Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation

Matteo Ruggero Ronchi
www.vision.caltech.edu/~mronchi

Pietro Perona
perona@caltech.edu

California Institute of Technology, Pasadena, CA, USA

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 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. Introduction

Estimating the pose of a person from a single monocular frame is a challenging task due to many confounding factors such as perspective projection, the variability of lighting and clothing, self-occlusion, occlusion by objects, and the simultaneous presence of multiple interacting people. Nevertheless, the performance of human pose estimation algorithms has recently improved dramatically, thanks to the development of suitable deep architectures [9, 11, 12, 17, 28, 29, 32, 33, 42, 44] and the availability of well-annotated image datasets, such as *MPII Human Pose Dataset* and *COCO* [4, 27]. There is broad consensus that performance is saturated on simpler single-person datasets [23, 24], and researchers' focus is shifting towards less constrained and more challenging datasets [4, 14, 27], where images may contain multiple instances of people, and a variable number of body parts (or keypoints) are visible. However, evaluation is challenging: more complex datasets make it harder to benchmark algorithms due to the many sources of error that may affect performance, and existing
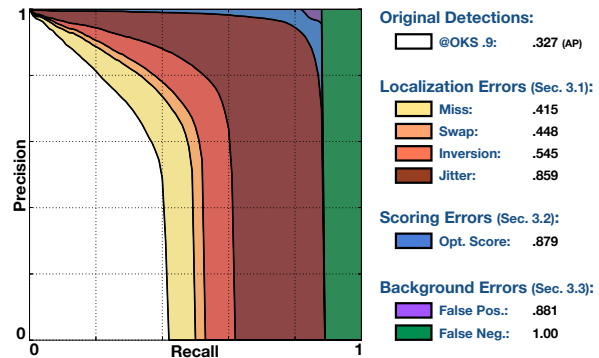

Figure 1. **Coarse to Fine Error Analysis.** We study the errors occurring in multi-instance pose estimation, and how they are affected by physical characteristics of the portrayed people. We build upon currently adopted evaluation metrics and provide the tools for a fine-grained description of performance, which allows to quantify the impact of different types of error at a single glance. The fine-grained Precision-Recall curves are obtained by fixing an OKS threshold and evaluating the performance of an algorithm after progressively correcting its mistakes.

metrics, such as Average Precision (AP) or mean Percentage of Correct Parts (mPCP), hide the underlying causes of error and are not sufficient for truly understanding the behaviour of algorithms.

Our goal is to propose a principled method for analyzing pose algorithms' performance. We make four contributions:
1. Taxonomization of the types of error that are typical of the multi-instance pose estimation framework;
2. Sensitivity analysis of these errors with respect to measures of image complexity;
3. Side-by-side comparison of two leading human pose estimation algorithms highlighting key differences in behaviour that are hidden in the average performance numbers;
4. Assessment of which types of datasets and benchmarks may be most productive in guiding future research.

Our analysis extends beyond humans, to any object category where the location of parts is estimated along with detections, and to situations where cluttered scenes may contain multiple object instances. This is common in fine-grained categorization [8], or animal behavior analysis [10, 16], where part alignment is often crucial.

2. Related Work

2.1. Error Diagnosis

Object Detection: Hoiem et al. [19] studied how a detailed error analysis is essential for the progress of recognition research, since standard benchmark metrics do not tell us *why* certain methods outperform others and *how* could they be improved. They determined that several modes of failure are due to different types of error and highlighted the main confounding factors for object detection algorithms. While [19] pointed out the value of discriminating between different errors, it did not show how to do so in the context of pose estimation, which is one of our contributions.

Pose Estimation: In their early work on pose regression, Dollár et al. [13] observed that unlike human annotators, algorithms have a distribution of the normalized distances between a part detection and its ground-truth that is typically bimodal, highlighting the presence of multiple error modes. The *MPII Human Pose Dataset* [4] Single-Person benchmark enables the evaluation of the performance of algorithms along a multitude of dimensions, such as 45 pose priors, 15 viewpoints and 20 human activities. However, none of the currently adopted benchmarks for Multi-Person pose estimation [14, 27, 31] carry out an extensive error and performance analysis specific to this framework, and mostly rely on the metrics from the Single-Person case. No standards for performing or compactly summarizing detailed evaluations has yet been defined, and as a result only a coarse comparison of algorithms can be carried out.

2.2. Evaluation Framework

We conduct our study on *COCO* [27] for several reasons: (i) it is the largest collection of multi-instance person keypoint annotations; (ii) performance on it is far from saturated and conclusions on such a large and non-iconic dataset can generalize to easier datasets; (iii) adopting their framework, with open source evaluation code, a multitude of datasets built on top of it, and annual competitions, will have the widest impact on the community. The framework involves simultaneous person detection and keypoint estimation, and the evaluation mimics the one used for object detection, based on Average Precision and Recall (AP, AR). Given an image, a distance measure is used to match algorithm detections, sorted by their confidence score, to ground-truth annotations. For bounding-boxes and segmentations, the distance of a detection and annotation pair is measured by their Intersection over Union. In the keypoint estimation task, a new metric called Object Keypoint Similarity (OKS) is defined. The OKS between a detection $\hat{\theta}^{(p)}$ and the annotation $\theta^{(p)}$ of a person p, Eq. 1, is the average over the labeled parts in the ground-truth ($v_i = 1, 2$), of the *Keypoint Similarity* between corresponding keypoint pairs, Fig. 2; unlabeled parts ($v_i = 0$) do not affect the OKS [2].
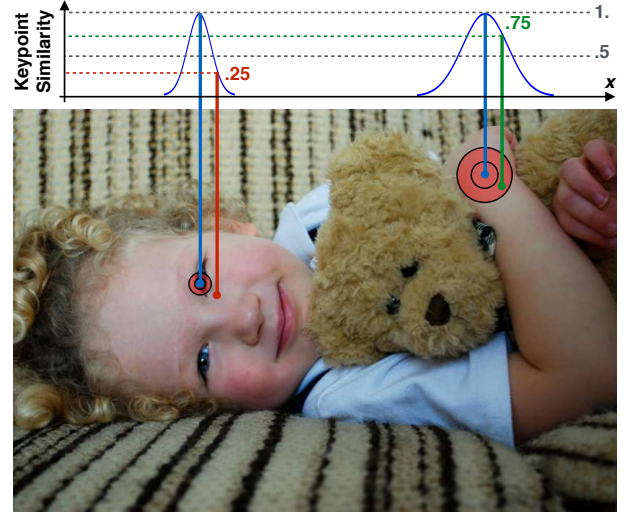

