Information Sciences 537 (2020) 284–301

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Perceiving heavily occluded human poses by assigning unbiased score

Lin Zhao [a,*], Jie Xu [a], Shanshan Zhang [a], Chen Gong [a], Jian Yang [a,*], Xinbo Gao [b]

[a] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China
[b] The Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

ARTICLE INFO

Article history:
Received 25 July 2019
Received in revised form 27 March 2020
Accepted 20 May 2020
Available online 30 May 2020

Keywords:
Occlusion detection
Human pose estimation
Multi-task learning

ABSTRACT

The problem of human pose estimation has been largely solved by the prevailing Deep Convolutional Neural Networks (DCNNs). However, heavily occluded human poses still represent great challenges. In this paper, we propose a new scoring method to perceive the detection of heavily occluded poses unbiasedly. The typical way of assigning scores to detected poses is to use the mean confidence of each joint. This makes poses with occlusion suppressed durng evaluation since invisible joints may have a much lower confidence than visible joints. We address this by identifying the visibility of each joint, an occlusion aware network is designed to predict both heatmaps and visible values of joints simultaneously. Thus, the degree of occlusion of a pose can be grasped, and a much fairer score is able to be set. Furthermore, a *KS*-net is proposed to predict the *KS* (*Keypoint Similarity*) between each estimated joint location and its matched ground-truth. The predicted *KS* calibrates localization accuracy better than the maximum heat value in heatmap. Pose score is calculated using the predicted visibility and *KS* value of each joint. The efficacy of our method is demonstrated on the most widely used MS-COCO pose dataset. Extensive experiments show that using our scoring approach can significantly improve the average precision of heavily occluded poses for the provided detections.

1. Introduction

Human pose estimation from a single monocular image is a fundamental task in computer vision. Having knowledge of a person's joint locations serves as a prerequisite for many high-level applications, such as person re-identification [1,2], action recognition [3,4], and human-computer interaction [5]. Though enormous difficulties lie in accurately estimating poses from images, high achievement has been seen in recent years. Due to powerful Deep Convolutional Neural Networks and the availability of large-scale in-the-wild image datasets with proper annotations, human pose estimation is very close to a solved problem. Recent single-person pose estimation methods [6–8] all produce acceptable results with very few errors, the performance has been saturated on *Leeds Sports Pose Dataset* [9] and *MPII Human Pose Dataset* [10]. Researchers' attention now

* Corresponding authors at: PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, China (L. Zhao).

E-mail addresses: linzhao@njust.edu.cn (L. Zhao), jiexu@njust.edu.cn (J. Xu), shanshan.zhang@njust.edu.cn (S. Zhang), chen.gong@njust.edu.cn (C. Gong), csjyang@njust.edu.cn (J. Yang), gaoxb@cqupt.edu.cn, xbgao@mail.xidian.edu.cn (X. Gao).

https://doi.org/10.1016/j.ins.2020.05.083
0020-0255/

focuses on multi-person pose estimation datasets, for example MS-COCO [11], where heavy self-or-outer occlusion still presents a major obstacle to strong performance.

Like object detection, localization and scoring are two necessary steps for multi-person pose estimation. While a great deal of work [12–14] has explicitly taken aim at localizing occluded poses more precisely, few studies have made the effort to provide appropriate scores. Scores determine the order of detections both locally in an image and globally across the whole dataset, directly affecting the computation of AP (Average Precision) and AR (Average Recall). Thus, it is important to assign an optimal score to detections, which exhibits these two characteristics: (1) higher score always reflects more precise detection; (2) score calibrates the probability of being a true positive [15].

Existing methods generally compute scores by calculating the mean of maximum heat values in heatmaps. Although this is a direct and simple strategy of scoring, it lacks the ability to result in optimal scores. Specifically, averaging over heatmaps does not take the occlusion of poses into consideration. Because the number of visible joints of a pose varies dramatically from few to all, utilizing the mean may bring about large partiality in setting scores. As widely adopted by the community, invisible joints are not annotated in datasets and do not take part in the evaluation. However, during training and testing, heatmaps of all joints will be generated for the sake of convenience. As a result, invisible joints inevitably get inaccurate heatmaps, which may respond with low values in a large area of locations. For this reason, heavily occluded poses often obtain lower scores, even its visible keypoints are correctly localized. Fig. 1 shows some examples of heavily occluded poses that are correctly detected but assigned with low scores.

In this paper, we propose a new scoring method that treats heavily occluded poses without unfairness. An important insight is that the visibility of each joint must be assessed. For this purpose, we propose an occlusion aware network for human pose estimation, which conducts heatmap generation and occlusion verification simultaneously. This is done in a multi-task learning framework. Because the two tasks share most of the network parameters and DCNN features, the proposed network only adds negligible computation to judge visibility. Moreover, owing to the benefits of multi-task learning, the network is capable of producing reliable predictions of both keypoints' heatmaps and visibility.

To calculate optimal scores, a good estimation of the localization accuracy of each keypoint is another crucial factor. As a conventional technique, the maximum heat values in heatmaps are broadly used as its measurement. However, since heat value indicates the possibility of the corresponding location to be a joint, it more naturally reflects the classification confidence. Higher classification confidence does not certainly correspond to higher localization accuracy. In Fig. 2, we show some cases where the locations that get high heat values are contrarily far from their ground truth. This misalignment between classification confidence and localization accuracy also can be found in object detection, which has been proved in [16]. Hence, the maximum heat values in heatmaps may not be the best choice to compute pose scores.

Based on the above observation, we investigate the feasibility of estimating localization accuracy from heatmaps of the predicted keypoints. Inspired by the *OKS* (*Object Keypoint Similarity*) [17] similarity measure for keypoint in MS-COCO, we adopt the *keypoint similarity* (*KS*) to measure the proximity of detected keypoints to their corresponding ground-truth locations. *KS*-net is therefore introduced in this paper, which predicts the *KS* as a replacement of the maximum heat value in heatmap to work out pose scores.

Fig. 1. Heavily occluded poses with correct detections but low scores. The red points give the ground truth locations of joints, and the blue points are detected joints. As can be seen, the detections are close to the ground truth, however, the pose scores given by the mean of maximum heat values are close to zero. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

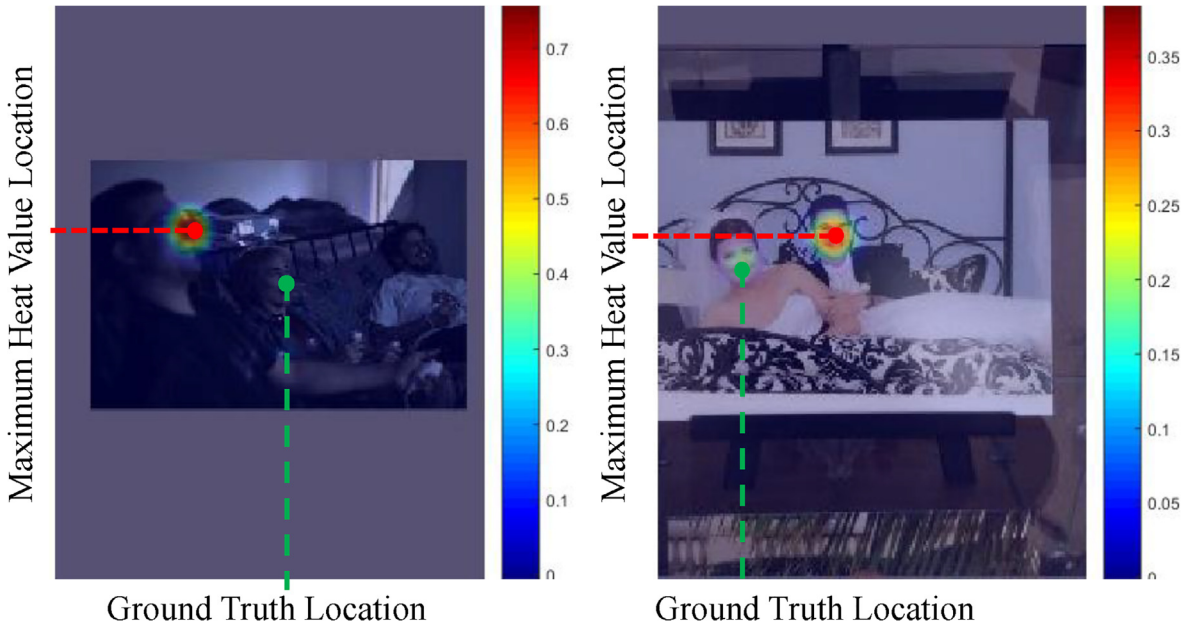


Fig. 2. Misalignment between the maximum heat value and localization accuracy. In both above figures, the ground truth locations of nose get low likelihoods, while the maximum heat values appear on the wrong persons' noses.

In summary, we develop a new scoring methodology for tackling the challenge of heavily occluded poses. The main contributions are threefold as follows:

1. We propose to assess the visibility of each keypoint while the network predicts heatmaps. This makes the process of scoring aware of occlusion, thus reducing the interference of invisible joints. AS a consequent, unbiased scores are able to be assigned to heavily occluded poses.
2. We design a novel *KS*-net for estimating the localization accuracy of predicted keypoints. As the *KS* is defined in the *OKS*, we are able to predict the *OKS* of each detected pose directly and use it as the score, which satisfies the requirements of an optimal score.
3. Based on the predicted visibility and localization accuracy of keypoints, we propose to assign reasonable scores for perceiving heavily occluded human poses.

The rest of this paper is organized as follows. In Section 2, we briefly introduce some related work of human pose estimation. In Section 3, we give an analysis of the properties of an optimal score and the drawbacks of conventional scoring methods. The details of the proposed scoring method are presented in Section 4. The experimental results are reported and discussed in Section 5. Finally, we summarize the paper and draw conclusions in Section 6.

2. Related work

Human pose estimation is one of the most attractive research topics in computer vision. Especially in recent years, the extraordinary ability of DCNNs has made it a promising task that can be used in real life. Conventional tree-structured models, *e.g.*, pictorial structures [18,19] and cascaded graphical models [20,21], provide an elegant framework to tackle the problem. Based on these methods, recent deep models push the state-of-the-art to a much higher level.

Single person pose estimation assumes that every person has been properly cropped out from images in advance. Toshev et al. firstly explore the idea of using CNN to solve the pose estimation problem. Their work DeepPose [22] obtains poses by straightforwardly regressing from images. Since this is a brute way of using CNN, Tompson et al. [23] combine CNN with graphical models, which replaces handcrafted features in conventional methods with deep features and predicts heatmaps of keypoints. Following this approach, Carreira et al. [24], Chen & Yuille [25] and many other works [26,27] make further improvements. Later works [28,29] utilize very deep CNNs to generate score maps of keypoints. Wei et al. [6] propose to refine heatmaps continuously by a multi-stage architecture. Newell et al. [7] introduce the module of Hourglass, which gets heatmaps through an encoder-decoder structure. This model represents an important milestone in the development of pose estimation. Because of its clever design and excellent performance, many following works add new techniques based on this model, *e.g.*, attention module [30], adversarial learning [31] and pyramid features [8]. As these very deep models implicitly learn pairwise relationships of joints during heatmap generation, the graphical model is no longer needed.

