On the Calibration of Human Pose Estimation

Kerui Gu* Rongyu Chen* Angela Yao
National University of Singapore
{keruigu, rchen, ayao}@comp.nus.edu.sg

Abstract

Most 2D human pose estimation frameworks estimate keypoint confidence in an ad-hoc manner, using heuristics such as the maximum value of heatmaps. The confidence is part of the evaluation scheme, e.g.*, AP for the MSCOCO dataset, yet has been largely overlooked in the development of state-of-the-art methods. This paper takes the first steps in addressing miscalibration in pose estimation. From a calibration point of view, the confidence should be aligned with the pose accuracy. In practice, existing methods are poorly calibrated. We show, through theoretical analysis, why a miscalibration gap exists and how to narrow the gap. Simply predicting the instance size and adjusting the confidence function gives considerable AP improvements. Given the black-box nature of deep neural networks, however, it is not possible to fully close this gap with only closed-form adjustments. As such, we go one step further and learn network-specific adjustments by enforcing consistency between confidence and pose accuracy. Our proposed Calibrated ConfidenceNet (CCNet) is a light-weight post-hoc addition that improves AP by up to 1.4% on off-the-shelf pose estimation frameworks. Applied to the downstream task of mesh recovery, CCNet facilitates an additional 1.0mm decrease in 3D keypoint error. Project page: comp.nus.edu.sg/keruigu/calibrate_pose/project.html.*

1. Introduction

Two common frameworks for human pose estimation are either based on heatmaps or direct regression. Heatmap methods [29, 34, 35] estimate a non-parametric spatial likelihood on the keypoints in the form of a dense heatmap. Direct regression methods regress the keypoint location either as a deterministic output [8, 30, 31] or as a distribution [17, 21].

The accuracy of the predicted pose is typically based on some distance measure with respect to the ground truth pose. Several metrics for accuracy exist, including end-point-error (EPE) and percentage of correct keypoints (PCK), but these have drawbacks due to their inability to account for the person size and annotation error differences among keypoints, leading to the dominance of Object Keypoint Similarity (OKS). By considering these factors, OKS is perceptually meaningful and a good interpretation of similarity [26].

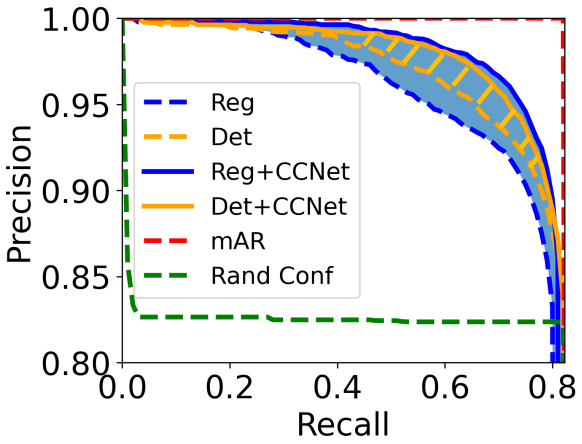

Figure 1. Confidence estimation plays a significant role in the Precision-Recall curve. The area under each curve represents the primary metric AP. As shown, adding CCNet on top of existing detection- and regression-based methods improves the AP. The orange dashed line area and blue area denote the improvement of detection- and regression-based methods, respectively.

A more comprehensive and widely accepted metric for evaluating pose is the mean Average Precision (mAP) based on the object keypoint similarity (OKS) [19]. This is fashioned directly after object detection evaluation schemes by viewing the pose estimation as a keypoint "detection". The mAP, as it is defined, goes beyond evaluating accuracy and also incorporates the corresponding confidence of the estimated pose. A simple experiment shows for SBL [34] that if we randomize the confidence estimation, the area under curve significantly drops (green dashed line in Fig. 1) and mAP drops from 72.4 to 67.5; instead, if we assign the confidence with the OKS, the area increases (red dashed line in Fig. 1) and mAP increases to 75.6 (see Sec. 4.1 for de-

*Equal contribution

tails). This clearly demonstrates that confidence has a significant impact on evaluation, but is currently overlooked.

First and foremost, the natural question that arises is: how calibrated are current pose estimation frameworks since most approaches use some heuristics such as taking the maximum value of the predicted heamaps [29, 34, 35] or uncertainty [17, 21]? To answer this question, we first analyze, from a statistical point of view, the expected confidence of pose estimation methods versus the ideal confidence given by the OKS under a common assumption that the ground truth annotations follow a Gaussian distribution. Our analysis reveals systematic miscalibrations based on the way mAP is an exponential envelope functioned by endpoint error, instance size, and keypoint annotation falloff. This manifests as a scaling gap for heatmap methods and a form gap for RLE-based methods. Empirically, we verify our analysis by provide a closed-form solution to align the expected OKS, which improves AP with only additionally predicting instance size and changing the confidence form.