Figure 2. **Keypoint Similarity (*ks*).** The *ks* between two detections, eye (red) and wrist (green), and their corresponding ground-truth (blue). The red concentric circles represent *ks* values of .5 and .85 in the image plane and their size varies by keypoint type, see Sec.2.2. As a result, detections at the same distance from the corresponding ground-truth can have different *ks* values.

$$\begin{cases} ks(\hat{\theta}_i^{(p)}, \theta_i^{(p)}) & = e^{-\frac{||\hat{\theta}_i^{(p)} - \theta_i^{(p)}||_2^2}{2s^2 k_i^2}} \\ OKS(\hat{\theta}^{(p)}, \theta^{(p)}) & = \frac{\sum_i ks(\hat{\theta}_i^{(p)}, \theta_i^{(p)}) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \end{cases} \quad (1)$$

The *ks* is computed by evaluating an un-normalized Gaussian function, centered on the ground-truth position of a keypoint, at the location of the detection to evaluate. The Gaussian's standard deviation k_i is specific to the keypoint type and is scaled by the area of the instance s, measured in pixels, so that the OKS is a perceptually meaningful and easy to interpret similarity measure. For each keypoint type, k_i reflects the consistency of human observers clicking on keypoints of type i and is computed from a set of 5000 redundantly annotated images [2].

To evaluate an algorithm's performance, its detections within each image are ordered by confidence score and assigned to the ground-truths that they have the highest OKS with. As matches are determined, the pool of available annotations for lower scored detections is reduced. Once all matches have been found, they are evaluated at a certain OKS threshold (ranging from .5 to .95 in [1]) and classified as True or False Positives (above or below threshold), and unmatched annotations are counted as False Negatives. Overall AP is computed as in the *PASCAL VOC Challenge* [15], by sorting the detections across all the images by confidence score and averaging precision over a predefined set of 101 recall values. AR is defined as the maximum recall given a fixed number of detections per image [20]. Finally, we will refer to cocoAP and cocoAR when AP and AR are additionally averaged over all OKS threshold values (.5:.05:.95), as done in the *COCO* framework [1].

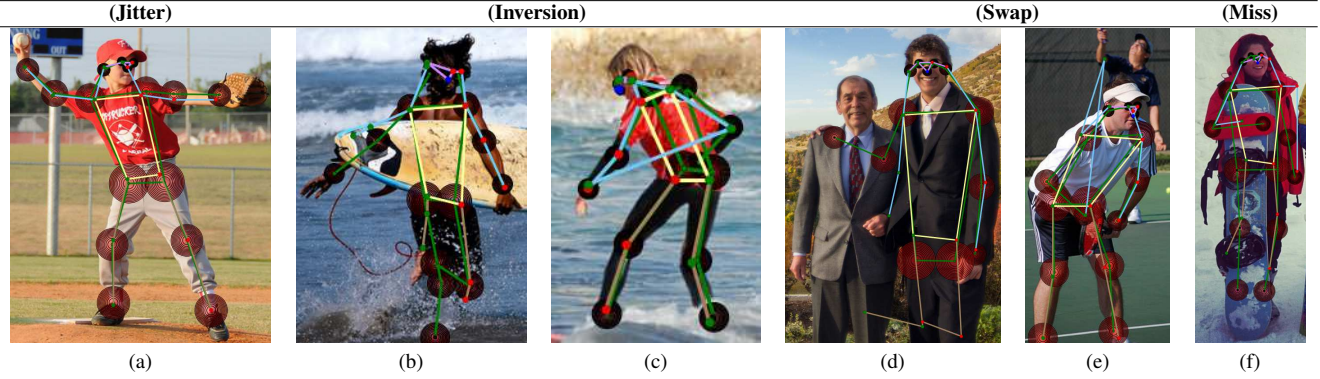| (Jitter)	(Inversion)		(Swap)		(Miss)
(a)	(b)	(c)	(d)	(e)	(f)