The problem of single person pose estimation has been largely solved by the above mentioned very deep models. Their performance on large-scale datasets like *MPII-Human Pose Dataset* [10] has reached a very high standard, and the discrepancy in detection accuracy between them is minor. Hence, more challenging multi-person pose estimation is attracting more researchers' attention.

Multi-Person Pose Estimation deals with images containing multiple instances of people, and the number of visible keypoints varies among different persons. Bottom-up methods detect all keypoints at first and make efforts to solve the problem of assembling. Most of the early methods [32,33] belong to this category. Deepcut [32] makes use of Integer Linear Program (ILP) to group joints detection candidates into separate persons. Deepercut [34] utilizes deeper networks to predict keypoints and improves the solution of ILP. However, solving the ILP problem globally is an NP-hard problem, it takes several minutes to get person clusters on one image. Then Cao et al. [35] introduce the estimation of Part Affinity Fields (PAFs) between each pair of body parts. With the help of PAFs, they are able to relax the global inference problem of assembling into a P-hard problem. They show remarkable results by impressive demos in real-time. Newell et al. [36] present competitive results by predicting associative embedding to group keypoints.

Top-down approaches utilize two stages to tackle the problem. Firstly, persons are detected and cropped by a top-performing object detector, and then single person pose estimation methods are employed for each cropped person patch. Papandreou et al. [12] use the Faster RCNN [37] for person detection and a fully convolutional ResNet [38] for pose estimation. Fang et al. [39] suggest to refine bounding boxes before conducting pose estimation. Mask-RCNN [40] appends a keypoint localization head on top of the ROI pooling layer, thus achieves person detection and pose estimation in one network. Recently, Chen et al. [13] propose to refine the localization of hard keypoints based on a Feature Pyramid Network (FPN) [41]. With the help of state-of-the-art object detectors, they further improve the results. Xiao et al. [14] also design the pose estimation network based on FPN, but replace upsampling with deconvolution. These methods produce better results than bottom-up ones as they can zoom into each person patch, but the speed is much slower.

Because of the superior performance of Top-down methods, they dominate the leading places of MS-COCO keypoints competition [42]. Their capability is amazing considering the challenges of the task. However, occlusion and invisible keypoints are still tough issues that need to be overcome. This paper follows the pipeline of Top-down methods and tries to perceive heavily occluded human poses accurately by assigning more reasonable scores.

3. Delving into the conventional scoring method

In this section, we will first analyse the desired properties of an optimal score and its formulation. Then, two drawbacks of current scoring methods will be explored, which are the ignorance of pose occlusion and the misalignment between the maximum heat value in heatmaps and localization accuracy. All analysis is done on the dataset of MS-COCO [11].

3.1. Properties of an optimal score

In multi-person pose estimation, a score must be given to each detected pose for evaluation. Scores represent the confidence of detections and are able to affect the result of evaluation largely. Locally, detections within each image are arranged in descending order of scores and assigned to ground-truths in sequence. Hence, while the matching process goes on, lower scored detections will have a smaller pool of available annotations to get matched with. Globally, after matches are determined, detections across all the images are sorted according to scores, and AP and AR are computed according to this ranking by using predefined criteria.

Therefore, we want scores to have these two properties: (i) monotonicity, a higher score always belongs to a better detection; (ii) calibration, a score represents the possibility of its corresponding detection to be a true positive. The first property guarantees that a better detection consistently ranks above poorer ones, so that the highest performance can be achieved for the provided detections. The second property allows to judge the quality of a detection directly from its score. Ideally, as has been presented in [15], the measurement for keypoint similarity (*i.e.*, *OKS*) is excellent for scoring, since the *OKS* perfectly meets both monotonicity and calibration. The formulation of the *OKS* is given as follows:

$$OKS = \frac{\sum_i e^{-\frac{d_i^2}{2s^2k_i^2}}\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{1}$$

where d_i is the Euclidean distance between each detected joint and its corresponding ground truth, v_i is the visibility flag of the ground truth, s is the object scale, and k_i is a per-keypoint constant.

As can be seen from Eq. (1), the *OKS* is calculated by averaging over all the visible keypoints in the ground truth of a pose ($v_i > 0$), and the un-normalized Gaussian function in the numerator can be regarded as *Keypoint Similarity* (*KS*). Given a pose detection $\hat{\boldsymbol{\theta}}^{(p)}$ and the corresponding annotation $\boldsymbol{\theta}^{(p)}$ of a person p, the i_{th} keypoint's $KS_i^{(p)}$ can be written as:

$$KS_i^{(p)} = e^{-\frac{\left\|\hat{\theta}_i^{(p)} - \theta_i^{(p)}\right\|_2^2}{2s^2k_i^2}}. \tag{2}$$

It is clear that $KS_i^{(p)}$ represents the localization accuracy of the i_{th} keypoint, and its value is affected by the Gaussian standard deviation k_i as well as the scale of the instance s. k_i is specific to the type of a keypoint, that is to say, in the case of the same value of numerator, keypoints on a person's body (*e.g.*, hips) tend to have a *KS* much larger than on a person's head (*e.g.*, eyes). We give an illustration of *KS* in Fig. 3.

Though the *OKS* provides the best solution of scoring, it is impossible to put to use in testing for the annotations are not given. Because a keypoint is located by the pixel with the maximum likelihood in its heatmap, current methods [13,14] typically utilize the mean of maximum heat values in heatmaps as the score. This is a convenient way, however, it brings drawbacks.

3.2. Ignorance of pose occlusion

Whether a pose is occluded or not, current methods for human pose estimation predict the locations of all keypoints, but ignore the prediction of visibility. Because invisible keypoints are not taken into account during evaluation, it seems that the prediction of visibility is not necessary. This may be true in single person pose estimation, but it is not appropriate for multi-person pose estimation. As analysed in Section 3.1, score affects the result of evaluation, and the visibility of keypoints is important for obtaining reasonable scores.

We first give an analysis about the conventional scoring method. Let $h_i^{(p)}$ denote the maximum likelihood in the i_{th} keypoint's heatmap of a person p and S_{MH}^p represent the score computed by taking the mean of maximum heat values in all heatmaps. Then the formulation can be written as:

$$S_{MH}^{(p)} = \frac{\sum_i h_i^{(p)} \cdot 1}{\sum_i 1}. \tag{3}$$

Comparing Eq. (3) with Eqs. (1) and (2), $h_i^{(p)}$ can be considered as an estimation of the localization accuracy (*i.e.*, $KS_i^{(p)}$), and S_{MH}^p can be regarded as a prediction of the *OKS*. Hence, it is obvious that $\delta(v_i > 0) \equiv 1$ in Eq. (3), *i.e.*, all keypoints are treated as visible.

Because occlusion is ubiquitous in multi-person pose estimation, neglecting keypoints' visibility may cause substantial scoring inequality between poses. As a widely used technique, the ground truth of invisible keypoints is given with a heatmap of zeros during training, thus the heat values in their predicted heatmaps are commonly very low. For this reason, the scores computed by Eq. (3) will heavily favour poses with more visible keypoints. Fig. 4 gives an illustration of the scoring bias.

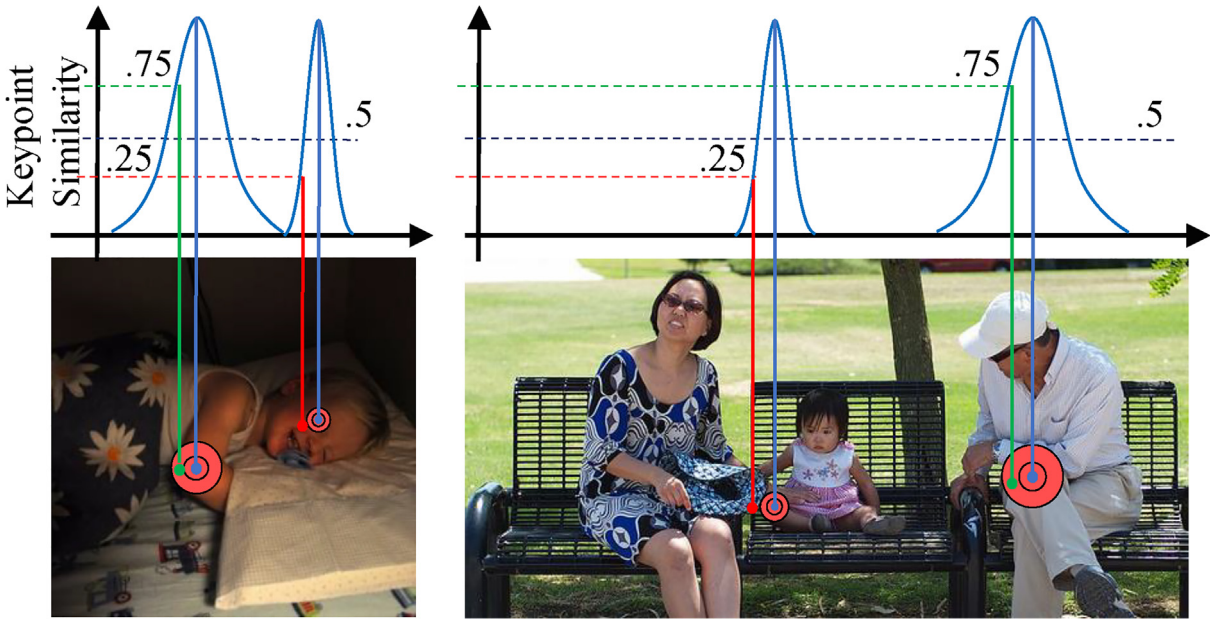


Fig. 3. Illustration of Keypoint Similarity (*KS*). The left figure shows the *KS* of two detections of different joints, eye (red) and elbow (green), on the same person. The right figure shows the *KS* of two detections of the same joint knee on the different persons, baby (red) and man (green). The red concentric circles represent *KS* values of .5 and .85, and their sizes vary by different keypoint types and different sizes of person instances. Hence, detections at the same distances from the corresponding ground truth can have different *KS* values (.25 *versus* .75 in the figure). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