However, only correcting the confidence form is not enough since the above analysis is conditioned on a perfect network, which assumes the predicted keypoint can always be located at the center of ground truth distributions. In practice, the network will vary depending on different backbones and datasets, so this assumption is hard to achieve and we aim to learn a network-specific adjustment to better calibrate the confidence.

In this way, we propose a simple yet effective calibration branch, Calibrated ConfidenceNet (CCNet), which is applicable to any pose estimation methods. Specifically, fixing the trained pose models, we base our calibration branch on the penultimate features of the original pose models, which contain rich pose-related information of the input image, and explicitly output score and visibility. We simply supervise the outputs of the calibration branch with the computed OKS and ground truth visibility. By doing so, the predicted confidence directly links to the practical OKS which addresses the issues of different confidence forms and different network characteristics simultaneously. With only a few epochs and negligible additional parameters, the network is much better calibrated and achieves better mAP on various pose models and various datasets.

Since the predictions from the 2D pose models serve as prior knowledge for downstream tasks, we are among one of the first methods that test the influence of confidence on downstream tasks. We take the example of 3D mesh recovery which fits a SMPL [20] model to human mesh according to 2D predictions. Experimentally, we show that with a better calibrated pose model, the 2D predictions can produce better 3D results.

Our contribution can be summarized as,

- We are the first to study the calibration of 2D pose estimation, which is overlooked in the literature but significantly counts in the evaluation of pose estimation models and the challenging downstream tasks.
- We mathematically formulate the ideal form of pose confidence and reveal a mismatch between this and the practical confidence form from current pose estimation methods. A simple solution is provided to verify and refine the misalignment.
- We propose a simple but effective method to explicitly model the calibration with minor addition of parameters and training time. Experiments show that adding the calibration branch gives significant improvement on the primary metric mAP and also benefits the downstream tasks.

2. Related Work

Pose Estimation. The past literature in 2D top-down based pose estimation mainly focused on how to improve the accuracy. However, only few works give some heuristic or empirical understandings regarding the pose confidence. Papandreou *et al*. [22] proposed a re-scoring strategy based on the detected bounding box, which is shown to be effective and applied in most of the following top-down based works [17, 34]. PETR [27] empirically found that changing the matching objective to OKS-based improves the AP under the same AR, indicating a better ranking over the samples. Poseur [21] following regression paradigm noticed previous regression scoring is heuristic and they rescore into likelihood presented by detection maxvals. Although there exist several works that try to change the form of confidence, they remain at the empirical level and lack actual understanding of the confidence. In our paper, we give theoretical understanding of the confidence for both heatmap- and RLE-based methods and analyze their calibration to OKS accordingly. As a result, we propose the Calibrated ConfidenceNet as the solution to achieve better calibration for pose models.

Confidence Estimation is essential in real-world applications and has pushed a lot of work in that direction [1, 12, 16]. Conventional calibration is broadly discussed in the classification [9] and regression [28]. It reflects the reliability of the confidence to indicate accuracy. The confidence in the context is interpreted as the probability of the prediction being correct. For instance, in classification, the softmax confidence is viewed as the probability of the image belonging to the class. Well-calibrated classifier is expected to have confidence approximate accuracy [9]. Similarly, recent work [23] introduced calibration to boundary box object classification in object detection. When it comes to regression the definition is vague, among which a quantile-based definition is common [28]. Evaluation becomes nontrivial in high dimensions. However, there is few work studying calibration in pose estimation [3, 24]. We argue that pose confidence is useful and informative only

when it first aligns well with true OKS, otherwise, it cannot be a helpful forecast [15]. mAP exactly evaluates this.

3. Preliminaries

3.1. Human Pose Estimation

We consider top-down 2D human pose estimation, where people are already localized and cropped from the scene. Given a single-person image $\mathbf{x}$, pose estimation methods estimate K keypoint coordinates $\hat{\boldsymbol{p}} \in \mathbb{R}^{K\times 2}$ and confidence score $\hat{\boldsymbol{s}} \in [0,1]^K$. The keypoint scores are aggregated into a person- or instance–wise confidence $\hat{c} \in [0,1]$ where higher values indicate higher confidence.

Heatmap methods [29, 34, 35] estimate K heatmaps $\hat{\boldsymbol{H}} \in \mathbb{R}^{K\times H\times W}$ to represent pseudo-likelihoods, *i.e.* unnormalized probabilities of each pixel being the k-th keypoint (Fig. 2 (a)). The heatmap $\hat{\boldsymbol{H}}_k$ can be decoded into the joint coordinate $\hat{\boldsymbol{p}}_k$ and confidence $\hat{s}_k$ with a simple $\arg\max$:

$$\hat{\boldsymbol{p}}_k = \operatorname{argmax}(\hat{\boldsymbol{H}}_k), \quad \hat{s}_k = \max(\hat{\boldsymbol{H}}_k), \tag{1}$$

although more complex forms of decoding have been proposed [38] in place of the $\arg\max$.