Figure 3. **Taxonomy of Keypoint Localization Errors.** Keypoint localization errors, Sec. 3.1, are classified based on the position of a detection as, *Jitter*: in the proximity of the correct ground-truth location, but not within the human error margin - left hip in (a); *Inversion*: in the proximity of the ground-truth location of the wrong body part - inverted skeleton in (b), right wrist in (c); *Swap*: in the proximity of the ground-truth location of the body part of a wrong person - right wrist in (d), right elbow in (e); *Miss*: not in the proximity of any ground-truth location - both ankles in (f). While errors in (b,d) appear to be more excusable than those in (c,e) they have the same weight. Color-coding: (ground-truth) - concentric red circles centered on each keypoint's location connected by a green skeleton; (prediction) - red/green dots for left/right body part predictions connected with colored skeleton, refer to the Appendix for an extended description.
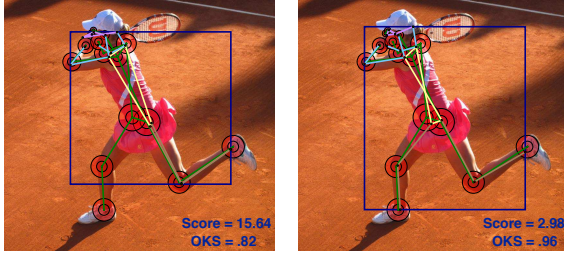

Figure 4. **Instance Scoring Error.** The detection with highest confidence score (Left) is associated to the closest ground-truth by the evaluation algorithm described in Sec. 2.2. However, its OKS is lower than the OKS of another detection (Right). This results in a loss in performance at high OKS thresholds, details in Sec. 3.2.

2.3. Algorithms

We conduct our analysis on the the top-two ranked algorithms [29, 11] of the *2016 COCO Keypoints Challenge* [1], and observe the impact on performance of the design differences between a top-down and a bottom-up approach.

Top-down (instance to parts) methods first detect humans contained in an image, then try to estimate their pose separately within each bounding box [14, 17, 32, 44]. The **Grmi** [29] algorithm is a two step cascade. In the first stage, a Faster-RCNN system [34] using ResNet-Inception architecture [37] combining inception layers [38] with residual connections [18] is used to produce a bounding box around each person instance. The second stage serves as a refinement where a ResNet with 101 layers [18] is applied to the image crop extracted around each detected person instance in order to localize its keypoints. The authors adopt a combined classification and regression approach [39, 34]: for each spatial position, first a classification problem is solved to determine whether it is in the vicinity of each of the keypoints of the human body, followed by a regression problem

Table 1. *2016 COCO Keypoints Challenge Leaderboard* [1]

	Cmu	Grmi	DL61	R4D	Umichvl
cocoAP	0.608	0.598	0.533	0.497	0.434

to predict a local offset vector for a more precise estimate of the exact location. The results of both stages are aggregated to produce highly localized activation maps for each keypoint in the form of a voting process: each point in a detected bounding box casts a vote with its estimate for the position of every keypoint, and the vote is weighted by the probability that it lays near the corresponding keypoint.

Bottom-up (parts to instance) methods first separately detect all the parts of the human body from an image, then try to group them into individual instances [9, 28, 33, 42]. The **Cmu** [11] algorithm estimates the pose for all the people in an image by solving body part detection and part association jointly in one end-to-end trainable network, as opposed to previous approaches that train these two tasks separately [22, 31] (typically part detection is followed by graphical models for the association). Confidence maps with gaussian peaks in the predicted locations, are used to represent the position of individual body parts in an image. Part Affinity Fields (PAFs) are defined from the confidence maps, as a set of 2D vector fields that jointly encode the *location* and *orientation* of a particular limb at each position in the image. The authors designed a two-branch VGG [36] based architecture, inspired from CPMs [42], to iteratively refine confidence maps and PAFs with global spatial contexts. The final step consists of a maximum weight bipartite graph matching problem [43, 26] to associate body parts candidates and assemble them into full body poses for all the people in the image. A greedy association algorithm over a minimum spanning tree is used to group the predicted parts into consistent instance detections.
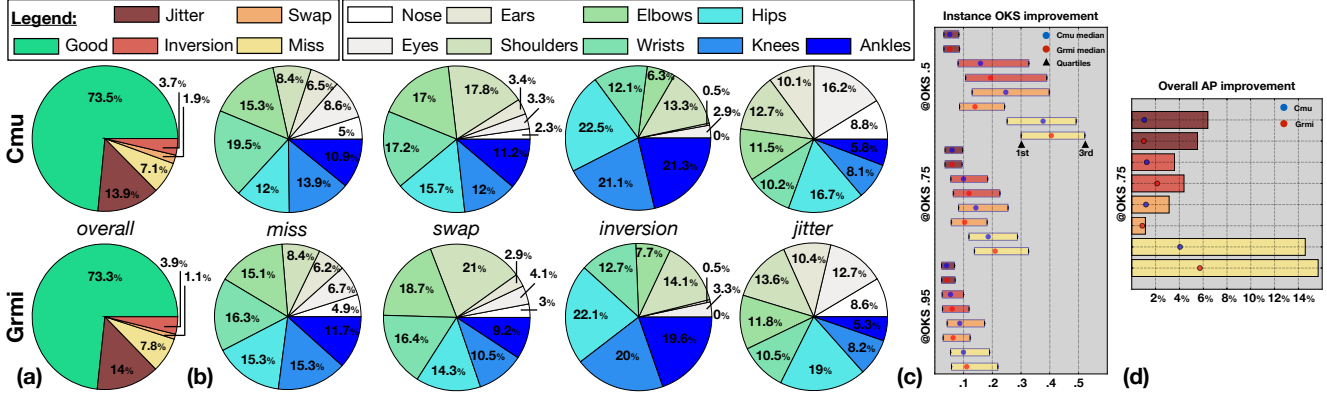
Figure 5. **Distribution and Impact of Localization Errors.** (a) Outcome for the predicted keypoints: *Good* indicates correct localization. (b) The breakdown of errors over body parts. (c) The algorithm's detections OKS improvement obtained after separately correcting errors of each type; evaluated over all the instances at OKS thresholds of .5, .75 and .95; the dots show the median, and the bar limits show the first and third quartile of the distribution. (d) The AP improvement obtained after correcting localization errors; evaluated at OKS thresholds of .75 (bars) and .5 (dots). A larger improvement in (c) and (d) shows what errors are more impactful. See Sec. 3.1 for details.