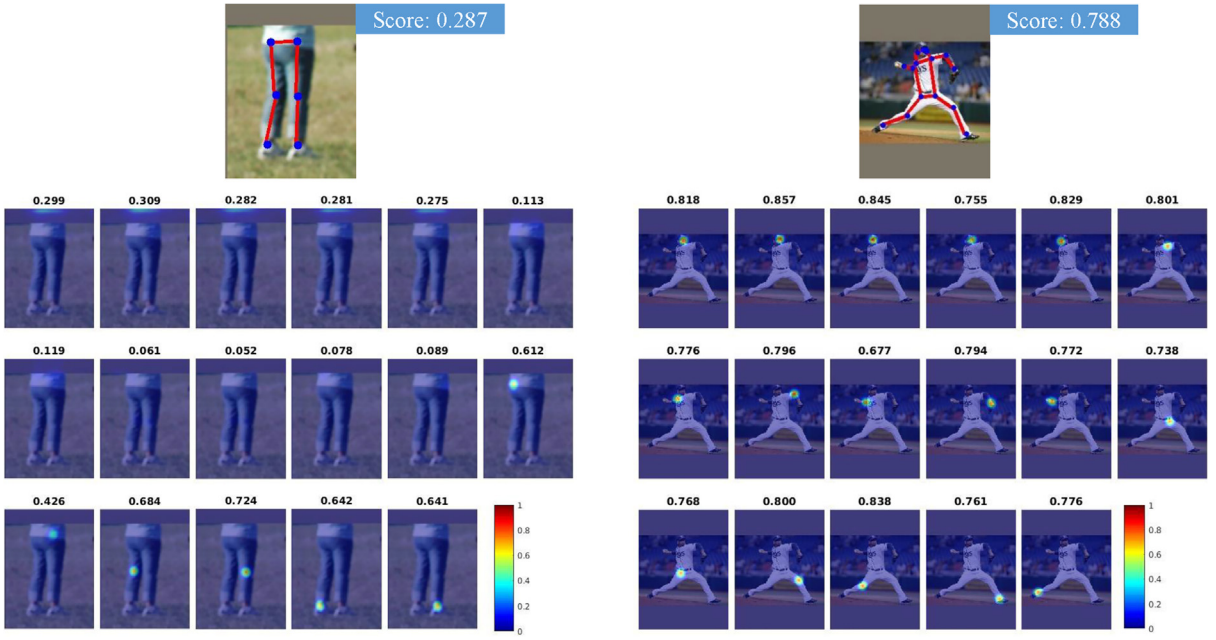


Fig. 4. Scoring bias. The two pose detections can both be considered as good. However, the right one with more visible keypoints gets a much higher score. The sub-figure beneath the detected pose presents the heatmap of each joint (from eye to ankle) and the corresponding maximum likelihood. It is clear, invisible keypoints produce much lower heat values than visible keypoints, which inevitably causes lower pose scores calculated by Eq. (3).

3.3. Misalignment of classification and localization accuracy

Detection based methods are the mainstream of human pose estimation algorithms, which depict joint locations with heatmaps. Denote the ground truth location of the i_{th} body joint of a person p by $\boldsymbol{\theta}_i^{(p)} = \left(x_i^{(p)}, y_i^{(p)}\right)$, the ground truth heatmap $\mathbf{H}_i^{(p)}$ is generated by computing a 2D Gaussian blob with the center on the ground truth location.

$$\mathbf{H}_i^{(p)}(\mathbf{o}) \sim \mathcal{N}\left(\theta_i^{(p)}, \boldsymbol{\Sigma}\right), \tag{4}$$

where $\mathbf{o} \in R^2$ denotes a location in heatmap $\mathbf{H}_i^{(p)}$, and $\boldsymbol{\Sigma}$ is the variance which is empirically set as an identity matrix $\mathbf{I}$.

Given a predicted heatmap $\hat{\mathbf{H}}_i^{(p)}$ for the i_{th} keypoint, the location $\mathbf{o}$ with the maximum heat value is regarded as the final joint position $\hat{\theta}_i^{(p)}$, and the maximum heat value $h_i^{(p)}$ is considered as the localization accuracy of the joint.

$$\hat{\theta}_i^{(p)} = \arg\max_{\mathbf{o}} \hat{\mathbf{H}}_i^{(p)}(\mathbf{o}), \tag{5}$$

$$h_i^{(p)} = \max_{\mathbf{o}} \hat{\mathbf{H}}_i^{(p)}(\mathbf{o}). \tag{6}$$

This solution is widely-adopted in detection based human pose estimation methods. Since the heat value of each location in the heatmap represents the probability of the location being the joint, it is convenient to use such a post-processing technique. However, it is problematic to utilize the maximum likelihood $h_i^{(p)}$ for computing the pose score, as formulated in Eq. (3).

We visualize the distribution of the maximum likelihoods of several detected joints in Fig. 5. The x-axis is the ground truth localization accuracy (*i.e.*, $KS_i^{(p)}$) of detected joints, while the y-axis denotes the corresponding maximum likelihoods in heatmaps (*i.e.*, $h_i^{(p)}$). As can be seen, $h_i^{(p)}$ is not well correlated with $KS_i^{(p)}$, the Pearson correlation coefficients between them are all very low, which indicates the maximum likelihood is not a good estimation of the localization accuracy.

This misalignment is caused by confusing classification accuracy with localization accuracy, as has been thoroughly analysed in [16]. The maximum likelihood indicates the classification accuracy of its location to be the joint, larger likelihoods do not surely correspond to better localization accuracy. Fig. 2 exhibits some examples of the misalignment.

To further demonstrate the misalignment, we substitute $KS_i^{(p)}$ in Eq. (1) with $h_i^{(p)}$ to predict *OKS*.

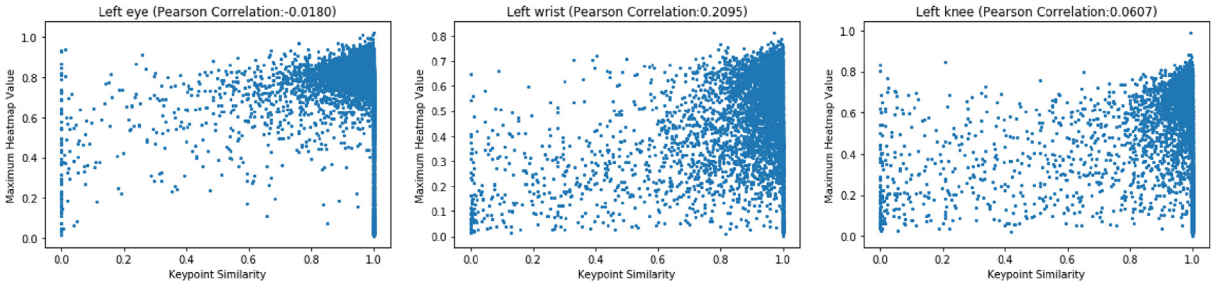


Fig. 5. The correlation between the maximum heat value and the localization accuracy. The correlation of left eye, left wrist, and left knee are presented. The Pearson correlation coefficients are: −0.0180 (left eye), 0.2095 (left wrist), and 0.0607 (left knee) respectively.

$$\widetilde{OKS}^{(p)} = \frac{\sum_i h_i^{(p)} \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}. \tag{7}$$

Fig. 6 shows the distribution of $\widetilde{OKS}$ and the ground truth *OKS* separately. As can be seen, the two distributions are quite different, which also proves $h_i^{(p)}$ misaligns $KS_i^{(p)}$. Comparing the formulation of $h_i^{(p)}$ in Eq. (6) with $KS_i^{(p)}$ in Eq. (2), it is obvious that $h_i^{(p)}$ does not take the type of a keypoint (*i.e.*, k_i) and the scale of the instance (*i.e.*, s) into account. Thus, $h_i^{(p)}$ lacks the ability to make a good estimation of $KS_i^{(p)}$, which brings errors in computing the score of pose.

4. Assigning an unbiased score

To assign as optimal scores as possible to detected poses, we present the details of our proposed scoring method in this section. In Section 4.1, we show the network design of occlusion aware human pose estimation. Then, the *KS*-net for predicting the localization accuracy of each detected joint is proposed in Section 4.2. Finally, we summarize the proposed scoring algorithm in Section 4.3.

4.1. Occlusion aware human pose estimation

Fig. 7 shows the architecture of our model, which consists of a shared backbone for feature extraction, a regression subnet for occlusion judgement and a detection subnet for keypoint estimation, we refer to this network as OA-HPE net. We utilize the Feature Pyramid Network (FPN) [41] as the backbone. Stacked hourglass [7] is another popular network for pose estimation, however, FPN based methods [13,14] exhibit the best performance on the MS-COCO keypoint leader board [42]. Because we aim to better perceive heavily occluded poses in multi-person pose estimation, for the convenience of fully taking advantage of the merits of state-of-the-art methods, FPN is employed in our method. Based on the visual features given by FPN, the subnet for keypoint detection can just apply the successful designs of popular methods. We do not make specific efforts to refine this part, because our focus is on improving the score of detected poses.

The subnet for occlusion judgement takes the visual features from the last residual block of FPN, which is typically denoted as C_5. A 1×1 convolutional kernel is applied on C_5 to generate the features for visibility regression. Then, two hidden 1024-d fully connected (*fc*) layers are attached before the final visibility regression layer. The output $\hat{\mathbf{V}} \in R^K$ is a vector

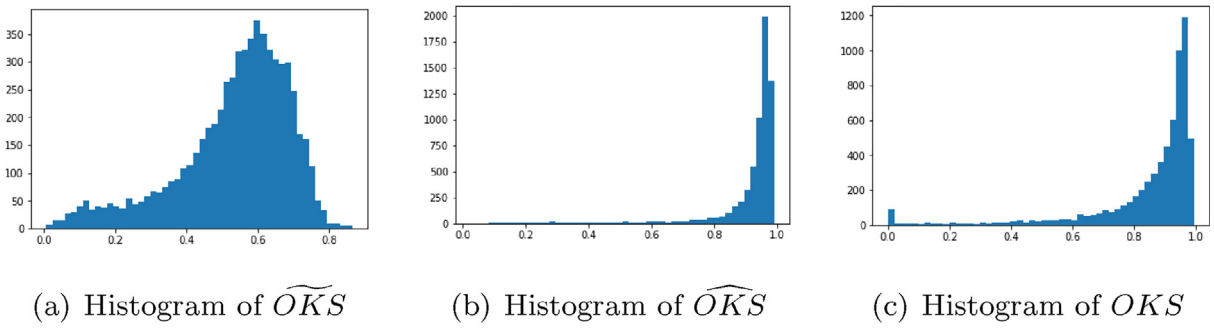


Fig. 6. Histogram of $\widetilde{OKS}$, *OKS*, and $\widehat{OKS}$. (a) The distribution of $\widetilde{OKS}$ is very different from *OKS*, which indicates the maximum heat value can not be interpreted as the localization accuracy. (b) To resolve the issue, we propose to predict the localization accuracy (*i.e.*, *KS*) for each detected keypoint.