Methods which estimate heatmaps are learned with an MSE loss with respect to a ground truth heatmap $\boldsymbol{H}_k$

$$\mathcal{L}_{\text{det}} = \sum_{k=1}^{K} \text{MSE}(\hat{\boldsymbol{H}}_k, \boldsymbol{H}_k) = \sum_{k=1}^{K}\sum_{i=1}^{H}\sum_{j=1}^{W}(\hat{h}_{kij} - h_{kij})^2, \tag{2}$$

where (i,j) are spatial indices, and $\hat{h}_{kij}, h_{kij}$ denotes the $(i,j)^{\text{th}}$ entry of $\hat{\boldsymbol{H}}_k$ and $\boldsymbol{H}_k$, Typically, the ground truth heatmap $\boldsymbol{H}_k$ is constructed as 2D Gaussian centered at the ground truth keypoint location and a fixed standard deviation $\tilde{l}$, *e.g.*, $\tilde{l}=2$ for input size 256×192.

Regression methods directly regress deterministic coordinates of the keypoints or likelihood distributions of the coordinates. We focus on the state-of-the-art RLE regression [17, 21], which generally model the likelihood as a distribution (*e.g.*, Normal, Laplace and Normalizing Flows) parameterized by mean and variance parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}$. Similarly, we can achieve the keypoint prediction and its confidence as

$$\hat{\mathbf{p}} = \hat{\boldsymbol{\mu}}, \quad \hat{s}_k = 1 - \hat{\sigma}. \tag{3}$$

The loss for formulated as a Negative Log-Likelihood:

$$\mathcal{L}_{\text{reg}} = -\sum_{k=1}^{K} \log \hat{p}(\boldsymbol{p}_k | \mathbf{x}; \hat{\boldsymbol{p}}_k, \hat{\sigma}_k), \tag{4}$$

which can be further expanded as an adaptive weighted error loss between $\hat{\mathbf{p}}_k$ and $\mathbf{p}_k$, plus a regularization such as $\log \hat{\sigma}_k^2$ in practice.

Other regression-based methods use heatmap maximum [33] or keypoint classification confidence from another head [18] or simply fill the confidence as 1 [30].

Instance-wise confidence scores are derived by aggregating the keypoint confidences with a weighted summation:

$$\hat{c} = \text{agg}(\hat{\boldsymbol{s}}) = \sum_{k=1}^{K} \hat{w}_k \hat{s}_k, \text{ where } \hat{w}_k = \frac{\mathcal{I}(\hat{s}_k > \tau_{\hat{s}})}{\sum_{k=1}^{K} \mathcal{I}(\hat{s}_k > \tau_{\hat{s}})}, \tag{5}$$

where $\mathcal{I}$ is an indicator function. The convention of the keypoint-to-instance aggregation $\text{agg}(\cdot)$ is an average function of only selected keypoint predictions with $\hat{s}_k > \tau_{\hat{s}}$.

3.2. Pose Estimation Evaluation

Several metrics have been proposed for evaluating keypoint accuracy, including end-point error (EPE), Percentage of Correct Keypoint (PCK) and Object Keypoint Similarity (OKS) [19]. Of these three error measures, OKS is the most comprehensive. EPE does not account for the scale of the person; PCK, while normalized with respect to the head size, does not account for keypoint difference. OKS factors in both instance size and keypoint variation as an instance measure. It is defined as a weighted sum of the exponential envelope of a scaled end-point error:

$$c = \sum_{k=1}^{K} w_k \exp\left(-\frac{\|\hat{\boldsymbol{p}}_k - \boldsymbol{p}_k\|^2}{2l_k^2}\right), \tag{6}$$

$$\text{where } w_k = \frac{v_k}{\sum_{k=1}^{K} v_k}, \quad \text{and } l_k^2 = \text{var}_k a, \tag{7}$$

where a is the body area, var_k is a per-keypoint annotation falloff constant, and v_k is a visibility indicator equal to 1 only if keypoint k is present in the scene[1]. The scaling l_k within the exponential envelope accounts for differences in scale across the different body joints and overall area of the pose while the weighting w_k takes presence in the scene into account.

A person instance is regarded as correct (positive) if its OKS exceeds some threshold. Over a dataset with N samples, we can tabulate the mean Average Recall (mAR) and mean Average Precision (mAP) as follows over T thresholds $\{\tau_t\}$:

$$\text{mAR} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{\mathcal{I}(c_i > \tau_t)}{N}. \tag{8}$$

This equation clearly states that mAR is **ranking-independent** to the predicted confidence and purely evaluates the accuracy of poses. However, the primary metric used for evaluating 2D pose estimation is mean Average

[1]Present includes both occluded and unoccluded keypoints within the bounding box crop.