3. Multi-Instance Pose Estimation Errors

We propose a taxonomy of errors specific to the multi-instance pose estimation framework: (i) **Localization**, Fig. 3, due to the poor localization of the keypoint predictions belonging to a detected instance; (ii) **Scoring**, Fig. 4, due to a sub-optimal confidence score assignment; (iii) **Background False Positives**, detections without a ground-truth annotation match; (iv) **False Negatives**, missed detections. We assess the causes and impact on the behaviour and performance of [11, 29] for each error type.

3.1. Localization Errors

A localization FP occurs when the location of the keypoints in a detection results in an OKS score with the corresponding ground-truth match that is lower than the evaluation threshold. They are typically due to the fact that body parts are difficult to detect because of self occlusion or occlusion by other objects. We define four types of localization errors, visible in Fig. 3, as a function of the keypoint similarity $ks(.,.)$, Eq. 1, between the keypoint i of a detection $\hat{\theta}_i^{(p)}$ and j of the annotation $\theta_j^{(p)}$ of a person p.

Jitter: small error around the correct keypoint location.

$$.5 \leq ks(\hat{\theta}_i^{(p)}, \theta_i^{(p)}) < .85$$

The limits can be chosen based on the application of interest; in the *COCO* framework, .5 is the smallest evaluation threshold, and .85 is the threshold above which also human annotators have a significant disagreement (around 30%) in estimating the correct position [2].

Miss: large localization error, the detected keypoint is not within the proximity of any body part.

$$ks(\hat{\theta}_i^{(p)}, \theta_j^{(q)}) < .5 \quad \forall q \in \mathcal{P} \quad \text{and} \quad \forall j \in \mathcal{J}$$

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.

$$ks(\hat{\theta}_i^{(p)}, \theta_i^{(p)}) < .5$$
$$\exists j \in \mathcal{J} \quad | \quad ks(\hat{\theta}_i^{(p)}, \theta_j^{(p)}) \geq .5$$

In our study we only consider inversions between the left and right parts of the body, however, the set of keypoints $\mathcal{J}$ can be arbitrarily defined to study any kind of inversion.

Swap: confusion between semantically similar parts of different instances. The detection is within the proximity of a body part belonging to a different person.

$$ks(\hat{\theta}_i^{(p)}, \theta_i^{(p)}) < .5$$
$$\exists j \in \mathcal{J} \quad \text{and} \quad \exists q \in \mathcal{P} \quad | \quad ks(\hat{\theta}_i^{(p)}, \theta_j^{(q)}) \geq .5$$

Every keypoint detection having a keypoint similarity with its ground-truth that exceeds .85 is considered *good*, as it is within the error margin of human annotators. We can see, Fig. 5.(a), that about 75% of both algorithm's detections are *good*, and while the percentage of *jitter* and *inversion* errors is approximately equal, [11] has twice as many *swaps*, and [29] has about 1% more *misses*. Fig. 5.(b) contains a breakdown of errors over keypoint type: faces are easily detectable (smallest percentage of *miss* errors); *swap* errors are focused on the upper-body, as interactions typically involve some amount of upper-body occlusion; the lower-body is prone to *inversions*, as people often self-occlude their legs, and there are less visual cues to distinguish left from right; finally *jitter* errors are predominant on the hips. There are no major differences between the two algorithms in the above trends, indicating that none of the methods contains biases over keypoint type. After defining and identifying localization errors, we measure the improvement in performance resulting from their correction.

Overall AP Improvement — Cmu, Grmi; average at all OKS values; @OKS .5 @OKS .75 @OKS .95 (a)

Histogram of Detection Scores — Low OKS Detections, Maximum OKS Detections; Cmu, Grmi (b) Original Confidence Scores (c) Optimal Confidence Scores
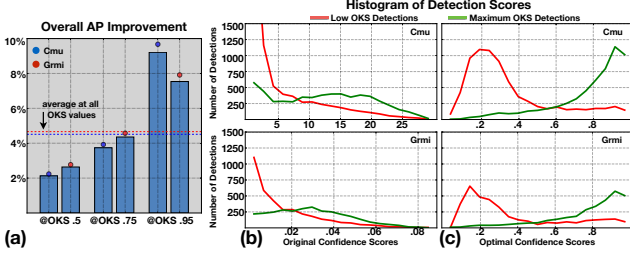
Figure 6. **Scoring Errors Analysis.** (a) The AP improvement obtained when using the optimal detection scores, as defined in Sec. 3.2. The histogram of detections' (b) original and (c) optimal confidence scores. We histogram separately the scores of detections achieving the maximum OKS with a given ground-truth instance (green) and the other detections achieving OKS of at least .1 (red). High overlap of the histograms, as in (b), is caused by the presence of many detections with high OKS and low score or vice versa; a large separation, as in (c), is indication of a better score.