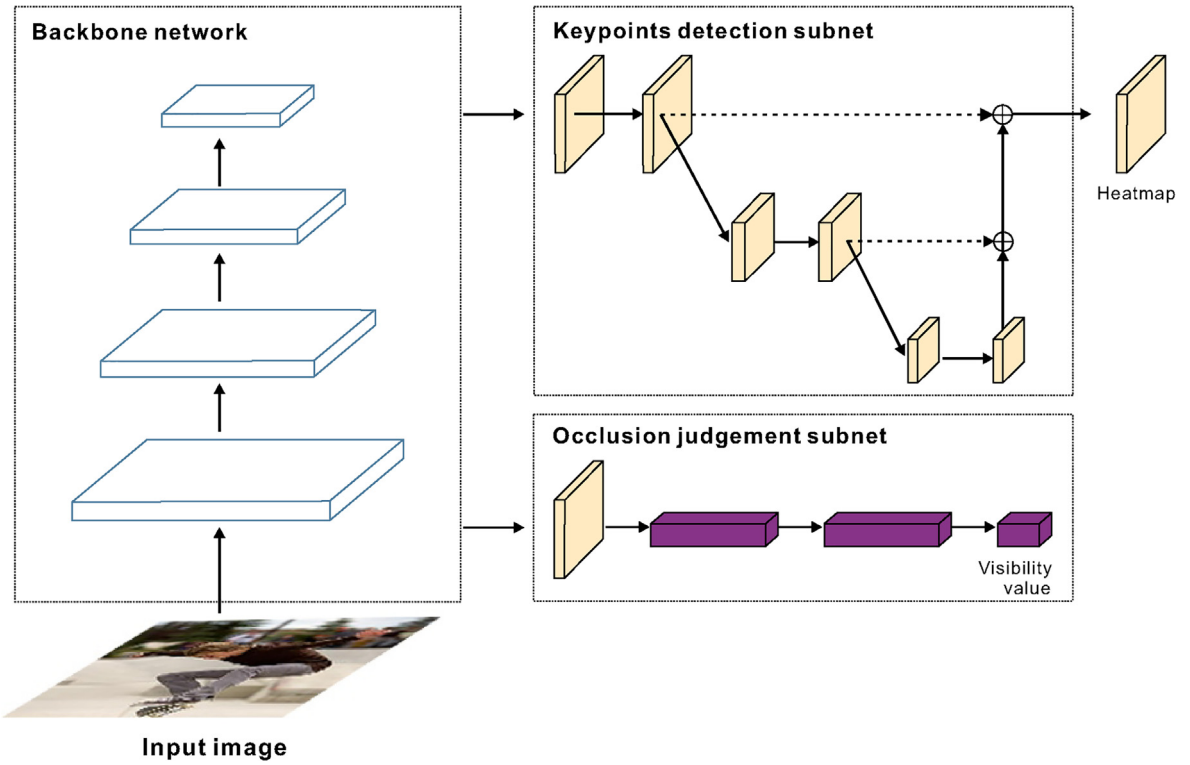


Fig. 7. Network architecture of the proposed OA-HPE net.

that estimates the occlusion of all K joints in a pose, where each element $0 < \hat{v}_i < 1$ represents the visibility value of the i_{th} joint.

The loss for training pose estimator is the squared error between predicted and ground truth heatmaps. For training visibility regressor, all keypoints annotated with coordinates are regarded as visible (*i.e.*, $v_i = 1$), and only keypoints without labels are treated as invisible (*i.e.*, $v_i = 0$). We use sigmoid cross-entropy to compute the visibility regression loss, as the visibility of each keypoint is independent and not mutually exclusive. Given the ground truth $\mathbf{V}$ and prediction $\hat{\mathbf{V}}$, the loss function is formulated as:

$$\mathcal{L}_{visible} = -\sum_{i=1}^{K}(v_i log(\hat{v}_i) + (1 - v_i) log(1 - \hat{v}_i)). \tag{8}$$

The total loss for training is the sum of pose estimation loss and visibility regression loss. Because the subnet for visibility regression is quite light, the added computation comparing to only training for pose estimation is almost negligible. And the training can be done just following the standard pipeline of detection based pose estimation methods [13,14].

4.2. KS-net for localization accuracy prediction

KS-net predicts the localization accuracy of keypoints from the corresponding heatmaps. Comparing to simply taking the maximum heat value, *KS*-net utilizes the information from all heatmaps of a pose and learns to estimate the most probable localization accuracy for all keypoints. We show the design of *KS*-net in Fig. 8. Because heatmaps can be regarded as high-level features, there is no need to conduct very deep convolutions again to learn powerful features for prediction. Here, a ResNet-18 [38] network is utilized as the backbone, then two hidden 1024-d fully connected (*fc*) layers are attached before the final localization accuracy regression layer.

The output of *KS*-net is a vector $\widehat{KS} \in R^K$. The ground truth localization accuracy *KS* can be computed according to Eqs. (2) and (5). We use L1-norm to compute the loss for training,

$$\mathcal{L} = \sum_{i=1}^{K}\left|\widehat{KS}_i - KS_i\right|. \tag{9}$$

As *KS*-net is a relatively small network, it can be trained very efficiently from scratch.

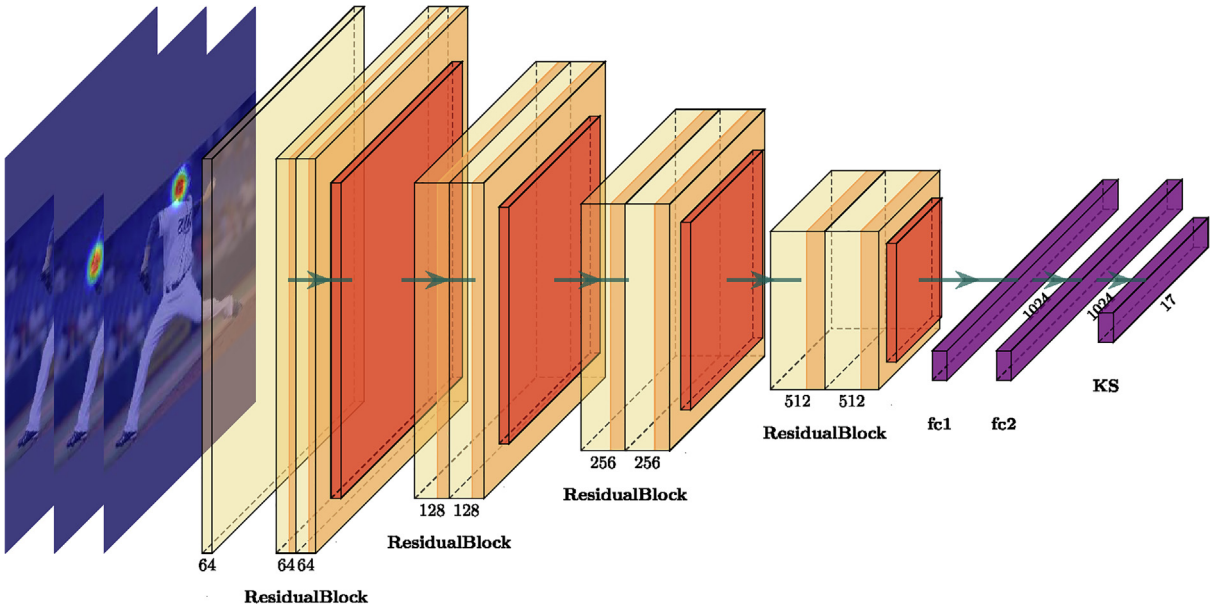


Fig. 8. Illustration the architecture of the proposed *KS*-net. It takes heatmaps as input and estimates the *KS* of each keypoint through a RestNet-18 backbone and two fully connected layers.

4.3. Summary of the proposed scoring method

After training the OA-HPE net and *KS*-net, unbiased scores are able to be assigned to all detected poses. We present a pseudo-code for the proposed scoring method in Algorithm 1.

Algorithm 1 Assigning unbiased scores for detected poses.

Input: The test image $\mathbf{I}$ containing one person p; The OA-HPE net $\mathcal{N}$; The *KS*-net $\mathcal{M}$.

Output: Detected human pose $\theta^{(p)}$; Score for the detected pose $s^{(p)}$.

Get the heatmaps and visibility values of keypoints:

$\hat{\mathbf{H}}^{(p)}, \hat{\mathbf{V}}^{(p)} \leftarrow \mathcal{N}(\mathbf{I})$;

Obtain the coordinates of each keypoint:

for $i \in [1, \cdots, K]$

$\hat{\theta}_i^{(p)} \leftarrow \arg\max \hat{\mathbf{H}}_i^{(p)}$

end for

Predict the localization accuracy of each keypoint:

$\widehat{KS}^{(p)} \leftarrow \mathcal{M}\left(\hat{\mathbf{H}}^{(p)}\right)$

Calculate the score for the detected pose:

$s^{(p)} \leftarrow \frac{\sum_{i=1}^{K} \widehat{KS}_i^{(p)} \hat{v}_i^{(p)}}{\sum_{i=1}^{K} \hat{v}_i^{(p)}}$

return $\theta^{(p)}, s^{(p)}$

5. Experiments

We conduct the experiments following the pipeline of top-down approaches for multi-person human pose estimation. However, our focus is on the validation of the proposed scoring method, especially on its performance of improving the perception of heavily occluded human poses for the provided pose detections.

5.1. Experimental setup

Dataset. Our experiments are conducted on MS COCO [11]. MPII Human Pose [10] is also a widely used multi-person human pose estimation dataset. But its evaluation metric mimics single person pose estimation, which does not evaluate in OKS-based AP and AR. Because we intend to improve the detection AP and AR of heavily occluded human poses, exper-

iments are not done on MPII Human Pose [10]. Both of the OA-HPE net in Section 4.1 and *KS*-net in Section 4.2 are trained on MS-COCO 2017 training dataset (includes 57 K images and 150 K person instances). Testing is done on MS-COCO 2017 validation dataset (includes 5 K images). Because we need the ground truth annotations to analyse the performance of our scoring method on heavily occluded human poses, the testing dataset of MS-COCO 2017 is not used in our experiments.

Evaluation metric. We follow the standard OKS-based AP and AR on MS COCO [17] to evaluate the performance of our scoring method. Further, to analyse the improvement on occluded human poses, the analysis tool released by Ronchi et al. [15] is utilized. They divide the COCO Dataset into twelve benchmarks according to occlusion (the number of visible keypoints) and crowding (the amount of overlap), and the performance is studied in each separate one by obtaining the PR (Precision Recall) curves at the evaluation threshold of .75 OKS. Based on their tool, we divide the MS-COCO 2017 validation dataset into four benchmarks only according to occlusion and also present the PR curves for analysis.

Implementation details. The OA-HPE net is trained by following the standard procedures as in [13]. We use random rotation ($[-45°, 45°]$), random scale ($[0.65, 1.35]$) and flipping to do data augmentation. The model is trained using the Adam optimizer, and the base learning rate is set as $1e-3$, which is dropped to $1e-4$ and $1e-5$ at the 120_{th} and 150_{th} epoch respectively. We train the model for 160 epochs with the batch size 32, and batch normalization is used. After training, the OA-HPE net is tested on the MS-COCO 2017 training dataset using the ground truth bounding boxes of all person instances. Then the estimated heatmaps and locations of all keypoints of each person instance are saved for the training of *KS*-net.