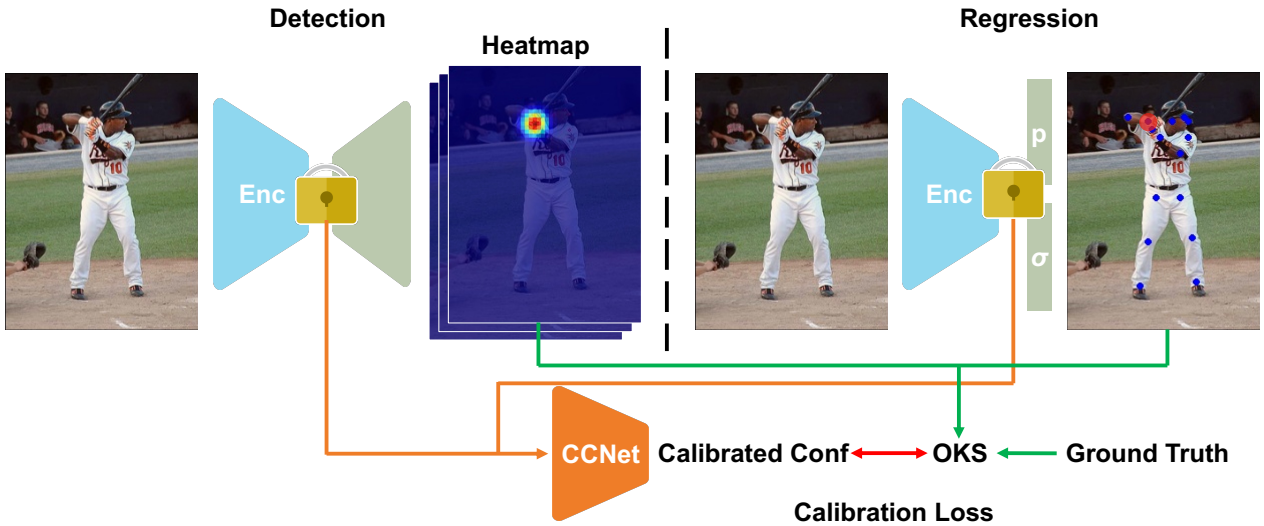

Figure 2. Our calibration method is a post-hoc addition to off-the-shelf pose estimation methods. CCNet estimates more calibrated confidences directly from latent pose representations and leads to improvement in the mAP.

Precision (mAP). The mAP is defined as

$$\text{mAP} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i'=1}^{N}\frac{\mathcal{I}(c_{i'} > \tau_t)}{N}\cdot\frac{\sum_{j=1}^{i'}\mathcal{I}(c_j > \tau_t)}{i'}, \quad (9)$$

where i' denotes an index based on the instances sorted according to their estimated confidences, *i.e.* $\hat{c}_1 \geq \ldots \hat{c}_{i'} \geq \cdots \geq \hat{c}_N$. The mAP therefore relies on the estimated confidences $\hat{c}$ to be consistent with the OKS in relative ordering and is **dependent** on the ranking of the predicted confidence. Note that this formulation of mAP is the same as the Area Under (maximum) Precision-Recall Curve (AUPRC) as in conventional classification [25][2].

4. An Analysis on Pose Calibration

4.1. Motivation

As stated in Equation 9, a high mAP not only requires an accurate prediction but also a good alignment between confidence and OKS. Intuitively, to achieve the upper bound of mAP, the ranking of confidence should be the same of that of OKS. However, no attempts have been made in the pose estimation literature to systematically learn the impact of the confidence estimation. As a sanity test, we replace the estimated score in heatmap-based and RLE-based method with a constant value. As shown in the second and third columns of Table 1, AP significantly drops around 5 point, which indicates the significant impact of confidence in the evaluation of AP. With this observation, we start our method by theoretically understanding how well current methods predict the confidence compared to the OKS.

[2]Note that another conventional metric AUROC=AR in the context.

4.2. Assumptions

For the analysis, we formulate the expected OKS and expected confidence of both heatmap- and RLE-based methods from a statistical perspective. Specifically, we follow two standard assumptions [17, 34]. First, the K keypoints of a person are conditionally independent, given the image itself. For clarity, we will drop the k indices in this section. Secondly, we assume that the ground truth location of each keypoint in an image follows a Gaussian distribution $\mathbf{p} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$. This Gaussian models ambiguities and errors in the annotation process [4, 17], where $\boldsymbol{\mu}$ specifies the true underlying location. For simplicity, we consider an isotropic Gaussian in our exposition and develop our analysis only in terms of variance σ^2, although the analysis can easily be extended for non-isotropic cases as well with a full covariance.