Localization errors are corrected by repositioning a keypoint prediction at a distance from the true keypoint location equivalent to a *ks* of .85 for *jitter*, .5 for *miss*, and at a distance from the true keypoint location equivalent to the prediction's distance from the wrong body part detected in the case of *inversion* and *swap*[1]. Correcting localization errors results in an improvement of the OKS of every instance and the overall AP, as some detections become True Positives (TP) because the increased OKS value exceeds the evaluation threshold. Fig. 5.(c) shows the OKS improvement obtainable by correcting errors of each type: it is most important to correct *miss* errors, followed by *inversions* and *swaps*, while *jitter* errors, although occurring most frequently, have a small impact on the OKS. We learn, Fig. 5.(d), that *misses* are the most costly error in terms of AP ($\sim 15\%$), followed by *inversions* ($\sim 4\%$), relative to their low frequency. We focus on the improvement at the .75 OKS threshold, as it has almost perfect correlation with the value of cocoAP (average of AP over all thresholds) [21]. Changing the evaluation threshold changes the impact of errors (for instance by lowering it to .5 more detections are TP so there is less AP improvement from their correction), but the same relative trends are verified, indicating that the above observations reflect the behavior of the methods and are not determined by the strictness of evaluation.

3.2. Scoring Errors

Assigning scores to instances is a typical task in object detection, but a novel challenge for keypoint estimation. A scoring error occurs when two detections $\hat{\theta}_1^{(p)}$ and $\hat{\theta}_2^{(p)}$ are in the proximity of a ground-truth annotation $\theta^{(p)}$ and the one with the highest confidence has the lowest OKS:

$$\begin{cases} Score(\hat{\theta}_1^{(p)}) & > Score(\hat{\theta}_2^{(p)}) \\ OKS(\hat{\theta}_1^{(p)}, \theta^{(p)}) & < OKS(\hat{\theta}_2^{(p)}, \theta^{(p)}) \end{cases}$$

---

[1]The Appendix contains examples showing how errors are corrected.

Table 2. Improvements due to the optimal rescoring of detections.

	Cmu [11]	**Grmi** [29]
Imgs. w. detections	11940	14634
Imgs. w. optimal detection order	7456 (62.4%)	9934 (67.8%)
Number of Scoring Errors	407	82
Increase of Matches	64	156
Matches with OKS Improvement	590	430

This can happen in cluttered scenes when many people and their detections are overlapping, or in the case of an isolated person for which multiple detections are fired, Fig. 4. Confidence scores affect evaluation, Sec. 2.2, *locally* by determining the order in which detections get matched to the annotations in an image, and *globally*, when detections are sorted across the whole dataset to compute AP and AR. As a result, it is important for the detection scores to be: (i) *'OKS monotonic increasing'*, so that a higher score always results in a higher OKS; (ii) calibrated, so that scores reflect as much as possible the probability of being a TP. A score possessing such properties is *optimal*, as it achieves the highest performance possible for the provided detections. It follows that the optimal score for a given detection corresponds to the maximum OKS value obtainable with any ground-truth annotation: monotonicity and perfect calibration are both guaranteed, as higher OKS detections would have higher score, and the OKS is an exact predictor of the quality of a detection. The optimal scores can be computed at evaluation time, by an oracle assigning to each detection a confidence corresponding to the maximum OKS score achievable with any ground-truth instance. To aid performance in the case of strong occlusion, we apply Soft-Non-Max-Suppression [5], which decays the confidence scores of detections as a function of the amount of reciprocal overlap.

Using optimal scores yields about 5% AP improvement, averaged at all the OKS evaluation thresholds, and up to 10% at OKS .95, Fig. 6.(a), pointing to the importance of assigning low scores to unmatched detections. A careful examination shows that the reason of the improvement is two-fold, Tab. 2: (i) there is an increase in the number of matches between detections and ground-truth instances (reduction of *FP* and *FN*) and (ii) the existing matches obtain a higher OKS value. Both methods have a significant amount of overlap, Fig. 6.(b), between the histogram of *original* scores for the detections with the highest OKS with a given ground-truth (green line) and all other detections with a lower OKS (red line). This indicates the presence of many detections with high OKS and low score or vice versa. Fig. 6.(c) shows the effect of rescoring: *optimal score* distributions are bi-modal and present a large separation, so confidence score is a better OKS predictor. Although the AP improvement after rescoring is equivalent, [29] provides scores that are in the same order as the optimal ones for a higher percentage of images and makes less errors, indicating that it is using a better scoring function.

N. of Keypoints across columns: ∈ [1, 5], ∈ [6, 10], ∈ [11, 15], ∈ [16, 17]

N. of Overlaps across rows: 0, ∈ [1, 2], ≥ 3

Figure 7. **Images from the *COCO* Dataset Benchmarks.** We separate the ground-truth instances in the *COCO* dataset into twelve benchmarks, based on number of visible keypoints and overlap between annotations; Fig. 10.(b) shows the size of each benchmark.
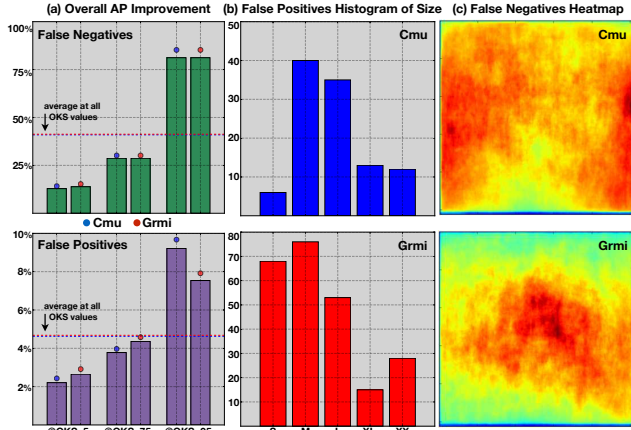

Figure 8. **Background Errors Analysis.** (a) The AP improvement obtained after *FN* (top) and *FP* (bottom) errors are removed from evaluation; horizontal lines show the average value for each method. (b) The histogram of the area size of *FP* having a high confidence score. (c) The heatmaps obtained by adding the resized ground-truth *COCO* segmentation masks of all the *FN*.