For each person instance, we first calculate the ground truth *KS* of each annotated keypoint according to Eq. (2). Using the *KS* of keypoints of a person instance as the ground truth and the corresponding heatmaps as the input, then the *KS*-net is able to be trained. We also train the model using the Adam optimizer, the base learning rate is set as $1e-3$ and dropped to $1e-4$ and $1e-5$ at the 30_{th} and 50_{th} epoch. The model is trained for 60 epochs using batch normalization. Because the *KS*-net is quite slim, the training can be done in hours following the standard procedures as in [38].

During testing, the heatmap from the OA-HPE net are computed by averaging the heatmaps of the original and flipped images. The final location of each keypoint is obtained by adjusting the highest heat value location with a quarter offset in the direction from the highest response to the second highest response. Then the score of a detected pose is obtained using the proposed method as described in Algorithm 1. We conduct all the experiments on a single Tesla P40 GPU.

5.2. Ablation experiment

In this subsection, we investigate the effectiveness of each component. To make the improvement produced by the proposed scoring method persuasive, all experiments are done using ResNet-101 as the backbone of FPN [41] and 384×288 cropped images as the input, which gives the best AP as reported in [13].

5.2.1. OA-HPE net

Here we discuss the performance of the OA-HPE net, especially how well it predicts the occlusion of each keypoint and how it may change the final AP.

Keypoints detection: affected or not? We first investigate whether doing occlusion judgement may affect the detection of keypoints. Table 1 gives the results on MS-COCO 2017 validation dataset. The first two rows show the results of HPE net, which has the same backbone and keypoints detection subnet with OA-HPE net but is absent of the occlusion judgement subnet. Both networks are trained with the same strategies, and the best results are reported. To do the comparison comprehensively, we present the results of both using the ground truth human boxes and the boxes given by a person detector [13]. The APs in this table are all calculated using the conventional scoring method as in [13,14] since we only analyse the performance of keypoints detection here. As can be seen, OA-HPE net almost gives the same results with HPE net, the little difference is acceptable considering that OA-HPE net also takes the task of occlusion judgement.

Occlusion judgement: viable or not? Because awareness of the visibility of keypoints is the basis of assigning unbiased scores, the accuracy of occlusion judgement can largely sway the performance of the proposed scoring method. Using the ground truth human boxes, Table 2 presents the accuracy of occlusion judgement on each keypoint by the OA-HPE net. It is clear that the predicted visibility of all keypoints is quite correct, most of the accuracy is above 90%. The occlusion of ears and wrists is relatively a little harder to judge properly, the reasons may be that they are at the end of a limb and the size is commonly small. However, even these keypoints still get an accuracy of about 88%.

Table 1
Comparison of the APs on MS-COCO 2017 validation dataset. HPE: The network only does human pose estimation; OA-HPE: The network does keypoints detection and occlusion judgement simultaneously. G: The input image of a person instance is cropped using the ground truth human box; D: The bounding boxes of human instances are obtained by a person detector [13].

Models	AP	$AP_{@.5}$	$AP_{@.75}$	AP_m	AP_l	AR	AR_m	AR_l
HPE(G)	74.8	92.4	81.4	71.8	79.9	77.9	74.4	83.2
HPE(D)	72.9	89.2	79.4	69.1	79.9	79.2	74.3	86.0
OA-HPE(G)	74.6	92.4	81.5	71.4	79.6	77.4	73.9	82.7
OA-HPE(D)	72.6	88.8	79.2	68.5	79.7	78.6	73.5	85.7

Table 2
Accuracy of the occlusion judgement on each keypoint by OA-HPE net. The labels of keypoints follow the annotation of MS-COCO dataset. L means left, R means right. For the sake of saving space, we only use the first three letters of each keypoint, *e.g.*, Sho means shoulder and Elb means elbow.

Nose	L-Eye	R-Eye	L-Ear	R-Ear	L-Sho	R-Sho	L-Elb	R-Elb
93.2	93.0	92.7	88.5	88.4	94.4	94.8	89.2	90.0
L-Wri	R-Wri	L-Hip	R-Hip	L-Kne	R-Kne	L-Ank	R-Ank	
87.6	88.5	90.3	90.6	92.2	92.3	93.4	93.0	

Qualitative comparison. To understand the benefit of occlusion judgement better, we compare the pose estimation results with and without the knowledge of keypoints' visibility. Fig. 9 shows some qualitative comparison. The first row exhibits the poses estimated by HPE net. Because HPE net lacks the ability to predict keypoints' visibility like the existing models [12,40], the detection results of all keypoints are delivered whether occlusion or not. The second row gives the results obtained by OA-HPE net, which only displays the keypoints that are predicted as visible. It is clear that the poses given by OA-HPE net are more attractive. For the occluded poses, the results of the HPE net may be weird and disobey the cognition of human poses, as the detected locations of invisible keypoints can be arbitrary and unreliable.

We also present some cases that the OA-HPE net gives wrong occlusion judgements in Fig. 10. In the first image, all keypoints are annotated as visible, the OA-HPE offers wrong assessments on four keypoints. However, the left wrist and elbow are actually unseen and the annotations are wrong. The visibility of the left and right ankles are ambiguous, the OA-HPE net estimates their visibility value (*i.e.*, $\hat{v}_i$) at 0.44, which also seems quite reasonable ($\hat{v}_i > 0.5$ is regarded as visible). The pose in the second image is a rare case, only two knees are visible. The wrong judgement made by OA-HPE net may be caused by little similar training images except for ambiguous keypoints.

From both the quantitative and qualitative results, the OA-HPE net demonstrates its ability to provide credible predictions about the visibility of keypoints. Comparing to the existing human pose estimation models that only offer keypoints' locations, this enriches the information of a detected human pose and may facilitate more applications of human poses in the real world.

What is the difference made on APs of knowing occlusion? With the knowledge of keypoints' visibility, we are able to assign unbiased scores for all the provided pose detections. Here, we give a detailed analysis of the improvement that can be delivered on the APs using unbiased scores. To exclude the effect of a person detector, we get pose detections from the OA-HPE net using the ground truth human boxes. According to the analysis in Section 3.1, the OKS (*i.e.*, Eq. (1)) gives the formu-

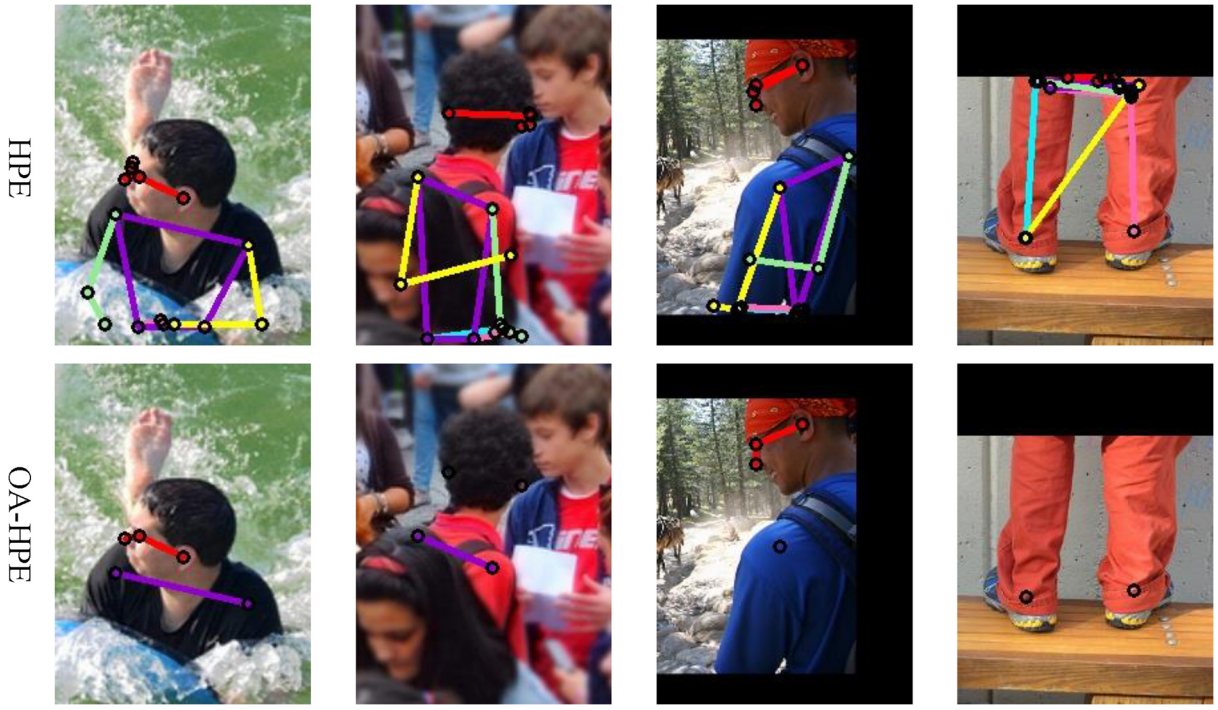


Fig. 9. Comparison of predicted poses by the HPE net and OA-HPE net. The OA-HPE net predicts both the location and visibility of each keypoint, and HPE net only gives the location of all keypoints ignoring the visibility. In the figure, each keypoint is presented using a black circle, limbs are depicted with lines in different colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

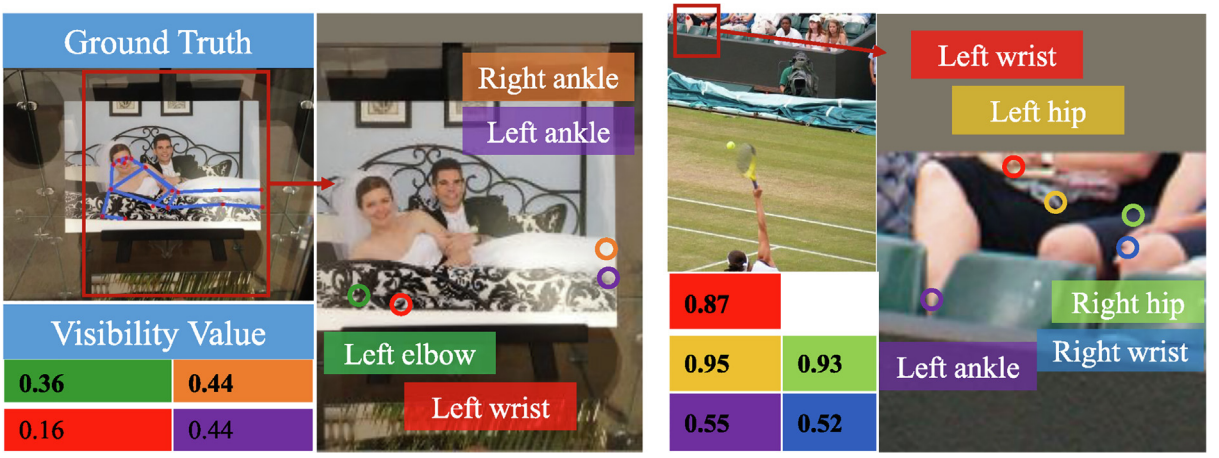


Fig. 10. Failure cases of occlusion judgement by OA-HPE net. The left part of each sub-figure gives the ground truth pose and the predicted visibility value by OA-HPE net; the right part shows the joints whose visibility is wrongly predicted by OA-HPE net. The wrongly predicted joint and its corresponding visibility value are presented in the same color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lation of an optimal score. Thus, we calculate the ground truth *KS* according to Eq. (2) and make analysis specifically on the role of keypoints' visibility.