4.3. Expected OKS

We study the pose calibration problem by first deriving the expected OKS for a given estimated pose $\hat{\mathbf{p}}$ and the assumed Gaussian distribution for the ground truth pose $\mathbf{p}$. Based on the definition of OKS in equation 6, the expected OKS is given as

$$\mathbb{E}_{\mathbf{p}}[\text{OKS}] = \mathbb{E}_{\mathbf{p}}\left[\exp\left(-\frac{\|\hat{\mathbf{p}}-\mathbf{p}\|^2}{2l^2}\right)\right] \quad (10)$$

$$= \frac{l^2}{\sigma^2+l^2}\exp\left(-\frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2(\sigma^2+l^2)}\right), \quad (11)$$

which is a function of $\{\boldsymbol{\mu}, \sigma, l, \hat{\mathbf{p}}\}$. When a network is perfectly trained, $\hat{\mathbf{p}}$ will approach $\boldsymbol{\mu}$ and the corresponding

OKS is represented as

$$s_{\text{OKS}} = \frac{l^2}{\sigma^2 + l^2} = 1 - \frac{\sigma^2}{\sigma^2 + l^2}. \tag{12}$$

4.4. Expected Confidence

Heatmap methods synthesize a ground truth heatmap $\boldsymbol{H}$ by constructing an isotropic Gaussian centered at the ground truth $\mathbf{p}$ and a standard deviation of $\tilde{l}$ set heuristically, *e.g.*, $\tilde{l} = 2$. Given our previous assumption on the distribution of $\mathbf{p}$, the effective ground truth can be expressed as $\tilde{\mathbf{p}} \sim \mathcal{N}(\mu, \sigma^2 + \tilde{l}^2)$ or in heatmap form as

$$\mathrm{h}_{\tilde{\mathbf{p}}} = 2\pi\tilde{l}^2 p(\tilde{\mathbf{p}}|\mathbf{p}) = \exp\left(-\frac{\|\tilde{\mathbf{p}} - \mathbf{p}\|^2}{2\tilde{l}^2}\right), \tag{13}$$

If we consider the predicted heatmap $\hat{h}$ as that which minimizes the MSE loss in equation 2, we arrive at the following:

$$\hat{h} = \arg\min_{\hat{h}} \mathbb{E}_{\mathbf{p}}[(\hat{h} - \mathrm{h})^2] = \mathbb{E}_{\mathbf{p}}[\mathrm{h}] \tag{14}$$

$$= \int_{\mathbf{p}} p(\mathbf{p}|\mathbf{x}) \cdot 2\pi^2 p(\tilde{\mathbf{p}}|\mathbf{p}) \mathbf{dp} = 2\pi^2 p(\tilde{\mathbf{p}}|\mathbf{x}). \tag{15}$$

The resulting optimal spatial heatmap $\hat{\boldsymbol{H}} = \{\hat{h}\} \overset{c}{\approx} \mathcal{N}(\boldsymbol{\mu}, \hat{\sigma}^2 \boldsymbol{I})$ with emerging $\hat{\sigma}^2 = \sigma^2 + \tilde{l}^2$, approximates rendered ground truth heatmap (see Supplementary for a similar derivation for the case with an imperfect mean). This derivation highlights that predicted heatmaps learned with a pixel-wise MSE loss exhibit a standard deviation slightly larger than $\tilde{l} = 2$ even if the coordinate prediction is accurate [7]. See Supp. A for a verification.

It follows from equation 15 that the predicted confidence, defined as the max from equation 1 and is located at $\hat{\boldsymbol{p}} \approx \boldsymbol{\mu}$ has a value given by,

$$\hat{s}_{\text{det}} = \hat{\mathrm{h}}_{\boldsymbol{\mu}} = 2\pi\tilde{l}^2 p(\tilde{\mathbf{p}} = \boldsymbol{\mu}|\mathbf{x}) \tag{16}$$

$$= \frac{2\pi\tilde{l}^2}{2\pi(\sigma^2 + \tilde{l}^2)} \exp\left(-\frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}\|^2}{2(\sigma^2 + \tilde{l}^2)}\right) = \frac{\tilde{l}^2}{\sigma^2 + \tilde{l}^2} = \frac{\tilde{l}^2}{\hat{\sigma}^2}. \tag{17}$$

The two expected values from equation 12 and equation 17 are different at the same location $\boldsymbol{\mu}$. This difference comes from that $\tilde{l}$ is a constant value whereas l changes according to different instance sizes and keypoints.