3.3. Background False Positives and False Negatives

FP and *FN* respectively consist of an algorithm's detections and the ground-truth annotations that remain unmatched after evaluation is performed. *FP* typically occur when objects resemble human features or when body parts of nearby people are merged into a wrong detection. Most of the *FP* errors could be resolved by performing better Non-Max-Suppression and scoring, since their impact is greatly reduced when using optimal scores, i.e. Fig. 1. Small size and low number of visible keypoints are instead the main cause of *FN*. In Fig. 8.(a) we show the impact

of background errors on the AP at three OKS evaluation thresholds: FN affect performance significantly more than *FP*, on average about 40% versus only 5%. For both methods, the average number of people in images containing *FP* and *FN* is about 5 and 7, compared to the dataset's average of 3, suggesting that cluttered scenes are more prone to having background errors. Interestingly, the location of *FN* errors for the two methods differs greatly, Fig.8.(c): [11] predominantly misses annotations around the image border, while [29] misses those at the center of an image. Another significant difference is in the quantity of *FP* detections having a high confidence score (in the top-20th percentile of overall scores), Fig.8.(b): [29] has more than twice the number, mostly all with small pixel area size ($< 32^2$).

4. Sensitivity to Occlusion, Crowding and Size

One of the goals of this study is to understand how the layout of people portrayed in images, such as the number of visible keypoints (*occlusion*), the amount of overlap between instances (*crowding*) and size, affects the errors and performance of algorithms. This section is focused on the properties of the data, so we analyze only on method, [11]. The *COCO* Dataset contains mostly visible instances having little overlap: Fig. 10 shows that only 1.7% of the annotations have more than two overlaps with an IoU $\geq .1$, and 86.6% have 5 or more visible keypoints. Consequently, we divide the dataset into twelve benchmarks, Fig. 7, and study the performance and occurrence of errors in each sepatate one. The PR curves obtained at the evaluation threshold of .75 OKS, after sequentially correcting errors of each type are shown in Fig. 9.(a). It appears that the performance of methods listed in Tab. 1 is a result of the unbalanced data
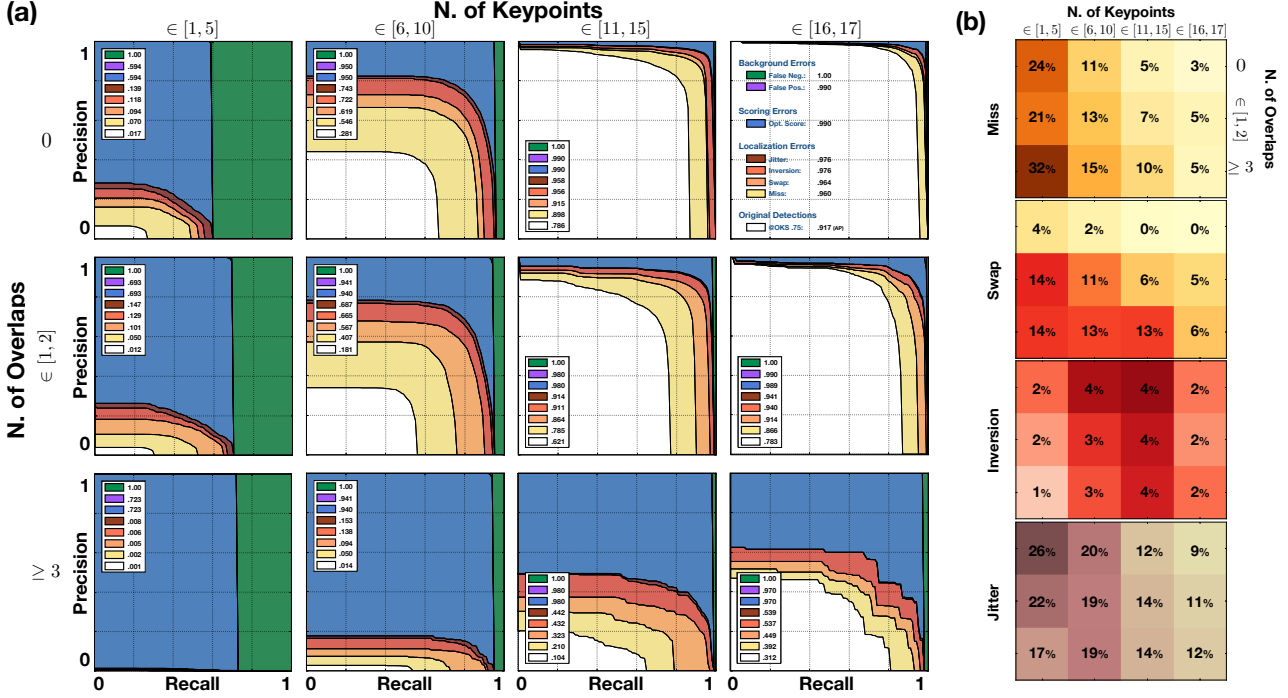
Figure 9. **Performance and Error Sensitivity to Occlusion and Crowding.** (a) The PR curves showing the performance of [11] obtained by progressively correcting errors of each type at the OKS evaluation threshold of .75 on the twelve Occlusion and Crowding Benchmarks described in Sec. 4; every legend contains the overall AP values. (b) The frequency of localization errors occurring on each benchmark set.
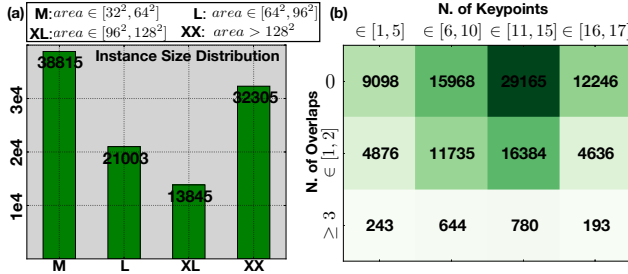