The APs obtained using different settings of the visibility value are given in Table 3. It is clear that ignoring occlusion will get the worst AP, which can be more than 2% lower than the upper bound even for the same provided pose detections. This suggests that quite a lot good pose detections are wrongly treated as undependable ones. Hence, it is obvious that heavily occluded poses can be the most possible ones that be mistakenly assigned low scores. For the visibility predicted by the OA-HPE net, the AP is improved by about 1% than ignoring occlusion. This indicates that occlusion judgement in the OA-HPE net plays its role, which helps set fair scores across poses and makes good detections of occluded poses perceivable.

5.2.2. KS-net

The accuracy of the predicted $\widehat{KS}$. Here, we analyse the localization accuracy predicted by the *KS*-net. For the heatmaps given by the OA-HPE net using ground truth human boxes, we get the prediction $\widehat{KS}$ for each keypoint. The accuracy of $\widehat{KS}$ is investigated based on the pose score that is calculated with it. Because the OKS (*i.e.*, Eq. (1)) gives the optimal score, we utilize the predicted $\widehat{KS}$ to compute $\widehat{OKS}$. To investigate only the reliability of $\widehat{KS}$, the ground truth occlusion of keypoints is used. Fig. 6 presents the distributions of $\widehat{OKS}$ and *OKS*. To demonstrate the advance of the *KS*-net, we also do the comparison with the distribution of $\widetilde{OKS}$, which is calculated using the maximum likelihood in heatmap as the prediction of *KS* (Eq. (7)). As can be seen, the distribution of $\widehat{OKS}$ is very similar to that of *OKS*, which shows pose detections getting high scores are in the majority. Nonetheless, the distribution of $\widetilde{OKS}$ indicates that the majority of poses have medium scores. Hence, it is clear that *KS*-net can make better predictions about the localization accuracy of keypoints than the existing methods [13,14].

We further manifest the effectiveness of the *KS*-net through the result of APs. Table 4 shows the APs acquired by using $\widetilde{OKS}$, $\widehat{OKS}$ and *OKS* as the pose score separately. For a fair comparison, we use the paired t-test with the significance level 0.05 [43] to statistically verify the superiority of the proposed *KS*-net. The proposed *KS*-net is trained for ten times, and the best model obtained in each training is used for testing. Obviously, the correctness of the prediction of *KS* holds sway over the final APs. Using $\widetilde{OKS}$ as the score causes a nearly 3% drop of AP comparing to the optimal score *OKS*, which implies that the maximum heat values of heatmaps make unsatisfied predictions of the localization accuracy. $\widehat{OKS}$ is able to improve the AP by about 1% by using the predictions given by the *KS*-net. This demonstrates that *KS*-net is capable of making better

Table 3
APs obtained using scores calculated on different settings of the visibility value for the same provided pose detections by OA-HPE net. In the first row, Visible indicates that all keypoints' visibility value is set as $\hat{v}_i = 1$, ignoring occlusion. In the second row, the predicted visibility $\hat{v}_i$ by OA-HPE net is used. GT means using the ground truth visibility, which calculates the optimal scores and produces the upper bound APs for the provided detections.

Visibility	AP	$AP_{@.5}$	$AP_{@.75}$	AP_m	AP_l	AR	AR_m	AR_l
Visible ($\hat{v}_i = 1$)	74.7	92.9	81.9	71.5	79.6	77.4	73.9	82.8
OA-HPE ($\hat{v}_i$)	75.6	93.0	82.4	72.0	81.0	77.4	73.9	82.8
GT ($\hat{v}_i = v_i$)	77.2	93.1	83.2	73.6	82.6	77.4	73.9	82.8

Table 4
APs obtained using $\widetilde{OKS}$, $\widehat{OKS}$ and *OKS* as the pose score separately for the same provided pose detections by OA-HPE net. $\surd$ means that $\widehat{OKS}$ is significantly better than the corresponding method.

Score	AP	$AP_{@.5}$	$AP_{@.75}$	AP_m	AP_l	AR
$\widetilde{OKS}$	74.3 $\surd$	92.5	81.3 $\surd$	70.8 $\surd$	79.6 $\surd$	77.4
$\widehat{OKS}$	75.1 ± .05	92.5 ± .05	81.8 ± .05	71.8 ± .04	80.2 ± .07	77.4 ± 0
OKS	77.2	93.1	83.2	73.6	82.6	77.4

predictions of the localization accuracy. However the AP obtained by $\widehat{OKS}$ is still about 2% lower than the upper bound, which indicates the difficulty of getting very precise predictions of *KS*.

Design choices of the *KS*-net. Here, we discuss the most suitable architecture of the *KS*-net. Specifically, four most famous backbone networks, AlexNet [44], VGGNet [45], GoogLeNet [46] and ResNet [38], are tested. The comparison of performance is given in Table 5. As can be seen, the results are not biased to one of the backbone networks. The discrepancies of the APs obtained by these four backbone networks are small, VGG-19 and ResNet-18 get a little better results than the other two. We choose ResNet-18 as the backbone network simply because it is the most widely used. Then, we investigate whether using a heavier backbone network ResNet-50 can improve the performance. The results reported in Table 5 show that adding the depth does not make the predictions better. The reason may be that heatmaps are already high-level features, it is unnecessary to use a very deep network to do the regression again. Lastly, ResNet-18 is replaced with just two convolution layers, we want to see if a very shallow backbone network would be qualified for the task. The APs obtained using this design are also presented in Table 5, which demonstrates a very shallow network is not competent to make precise predictions.

5.3. Performance of perceiving occluded human poses

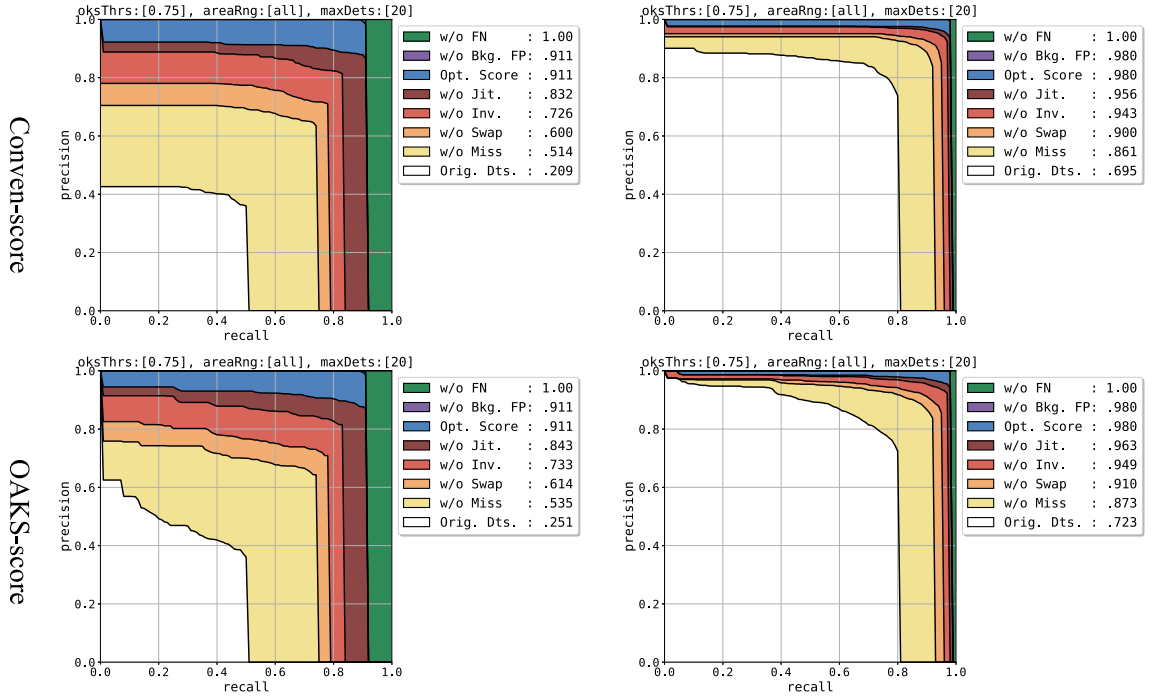
We evaluate the proposed scoring method on its performance of perceiving occluded human poses thoroughly in this subsection. Following the analysis tool for performing a detailed breakdown of the errors on MS-COCO dataset by Ronchi et al. [15], the MS-COCO 2017 validation dataset is separated into four benchmarks according to the number of Visible Keypoints (VK), which are Almost Occluded (VK$\in [1,5]$), Largely Occluded (VK$\in [6,10]$), Slightly Occluded (VK$\in [11,15]$) and No Occlusion (VK$\in [16,17]$). In this paper, we refer to poses with less than 10 visible keypoints as heavily occluded human poses. The PR curves obtained at the evaluation threshold of .75 OKS on each benchmark will be presented to investigate the performance to occlusion. The evaluation threshold of .75 OKS is chosen, because it is the median of the OKS thresholds (ranging from .5 to .95 in [17]) used for evaluation on MS-COCO dataset, which is also the default threshold to get the PR curves in the analysis tool [15].

The PR curves are obtained based on the pose detections given by OA-HPE net, both the poses detected using the ground truth human boxes and the detected bounding boxes [13] will be utilized for conducting the investigation. We will compare the proposed scoring method with the widely utilized scoring strategy by current methods, which uses the maximum likelihoods in heatmaps as the predictions of the localization accuracy of keypoints and ignores pose occlusion (the formulation is given in Eq. (3)). For convenience, we dub the proposed scoring method as OAKS-score, the conventional scoring strategy as Conven-score, and the optimal score using the ground truth OKS as Oracle-score.