RLE-based regression methods are learned by minimizing a negative log-likelihood over the predicted distribution as shown in equation 4. For simplicity, we consider a normal distribution (an alternative distribution such as the Laplace or Normalizing Flow do not change the conclusions) $\mathbf{p}' \sim \mathcal{N}(\hat{\boldsymbol{p}}, \hat{\sigma}^2 \boldsymbol{I})$ are trained by over distribution, which is treated as a maximum optimization,

$$\max_{\hat{p}, \hat{\sigma}} \mathbb{E}_{\mathbf{p}}\left[\log \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(-\frac{(\mathbf{p} - \hat{\mathbf{p}})^2}{2\hat{\sigma}^2}\right)\right]. \tag{18}$$

Method	Orig	*const.*	OKS			mAR↑
			Mean	Pred	GT	
SBL [34]	72.4	67.5	72.2	73.0	73.6	75.6
RLE [17]	72.2	67.2	71.8	73.2	73.3	75.4

Table 1. mAP under different confidence estimation except the last column. 'Orig' means applying their original ways of estimating confidence. '*consts.*' means replacing the confidence with a constant value. 'Mean', 'Pred', 'GT' corresponds to adjusting the confidence prediction with equation 12 of mean area, predicted area (with additional supervision), and ground truth area. Results show that confidence estimation significantly influences the final evaluation and our closed-form adjustment can increase the mAP.

When equation 18 is maximized, the predictive distribution approximates the optimal, $\hat{p} \approx \mu, \hat{\sigma} \approx \sigma, \mathbf{p}' \approx \mathbf{p}$. We give detailed derivations in the Supplementary.

Subsituting the $\hat{\sigma}$ from above into the heuristic score for RLE given in equation 3, we arrive at

$$\hat{s}_{\text{reg}} = 1 - \hat{\sigma} = \mathbb{E}\left[1 - \sqrt{\frac{\pi}{8}}\|\hat{\mathbf{p}} - \mathbf{p}\|_1\right]. \tag{19}$$

Comparing equation 12 with equation 19, the expected confidence of RLE-based methods takes a linear form of $\hat{\sigma}$ and only models the annotation variation but ignores the instance size. Besides, it averages across all keypoints not excluding invisible ones, *i.e.*, $\hat{\boldsymbol{v}} = \mathbf{1}$, leading to more inconsistencies with OKS (equation 7).

4.5. Discussion

Misordered Ranking. Although the three confidence forms all give descending curves, the different decay rates will influence the ranking. One explanation is that when it comes to a specific sample, the actual OKS will vary according to different samples but the expected confidences of both heatmap- and RLE-based methods remain unchanged since they don't consider the instance size and keypoint falloff constants. Image two similar σs, OKS is likely to have different rankings depending on l, which becomes inconsistent with the ranking of expected confidences.

Confidence Correction. Motived by the above analysis, we provide a simple confidence correction to make the pose network better calibrated to the OKS based on the different confidence forms, which, at the same time, serves as the empirical verification of our theoretical derivations. With original $\hat{\sigma}$ (for detection one way is to fit heatmap with Gaussian to get) and estimated or ground truth per sample l, scores can be adjusted to equation 11. Table 1 demonstrates that appropriately adjusting the confidence form to the ideal one improves mAP. Yet, this rescoring is based on the dismantling of metrics in ideal assumptions. We give the limitations as follows.

Method	Confidence	Backbone	Input Size	#Params (M)	#GFLOPs	mAP↑	AP.5↑	AP.75↑	AP (M)↑	AP (L)↑	mAR↑
Detection											
SBL [34]	Hm	ResNet-50	256×192	34.00	5.46	72.4	91.5	80.4	69.8	76.6	75.6
+CCNet		ResNet-50	256×192	34.08	5.52	73.3 (+0.9)	92.6	80.9	70.4	77.5	75.6
SBL [34]	Hm	ResNet-152	384×288	68.64	12.77	76.5	92.5	83.6	73.6	81.2	79.3
+CCNet		ResNet-152	384×288	68.71	12.83	77.3 (+0.8)	93.5	84.1	74.0	81.6	79.3
HRNet [29]	Hm	HRNet-W32	256×192	28.54	7.7	76.0	93.5	83.4	73.7	80.0	79.3
+CCNet		HRNet-W32	256×192	28.62	7.76	77.0 (+1.0)	93.7	84.0	74.0	81.0	79.3
HRNet [29]	Hm	HRNet-W48	384×288	63.62	15.31	77.4	93.4	84.4	74.8	82.1	80.9
+CCNet		HRNet-W48	384×288	73.69	15.36	78.3 (+0.9)	93.6	85.1	75.5	83.4	80.9
ViTPose [35]	Hm	ViT-Base	256×192	89.99	17.85	77.3	93.5	84.5	75.0	81.6	80.4
+CCNet		ViT-Base	256×192	90.07	17.91	78.1 (+0.8)	93.7	85.0	75.4	83.3	80.4
Regression											
RLE [17]	Reg	ResNet-50	256×192	23.6	4.0	72.2	90.5	79.2	71.8	75.3	75.4
+CCNet		ResNet-50	256×192	23.6	4.0	73.6 (+1.4)	91.6	80.2	72.0	77.6	75.4
RLE [17]	Reg	ResNet-152	384×288	58.3	11.3	76.3	92.4	82.6	75.6	79.7	79.2
+CCNet		ResNet-152	384×288	58.3	11.3	77.1 (+0.8)	92.6	83.2	75.6	81.3	79.2
RLE [17]	Reg	HRNet-W32	256×192	39.3	7.1	76.7	92.4	83.5	76.0	79.3	79.4
+CCNet		HRNet-W32	256×192	39.3	7.1	77.5 (+0.8)	92.6	84.2	75.9	81.3	79.4
RLE [17]	Reg	HRNet-W48	384×288	75.6	33.3	77.9	92.4	84.5	77.1	81.4	80.6
+CCNet		HRNet-W48	384×288	75.6	33.3	78.8 (+0.9)	92.6	85.1	77.0	82.9	80.6
Poseur [21]	Reg	ResNet-50	256×192	33.1	4.6	76.8	92.6	83.7	74.2	81.4	79.7
+CCNet		ResNet-50	256×192	33.1	4.6	77.7 (+0.9)	92.7	84.2	74.9	82.3	79.7
IPR [30]	*const.*	ResNet-50	256×192	34.0	5.5	65.6	88.1	71.8	61.3	70.2	74.9
IPR [30]	Hm	ResNet-50	256×192	34.0	5.5	69.5	88.9	74.6	67.2	74.7	74.9
+CCNet		ResNet-50	256×192	34.1	5.5	70.8 (+1.3)	90.5	78.1	68.1	75.8	74.9