Figure 10. **Benchmarks of the *COCO* Dataset.** The number of instances in each benchmark of the *COCO* training set based on (a) the size of instances, or (b) the number of overlapping ground-truth annotations with IoU $\geq$ .1 and visible keypoints, Fig. 7.

distribution, and that current algorithms still vastly underperform humans in detecting people and computing their pose, specifically when less than 10 keypoints are visible and overlap is present. Localization errors degrade the performance across all benchmarks, but their impact alone does not explain the shortcomings of current methods. Over 30% of the annotations are missed when the number of visible keypoints is less than 5 (regardless of overlap), and background *FP* and *scoring* errors account for more than 40% of the loss in precision in the benchmarks with high overlap. In Fig. 9.(b), we illustrate the frequency of each localization error. *Miss* and *jitter* errors are predominant when there are few keypoints visible, respectively with high and low overlap. *Inversions* are mostly uncorrelated with the amount of overlap, and occur almost always in mostly vis-

ible instances. Conversely, *swap* errors depend strongly on the amount of overlap, regardless of the number of visible keypoints. Compared to the overall rates in Fig. 5.(a-cmu) we can see that *inversion* and *jitter* errors are less sensitive to instance overlap and number of keypoints. A similar analysis can be done by separating *COCO* into four size groups: *medium*, *large*, *extra-large* and *extra-extra large*, Fig. 10.(a). The performance at all OKS evaluation thresholds improves with size, but degrades when instances occupy such a large part of the image that spatial context is lost, Fig. 11.(a). AP is affected by size significantly less than by the amount of overlap and number of visible keypoints. In Fig. 11.(b) we show the AP improvement obtainable by separately correcting each error type in all benchmarks. Errors impact performance less (they occur less often) on larger instances, except for *scoring* and *FP*. Finally, while *FN*, *miss* and *jitter* errors are concentrated on medium instances, all other errors are mostly insensitive to size.

5. Discussion and Recommendations

Multi-instance pose estimation is a challenging visual task where diverse errors have complex causes. Our analysis defines three types of error - *localization*, *scoring*, *background* - and aims to discover and measure their causes, rather than averaging them into a single performance metric. Furthermore, we explore how well a given dataset may be used to probe methods' performance through its statistics of instances' visibility, crowding and size.
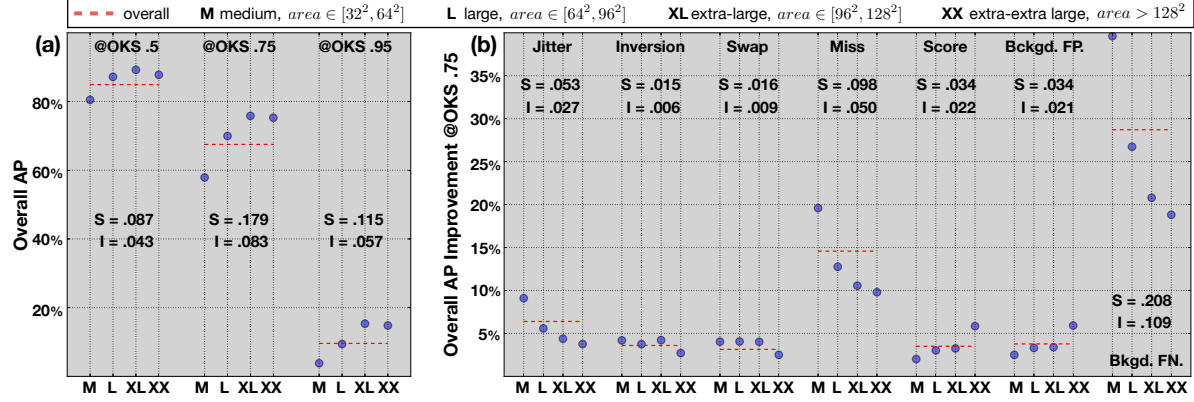
Figure 11. **Performance and Error Sensitivity to Size.** (a) The overall AP obtained by evaluating [11] at three OKS evaluation thresholds on the four Size Benchmarks described in Sec. 4. (b) The AP improvement at the OKS threshold of .75 obtained after separately correcting each error type on the benchmarks. In both figures, the dashed red line indicates evaluation over all the instance sizes, Sensitivity (S) and Impact (I) are respectively computed as the difference between the maximum and minimum, and the maximum and average, values.

The biggest problem for pose estimation is localization errors, present in about 25% of the predicted keypoints in state of the art methods, Fig. 5.(a). We identify four distinct causes of localization errors, *Miss*, *Swap*, *Inversion*, and *Jitter*, and study their occurrence in different parts of the body, Fig. 5.(b). The correction of such errors, in particular *Miss*, can bring large improvements in the instance OKS and AP, especially at higher evaluation thresholds, Fig. 5.(c-d).