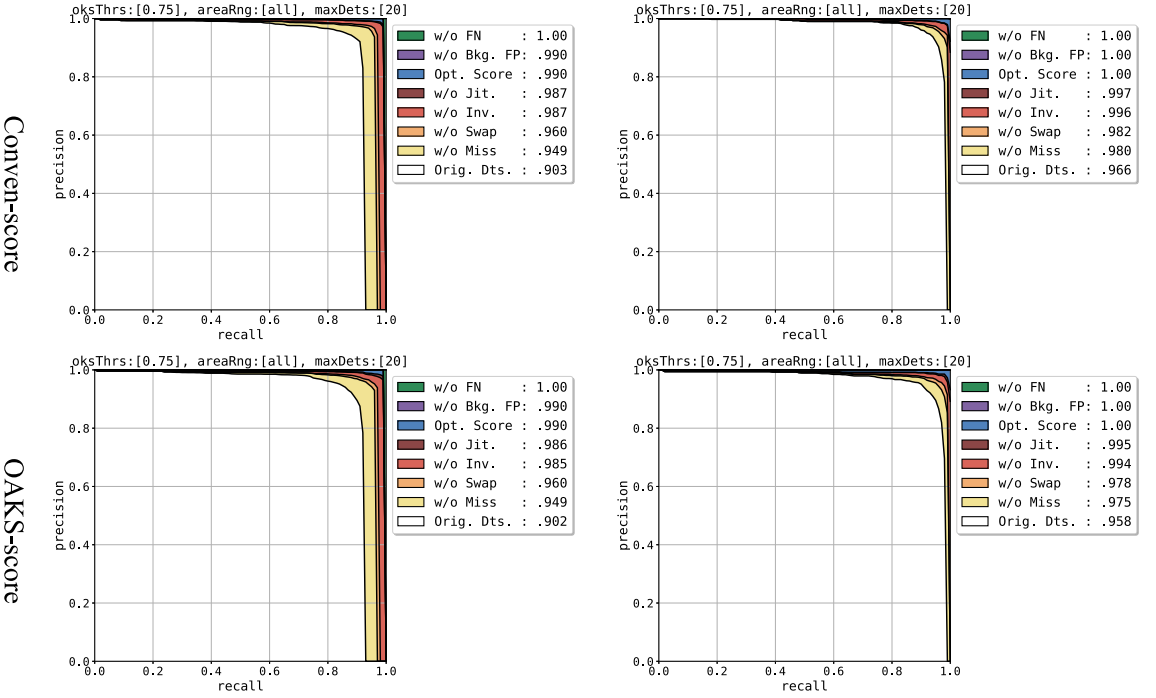
Fig. 11 exhibits the PR curves obtained using OAKS-score and Conven-score on the four different pose occlusion benchmarks. The legend in each sub-figure gives the overall AP values after progressively correcting errors of each type (details of the errors are explained in [15]). As can be seen, heavily occluded human poses (*i.e.*, Almost Occluded and Largely Occluded) create significant challenges. While the performance of pose estimation on Slightly Occluded and No Occlusion are near saturation, there are considerable errors of each type on the heavily occluded benchmarks, which includes the error of scoring (the area in blue color). Comparing the PR curves of the two different scoring methods, Conven-score is unable to assign high scores to good detections of heavily occluded poses, which leads to low precision even when the recall is still low. OAKS-score largely corrects this situation and improves the APs (the areas in white color) by 4.2% and 2.8% on the Almost Occluded and Largely Occluded benchmarks separately. The performance of OAKS-score and Conven-score on the other two benchmarks are comparable, this is reasonable since biased scoring cannot happen on visible poses.

Table 5
The comparison of APs obtained using different backbone networks of the *KS*-net. The APs are all calculated based on the settings used in Table 4.

Backbone	AP	$AP_{@.5}$	$AP_{@.75}$	AP_m	AP_l	AR	AR_m	AR_l
2 Conv	74.2	92.3	81.2	71.1	79.3	77.4	73.9	82.8
AlexNet	74.8	92.4	81.6	71.6	79.9	77.4	73.9	82.8
VGG-19	75.0	92.4	81.8	71.7	80.0	77.4	73.9	82.9
GoogLeNet	74.8	92.4	81.7	71.5	79.9	77.4	73.9	82.9
ResNet-18	75.1	92.5	81.9	71.9	80.3	77.4	73.9	82.8
ResNet-50	74.6	92.4	81.6	71.5	79.7	77.4	73.9	82.8

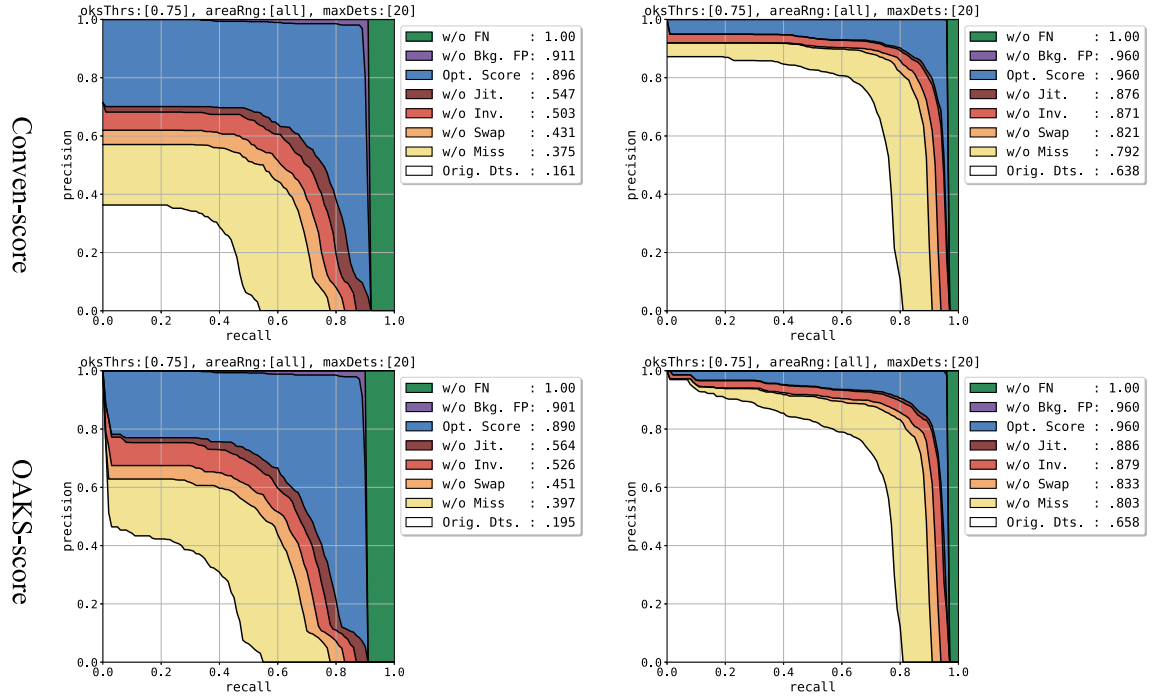


(a) Almost Occlusion (b) Largely Occlusion

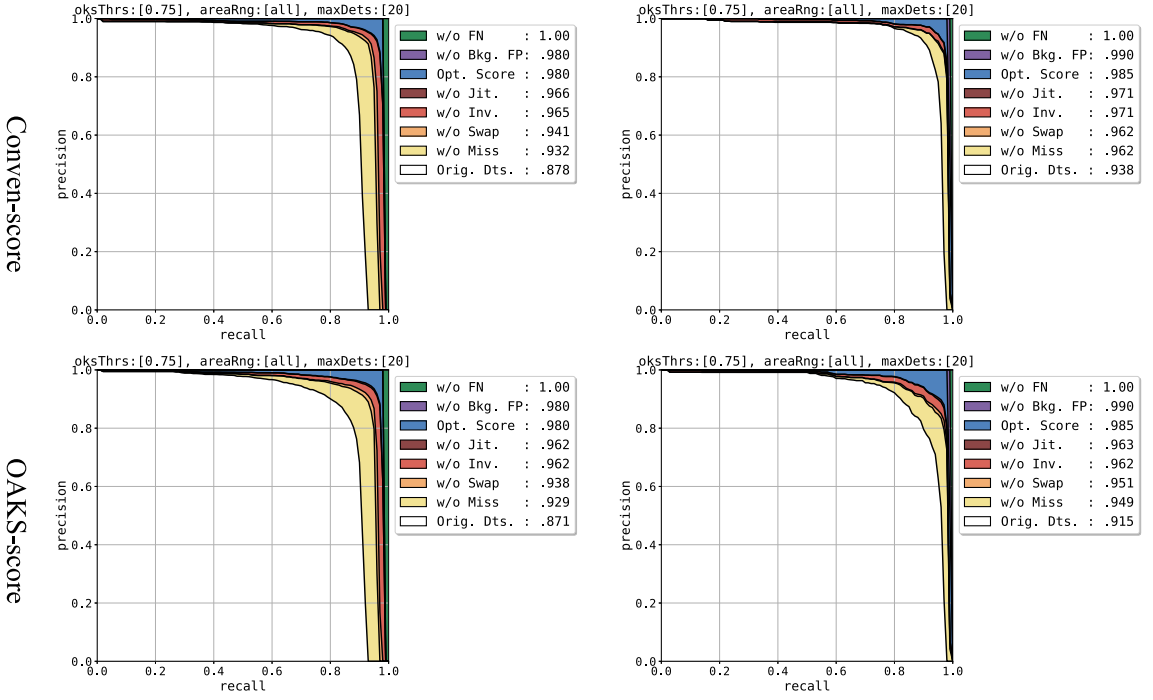


(c) Slightly Occlusion (d) No Occlusion

Fig. 11. Comparison of performance to Occlusion of OAKS-score and Conven-score based on the ground-truth human bounding boxes. The legend contains the overall AP values, the white color indicates the original AP obtained by Conven-score or OAKS-score. The AP values in other colors are obtained by progressively correcting different types of errors defined in [15]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Almost Occlusion (b) Largely Occlusion



(c) Slightly Occlusion (d) No Occlusion

Fig. 12. Comparison of performance to Occlusion of OAKS-score and Conven-score based on the detected human bounding boxes. In this case, though the score of the detected bounding box is needed to be considered for scoring the pose, OAKS-score is still able to get better AP values than Conven-score on the Almost Occlusion and Largely Occlusion Benchmarks.

To make the experiments comprehensive, we also do comparisons based on the pose detections using the detected human boxes [13], as shown in Fig. 12. In this case, the assigned score to each detected pose needs to multiply the score of its corresponding bounding box. Considering the detection of occluded persons is also a very challenging task [47], this makes refining the score of a detected pose more difficult. Nevertheless, OAKS-score also demonstrates better ability to perceive heavily occluded human poses, which outperforms Conven-score by 3.4% and 2.0% respectively on the Almost Occluded and Largely Occluded benchmarks.

Finally, we have a look at the APs obtained by the three different scoring methods on the whole MS-COCO 2017 validation set. Tables 6 and 7 present the results based on poses estimated using the ground truth human boxes and the detected bounding boxes [13] separately. Comparing these numerical APs, OAKS-score is only able to outdo Conven-score slightly in both cases. It seems that assigning unbiased scores by OAKS-score will not make much difference. However, the unbalanced data distribution of the MS-COCOC dataset covers up the truth. Table 8 shows the number of instances in each benchmark of the MS-COCO 2017 validation set. It appears that heavily occluded poses just make up a small proportion of the whole dataset. Thus, the improvement obtained by assigning unbiased scores on the heavily occluded pose benchmarks can be significantly diluted on the whole dataset. However, it is never wise to overlook the unfairness of scoring on heavily occluded poses. Considering that it has already been so hard to achieve a good pose estimation hypothesis on heavily occluded poses, we can not afford to discard it just because of a biased score.