Table 2. Comparisons with state-of-the-art methods on the COCO validation set. The blue color depicts improved value after applying the proposed CCNet. "Hm", "Reg", and "*const.*" represent confidence functions originating from heatmap maximum, direct regression, and constant value, respectively. This table demonstrates that our CCNet considerably improves the AP of all methods.

5. ConfidenceNet

The above analysis shows that changing the score form will improve the ranking and mAP. However, it is not sufficient because the analysis is conditioned on a perfect network that will achieve the optimum value w.r.t. the optimization objective, whereas in practice, different models present different correlations between prediction and $\hat{\sigma}$, which needs extra modeling.

Motivated by this, we propose a calibrated ConfidenceNet (CCNet), which designs a general, efficient, and effective calibration branch for existing pose estimation methods. Specifically, denoting the previous pose network as PredNet $\mathcal{P}$ which outputs keypoint locations, we apply a lightweight network ConfidenceNet $\mathcal{C}$ to predict confidence based on the features in $\mathcal{P}$. For instance, for heatmap-based methods, we detach and utilize the penultimate features after deconvolution layers; for RLE-based methods, we similarly use the features after the Global Average Pooling layer. In this way, it does not require re-training and allows ConfidenceNet $\mathcal{C}$ to get access to the rich features from the PredNet $\mathcal{P}$. Also, the PredNet $\mathcal{P}$ is fixed so mAR will not be affected.

Formally, the ConfidenceNet $\mathcal{C}$ outputs a calibrated confidence $\hat{s}_k \in [0, 1]$ for each keypoint. Apart from keypoint confidence, it also predicts keypoint visibility $\hat{v}_k \in [0, 1]$. This corrects the bias caused by thresholding confidence as visibility in existing practice (equation 5). We observe that accuracy may not be well aligned with visibility (Sec. 6.4). For confidence, a simple yet effective MSE loss is applied to calibrate predictions with ground truth keypoint as,

$$\mathcal{L}_{\text{conf}} = \sum_{k=1}^{K} (\hat{s}_k - s_k)^2. \tag{20}$$

where s_k is the OKS for this keypoint. The experiments find that the loss choices of calibration are robust to others including aleatory and Cross-Entropy. For visibility, we commonly treat it as a binary classification to use a Binary Cross-Entropy (BCE) loss as,

$$\mathcal{L}_{\text{vis}} = -\sum_{k=}^{K} (v_k \log \hat{v}_k + (1 - v_k) \log(1 - \hat{v}_k)). \tag{21}$$

The total loss is then a weighted sum as

$$\mathcal{L} = \mathcal{L}_{\text{conf}} + \lambda \mathcal{L}_{\text{vis}}. \tag{22}$$

Following the OKS form (equation 7), we similarly obtain the instance-level confidence by aggregating the predicted visibility and confidence.

6. Experiments

6.1. Datasets

Datasets & Evaluation Metrics. We evaluate the pose estimation task on three human pose benchmarks, MSCOCO [19], MPII [2], MSCOCO-WholeBody [11]. For the downstream tasks, we evaluate the 3D fitting task on 3DPW.

The **MSCOCO** dataset consists of 250k person instances annotated with 17 keypoints. We follow the standard and evaluate the model with mAP over 10 OKS thresholds. To demonstrate the generalization ability of our method, we also evaluate our method on the **MPII** dataset with Percentage of Correct Keypoints (PCK) and the **MSCOCO-WholeBody** dataset includes face and hand annotations for each human instance, we test our method with the common metric mAP to show its capability on face and hand keypoint detection apart from the body.