Another important source of error is noise in the detection's confidence scores. To minimize errors, the scores should be (i) 'OKS monotonic increasing' and (ii) calibrated over the whole dataset, Sec. 3.2. The *optimal score* of a given detection corresponds to the maximum OKS value obtainable with any annotation. Replacing a method's scores with the optimal scores yields an average AP improvement of 5%, Fig. 6.(a), due to the fact that ground-truth instances match detections that obtain higher OKS, and the overall number of matches is increased, Tab. 2. A key property of good scoring functions is to separate as much as possible the distribution of confidence scores for detections obtaining high OKS versus low OKS, Fig. 6.(c).

Characteristics of the portrayed people, such as the amount of overlap with other instances and the number of visible keypoints, substantially affects performance. A comparison between Fig. 9.(a) and Tab. 1, shows that average performance strongly depends on the properties of the images, and that state of the art methods still vastly underperform humans when multiple people overlap and significant occlusion is present. Since *COCO* is not rich in such challenging pictures, it remains to be seen whether poor performance is due to the low number of training instances, Fig. 10.(b), and a new collection and annotation effort will be needed to investigate this question. The size of instances also affects the quality of the detections, Fig. 11.(a), but is less relevant than occlusion or crowding. This conclusion may be biased by the fact that small instances are not annotated in *COCO* and excluded from our analysis.

In this study we also observe that despite their design differences, [11, 29] display similar error patterns. Nonetheless, [11] is more sensitive to *swap* errors, as keypoint predictions from the entire image can be erroneously grouped into the same instance, while [29] is more prone to *misses*, as it only predicts keypoint locations within the detected bounding box. [29] has more than twice the number of high confidence *FP* errors, compared to [11]. Finally, we observe that *FN* are predominant around the image border for [11], where grouping keypoints into consistent instances can be harder, and concentrated in the center for [29], where there is typically clutter and bounding boxes accuracy is reduced.

Improving Localization: 3D reasoning along with the estimation of 2D body parts [40] can improve localization by both incorporating constraints on the anatomical validity of the body part predictions, and learning priors on where to expect visually occluded parts. Two promising directions for improvement are possible: (i) collecting 3D annotations [7] for the humans in *COCO* and learning to directly regress 3D pose end-to-end [30]; (ii) modeling the manifold of human poses [3, 6, 35] and learning how to jointly predict the 3D pose of a person along with its 2D skeleton [41].

Improving Scoring: Graphical models [25] can be used to learn a scoring function based on the relative position of body part locations, improving upon [11, 29] which only use the confidence of the predicted keypoints. Another promising approach is to use the validation set to learn a regressor for estimating optimal scores (Sec. 3.2) from the confidence maps of the predicted keypoints and from the sub-optimal detection scores generated by the algorithm. Comparing scores of detections in the same image relatively to each other will allow optimizing their order.

We release our code[2] for future researchers to analyze the strengths and weaknesses of their methods.

---

[2] https://goo.gl/9EyDyN

References

[1] COCO Keypoints Challenge, ECCV 2016. http://image-net.org/challenges/ilsvrc+coco2016. October, 2016. 2, 3

[2] COCO Keypoints Evaluation. http://mscoco.org/dataset/#keypoints-eval. October, 2016. 2, 4

[3] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1455, 2015. 8

[4] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2

[5] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Improving object detection with one line of code. *arXiv preprint arXiv:1704.04503*, 2017. 5

[6] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 8

[7] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1365–1372. IEEE, 2009. 8

[8] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 1

[9] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016. 1, 3

[10] X. Burgos-Artizzu, P. Dollár, D. Lin, D. Anderson, and P. Perona. Social behavior recognition in continuous videos. In *CVPR*, 2012. 1

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016. 1, 3, 4, 5, 6, 7, 8

[12] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2014. 1

[13] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1078–1085. IEEE, 2010. 2

[14] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *European Conference on Computer Vision*, pages 228–242. Springer, 2010. 1, 2, 3

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2

[16] E. Eyjolfsdottir, S. Branson, X. P. Burgos-Artizzu, E. D. Hoopfer, J. Schor, D. J. Anderson, and P. Perona. Detecting social actions of fruit flies. In *European Conference on Computer Vision*, pages 772–787. Springer, 2014. 1

[17] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3582–3589, 2014. 1, 3

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[19] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. 2

[20] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *IEEE transactions on pattern analysis and machine intelligence*, 38(4):814–830, 2016. 2

[21] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv preprint arXiv:1611.10012*, 2016. 5

[22] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016. 3

[23] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12. 1

[24] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[25] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 8

[26] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1, 2

[28] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *arXiv preprint arXiv:1603.06937*, 2016. 1, 3

[29] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. *arXiv preprint arXiv:1701.01779*, 2017. 1, 3, 4, 5, 6, 8

[30] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *arXiv preprint arXiv:1611.07828*, 2016. 8

[31] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016. 2, 3

[32] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3178–3185. IEEE, 2012. 1, 3

[33] V. Ramakrishna, D. Munoz, M. Hebert, A. J. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 2014. 1, 3

[34] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3

[35] M. R. Ronchi, J. S. Kim, and Y. Yue. A rotation invariant latent factor model for moveme discovery from static poses. *arXiv preprint arXiv:1609.07495*, 2016. 8

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016. 3

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 3

[39] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014. 3

[40] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 677–684. IEEE, 2000. 8

[41] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *arXiv preprint arXiv:1701.00295*, 2017. 8

[42] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *arXiv preprint arXiv:1602.00134*, 2016. 1, 3

[43] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001. 3

[44] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011. 1, 3