5.4. The generality of the proposed scoring method

The efficiency of the proposed scoring method to improve the perception of heavily occluded human poses has been demonstrated in the above experiments. Here, we show its generality by applying to other methods for human pose estimation. Without loss of generality, the proposed scoring method is applied to two recently proposed human pose estimation methods [48,49]. Cheng et al. [48] detect 2D human poses in each frame of a video, then temporal information is exploited to produce 3D poses and diminish the affection of occluded joints. Based on their method of 2D human pose estimation, we employ the proposed scoring method. Kishore et al. [49] first detect and segment pedestrians by giving bounding boxes and

Table 6
The comparison of APs obtained by the three different scoring methods based on the pose detections using the ground truth human boxes. Conven-score is the method conventionally used by the existing pose estimation methods, OAKS-score is the proposed scoring method, Oracle-score gives the upper bound by using the optimal score (*i.e.*, OKS).

Method	AP	$AP_{@.5}$	$AP_{@.75}$	AP_m	AP_l	AR	AR_m	AR_l
Conven-score	74.6	92.4	81.5	71.4	79.6	77.4	73.9	82.8
OAKS-score	**74.8**	92.4	**81.7**	**71.6**	**80.0**	77.4	73.9	82.8
Oracle-score	77.2	93.1	83.2	73.6	82.6	77.4	73.9	82.8

The significance of bold in the Tables 6 and 7 is the best result among these methods for comparison.

Table 7
The comparison of APs based on the pose detections using the detected bounding boxes [13]. There is no Oracle-score in this case, because the corresponding of detected human boxes and the ground truth is unsure.

Method	AP	$AP_{@.75}$	AP_m	AP_l	AR	$AR_{@.5}$	$AR_{@.75}$	AR_m	AR_l
Conven-score	72.6	79.2	68.5	79.7	78.6	92.9	84.2	73.5	85.7
OAKS-score	72.6	79.2	68.5	**79.9**	**78.7**	**93.0**	**84.4**	**73.7**	85.7

The significance of bold in the Tables 6 and 7 is the best result among these methods for comparison.

Table 8
The number of instances in each benchmark of the MS-COCO 2017 validation set. "Occluded" is abbreviated to"Occlu" in the first row.

Benchmark	Almost Occlu	Largely Occlu	Slightly Occlu	No Occlu
Numbers	922	1775	2745	910
Proportion	14.5%	27.9%	43.2%	14.4%

Table 9
Comparison of $AP_{@.75}$ evaluation results on the four occlusion benchmarks by the original models and employing the proposed scoring method.

Method	OAKS-score	Almost Occ	Largely Occ	Slightly Occ	No Occ
Cheng [48]		27.4	82.3	93.3	98.1
	✓	38.6	88.3	93.8	97.3
Kishore [49]		23.7	82.4	94.1	98.1
	✓	40.2	90.0	95.0	98.1

masks in images, then 2D human poses are estimated by using the detection. Our scoring method can be utilized in their framework of human pose estimation.

We use MS COCO dataset [11] to validate the generality of the proposed scoring method, because it is the most widely used dataset for person keypoints localization. Both the 2D human pose estimators of Cheng et al. [48] and Kishore et al. [49] are trained from scratch on MS COCO 2017 training set, and the testing is done on MS COCO 2017 validation set to exhibit the superiority of our scoring method over their origins. We train both models for 160 epochs, and data augmentation strategies [13] are used. The Adam optimizer with the base learning rate $1e-3$ is utilized to train the models, the learning rate is decreased by a factor of 0.1 after 120 and 150 epochs respectively. Following the analysis tool of Ronchi et al. [15], we report the precision $AP_{@.75}$ (AP at OKS = 0.75) on the four occlusion benchmarks to compare the performance.

The results are presented in Table 9. The original models of Cheng et al. [48] and Kishore et al. [49] get $AP_{@.75}$ 27.4% and 23.7% on the Almost Occluded benchmark, employing the proposed scoring method is able to boost the performance of both models to 38.6% and 40.2% respectively. The improvement on the Largely Occluded benchmark is also considerable, the precision increases by 6.0% and 7.6% on the two models respectively. This clearly shows that the proposed scoring method can be applied to different models and make the estimation of heavily occluded human poses better.

6. Conclusion

This paper proposes to better perceive heavily occluded human poses by assigning an unbiased score. We give a detailed analysis of the conventional scoring method used by existing frameworks, and find two reasons that make it unsuitable for occluded poses, which are ignorance of pose occlusion and misalignment of classification and localization accuracy. Thus, a novel occlusion aware human pose estimation network is proposed to acquire the visibility of keypoints while heatmaps are regressed. Moreover, *KS*-net is proposed to predict the keypoint similarity between detected keypoints and their corresponding ground truth, which replaces the maximum heat value of heatmaps and naturally represents localization accuracy. Experimental results on MS-COCO dataset demonstrate that the proposed scoring method significantly improves the detection accuracy of heavily occluded human poses. In future work, we plan to design an end-to-end framework to give detections of human poses and assign the corresponding unbiased scores simultaneously.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Lin Zhao: Conceptualization, Methodology, Software, Writing - original draft. **Jie Xu:** Software, Validation. **Shanshan Zhang:** Formal analysis, Resources. **Chen Gong:** Writing - review & editing. **Jian Yang:** Project administration, Funding acquisition. **Xinbo Gao:** Supervision.

Acknowledgments

This work was supported by NSF of China (Grant Nos. 61802189, 61973162, 61702262, U1713208), NSF of Jiangsu Province (Grant Nos. BK20180464, BK20171430, BK20181299), the Fundamental Research Funds for the Central Universities (Grant Nos. 30918014107, 30919011280, 30920032202, 30920032201), the "Young Elite Scientists Sponsorship Program" by Jiangsu Province, and the "Young Elite Scientists Sponsorship Program" by CAST (Grant No. 2018QNRC001), Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 61861136011), Science and Technology on Parallel and Distributed Processing Laboratory (PDL) Open Fund (Grant No. WDZC20195500106).

References

[1] H. Li, J. Xu, J. Zhu, D. Tao, Z. Yu, Top distance regularized projection and dictionary learning for person re-identification, Inf. Sci. 502 (2019) 472–491.
[2] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 650–667.
[3] T. Huynh-The, C.-H. Hua, N.A. Tu, T. Hur, J. Bang, D. Kim, M.B. Amin, B.H. Kang, H. Seung, S.-Y. Shin, et al, Hierarchical topic modeling with pose-transition feature for action recognition using 3d skeleton data, Inf. Sci. 444 (2018) 20–35.
[4] D.C. Luvizon, D. Picard, H. Tabia, 2d/3d pose estimation and action recognition using multitask deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5137–5146.
[5] R. Huang, H. Cheng, H. Guo, X. Lin, J. Zhang, Hierarchical learning control with physical human-exoskeleton interaction, Inf. Sci. 432 (2018) 584–595.
[6] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.
[7] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 483–499.
[8] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1281–1290.

[9] S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, in: Proceedings of the British Machine Vision Conference, 2010, pp. 12.1–12.11.
[10] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014, pp. 3686–3693..
[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[12] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, Towards accurate multi-person pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4903–4911.
[13] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103–7112.
[14] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 466–481.
[15] M. Ruggero Ronchi, P. Perona, Benchmarking and error diagnosis in multi-instance pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 369–378.
[16] B. Jiang, R. Luo, J. Mao, T. Xiao, Y. Jiang, Acquisition of localization confidence for accurate object detection, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 784–799.
[17] Coco keypoints evaluation, URL: http://cocodataset.org/#keypoints-eval, october, 2016..
[18] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2878–2890.
[19] L. Zhao, X. Gao, D. Tao, X. Li, Tracking human pose using max-margin markov models, IEEE Trans. Image Process. 24 (12) (2015) 5274–5287.
[20] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, Poselet conditioned pictorial structures, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 588–595.
[21] L. Zhao, X. Gao, D. Tao, X. Li, A deep structure for human pose estimation, Signal Process. 108 (2015) 36–45.
[22] A. Toshev, C. Szegedy, Deeppose: human pose estimation via deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.
[23] J.J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 1799–1807..
[24] J. Carreira, P. Agrawal, K. Fragkiadaki, J. Malik, Human pose estimation with iterative error feedback, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4733–4742.
[25] X. Chen, A.L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 1736–1744..
[26] I. Lifshitz, E. Fetaya, S. Ullman, Human pose estimation using deep consensus voting, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 246–260.
[27] U. Rafi, B. Leibe, J. Gall, I. Kostrikov, An efficient convolutional network for human pose estimation, in: Proceedings of the British Machine Vision Conference, 2016, pp. 109.1–109.11.
[28] G. Ning, Z. Zhang, Z. He, Knowledge-guided deep fractal neural networks for human pose estimation, IEEE Trans. Multimedia 20 (5) (2018) 1246–1259.
[29] L. Ke, M.-C. Chang, H. Qi, S. Lyu, Multi-scale structure-aware network for human pose estimation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 713–728.
[30] X. Chu, W. Yang, W. Ouyang, C. Ma, A.L. Yuille, X. Wang, Multi-context attention for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1831–1840.
[31] Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial posenet: a structure-aware convolutional network for human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1212–1221.
[32] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P.V. Gehler, 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, 2016, pp. 4929–4937.
[33] U. Iqbal, A. Milan, J. Gall, Posetrack: joint multi-person pose estimation and tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2011–2020.
[34] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, Deepercut: a deeper, stronger, and faster multi-person pose estimation model, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 34–50.
[35] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.
[36] A. Newell, Z. Huang, J. Deng, Associative embedding: end-to-end learning for joint detection and grouping, in: Advances in Neural Information Processing Systems, 2017, pp. 2277–2287..
[37] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99..
[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[39] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, Rmpe: regional multi-person pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2334–2343.
[40] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969..
[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[42] Coco keypoints leaderboard challenge17. URL: http://cocodataset.org/#keypoints-leaderboard, october, 2017..
[43] S. García, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power, Inf. Sci. 180 (10) (2010) 2044–2064.
[44] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105..
[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556..
[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[47] S. Zhang, J. Yang, B. Schiele, Occluded pedestrian detection through guided attention in cnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6995–7003.
[48] Y. Cheng, B. Yang, W. Yan, T.R. Tan, Occlusion-aware networks for 3d human pose estimation in video, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 723–732.
[49] P.S.R. Kishore, S. Das, P.S. Mukherjee, U. Bhattacharya, A deep framework for occluded pedestrian pose estimation, in: Proceedings of British Machine Vision Conference, 2019, pp. 1–15..