indicate that the two predominant errors are jitter and miss; swap errors have a very limited impact and occur during interactions which results in some amount of occlusion, Fig. 4.(Right); inversion errors are mostly absent. The scoring of detections is not critical, as in the human case, since images always contains exactly two mice. Because of the fairly simple, clutter-free and fixed-viewpoint image capture settings, background errors (False Positives and False Negatives) are mostly absent.
1.5. Performance Analysis Reports

In the following pages we include the performance reports\(^1\) generated by the released analysis code\(^2\) on the following algorithms:

- **Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [3]**
- **Towards Accurate Multi-person Pose Estimation in the Wild [8]**
- **Mask R-CNN [5]**

---

\(^1\)Some values have been obfuscated to preserve the sanctity of the COCO test-dev split. Check the Main Paper to find the corresponding values on the COCO training set.

\(^2\)Available for download at: [https://goo.gl/9EyDyN](https://goo.gl/9EyDyN)
1 Human Pose and 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.

Figure 1: Detection’s Skeleton Color Coding.
2 Overall Detector Characteristics

- Num. Detections: 46256
- Num. Images [with Detections]: 20288 [11940]

![Graphs showing precision recall curves for different OKS thresholds and area ranges.](image)

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

3 Error Impact on AP

![Graphs showing AP improvement for different error types.](image)

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.

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 5: Human Keypoint Breakdown. The frequency of each localization error for every keypoint of the human body.
<table>
<thead>
<tr>
<th>Miss</th>
<th>Swap</th>
<th>Inversion</th>
<th>Jitter</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Miss Image" /></td>
<td><img src="image2.png" alt="Swap Image" /></td>
<td><img src="image3.png" alt="Inversion Image" /></td>
<td><img src="image4.png" alt="Jitter Image" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Miss Image" /></td>
<td><img src="image6.png" alt="Swap Image" /></td>
<td><img src="image7.png" alt="Inversion Image" /></td>
<td><img src="image8.png" alt="Jitter Image" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Miss Image" /></td>
<td><img src="image10.png" alt="Swap Image" /></td>
<td><img src="image11.png" alt="Inversion Image" /></td>
<td><img src="image12.png" alt="Jitter Image" /></td>
</tr>
</tbody>
</table>

**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).

<table>
<thead>
<tr>
<th>High Score</th>
<th>Low Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="High Score Image" /></td>
<td><img src="image2" alt="Low Score Image" /></td>
</tr>
<tr>
<td><img src="image3" alt="High Score Image" /></td>
<td><img src="image4" alt="Low Score Image" /></td>
</tr>
<tr>
<td><img src="image5" alt="High Score Image" /></td>
<td><img src="image6" alt="Low Score Image" /></td>
</tr>
<tr>
<td><img src="image7" alt="High Score Image" /></td>
<td><img src="image8" alt="Low Score Image" /></td>
</tr>
<tr>
<td><img src="image9" alt="High Score Image" /></td>
<td><img src="image10" alt="Low Score Image" /></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>High Num. Keypoints</th>
<th>Low Num. Keypoints</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
</tbody>
</table>

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.
1 Human Pose and 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.

Figure 1: Detection’s Skeleton Color Coding.
2 Overall Detector Characteristics

- Num. Detections: 64438
- Num. Images [with Detections]: 20288 [14634]

![Precision Recall Curves](image)

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

3 Error Impact on AP

![AP Improvement](image)

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.

Localization Errors Taxonomy:

- **Jitter**: small error around the correct keypoint location.
- **Miss**: large localization error, the detected keypoint is not within the vicinity 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 vicinity of a body part belonging to a different person.

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

Figure 1: Detection’s Skeleton Color Coding.
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.

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 5: Human Keypoint Breakdown. The frequency of each localization error for every keypoint of the human body.
<table>
<thead>
<tr>
<th>Miss</th>
<th>Swap</th>
<th>Inversion</th>
<th>Jitter</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Miss Image" /></td>
<td><img src="image2" alt="Swap Image" /></td>
<td><img src="image3" alt="Inversion Image" /></td>
<td><img src="image4" alt="Jitter Image" /></td>
</tr>
<tr>
<td><img src="image5" alt="Miss Image" /></td>
<td><img src="image6" alt="Swap Image" /></td>
<td><img src="image7" alt="Inversion Image" /></td>
<td><img src="image8" alt="Jitter Image" /></td>
</tr>
<tr>
<td><img src="image9" alt="Miss Image" /></td>
<td><img src="image10" alt="Swap Image" /></td>
<td><img src="image11" alt="Inversion Image" /></td>
<td><img src="image12" alt="Jitter Image" /></td>
</tr>
<tr>
<td><img src="image13" alt="Miss Image" /></td>
<td><img src="image14" alt="Swap Image" /></td>
<td><img src="image15" alt="Inversion Image" /></td>
<td><img src="image16" alt="Jitter Image" /></td>
</tr>
<tr>
<td><img src="image17" alt="Miss Image" /></td>
<td><img src="image18" alt="Swap Image" /></td>
<td><img src="image19" alt="Inversion Image" /></td>
<td><img src="image20" alt="Jitter Image" /></td>
</tr>
<tr>
<td><img src="image21" alt="Miss Image" /></td>
<td><img src="image22" alt="Swap Image" /></td>
<td><img src="image23" alt="Inversion Image" /></td>
<td><img src="image24" alt="Jitter Image" /></td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>High Score</th>
<th>Low Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="High Score Example" /></td>
<td><img src="image2.png" alt="Low Score Example" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="High Score Example" /></td>
<td><img src="image4.png" alt="Low Score Example" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="High Score Example" /></td>
<td><img src="image6.png" alt="Low Score Example" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="High Score Example" /></td>
<td><img src="image8.png" alt="Low Score Example" /></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>High Num. Keypoints</th>
<th>Low Num. Keypoints</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
</tbody>
</table>

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.