For the downstream benchmarks, **3DPW** is a more challenging outdoor benchmark, which provides around 3k SMPL annotations for testing. We follow the convention [14] and use MPJPE, PA-MPJPE, and MVE to measure the quality of the predicted 3D mesh. Additional implementation details and pseudo-code is provided in the Supplementary.

6.2. Comparions with SOTAs

Evaluation on MSCOCO [19]. Since our method is a plug-and-play module after the training of pose models, we evaluate our method on several baselines, including SBL [34], HRNet [29], VitPose [35] for heatmap-based pipelines, RLE [17], IPR [30], Poseur [21] for regression-based pipelines. We perform our method based on the checkpoints released on the official website. We report our results in Tab. 2. Our simple yet effective method universally improves across varying backbones, learning pipelines, and scoring functions, which means it can be applied model-agnostically even when the uncertainty estimation capabilities of different networks may vary from PredNets to PredNets. Also, we argue that pose estimation methods should be aware of confidence estimation and report both improved mAP and mAR. The gap between the two reasonably shows how calibrated the pose model is.

COCO-WholeBody [11] dataset evaluates the task of whole-body pose estimation, which includes not only body keypoints but also face and hand keypoints. The confidence estimation is even less considered. For example, the convention is to assign the whole instance confidence to each part, which is not reasonable when evaluating the AP of the corresponding part. By simply changing the confidence of each part to the aggregation of the predicted part (the third row in Tab. 3) instead of all keypoints (the second row in Tab. 3), the AP is significantly improved up to 5.3 for foot. Applying the proposed CCNet, we further improve the AP on every part and the whole body.

		Body	Foot	Face	Hand	Whole
mAP↑	Whole [21]	67.2	63.6	84.6	58.3	61.0
	Part	68.5	68.9	85.9	62.5	61.0
	+CCNet	69.9	69.2	86.4	62.7	63.3 (+2.3)
mAR↑		72.3	72.9	88.1	65.4	67.2

Table 3. mAP evaluation on the COCO-WholeBody validation set based on Poseur [21]. We base our CCNet on the part confidence and improve the whole body AP by 2.3.

	PCK.5↑	PCK.1↑	mAP↑	mAR↑	AUSE↓	
					PCK.5	PCK.1
RLE [17]	86.2	32.9	75.4	78.8	3.35	1.76
+CCNet			76.6 (+1.2)		2.98	1.49
SBL [17]	88.5	33.9	77.3	80.5	3.90	2.36
+CCNet			77.7 (+0.4)		3.52	1.95

Table 4. The proposed CCNet improves all mAP and AUSE-PCK evaluations on the MPII validation set.

	mAP↑		mAR↑	Pearson Corr↑		AUSE-OKS↓	
	KP			Ins	KP	Ins	KP
RLE [17]	77.9		82.7	0.700	0.637	2.72	5.03
+CCNet	78.7 (+0.8)			0.782	0.636	1.72	4.22
SBL [34]	76.7		82.8	0.643	0.543	2.77	6.47
+CCNet	78.9 (+2.2)			0.718	0.628	2.13	4.11

Table 5. Other confidence quantification evaluations except mAP on the COCO validation set, where "Ins" and "KP" are short for instance and keypoint, respectively.

MPII [2] is a single-person dataset and another commonly used benchmark (Tab. 4). OKS is defined similarly on MPII, where the tight boundary box is used as the area [11, 19]. Though PCK [2], average of visible $\mathcal{I}(\|\boldsymbol{r}_k\| \leq \tau)$, as a simpler metric is less comprehensive than OKS [19, 26, 27], we also try to calibrate the model w.r.t. it to verify the applicability of our method to other metrics and datasets. Since no mAP is defined on PCK, we evaluate with another metric AUSE [10].

Other Confidence Testings. Apart from the most established mAP, (1) Pearson Correlation testing between instance/keypoint OKS and their confidence is also adopted to evaluate the reliability of confidence [3, 7, 17]. (2) We measure Area Under Sparsification Error (AUSE) [6, 10] plotted by gradually removing the most uncertain samples and computing the remaining error ($1-$OKS). (3) As the occlusion patch expands, confidence is expected to shrink along with OKS (Fig. 3). The outperformance of our method

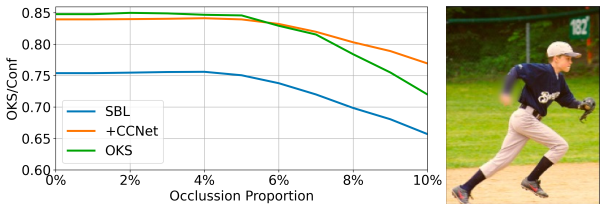

Figure 3. OKS and estimated confidence decrease are well correlated with the occluder size proportional to the input size. The right panel gives an illustration of a Gaussian blur placed on the right wrist.

Method	PA-MPJPE↓	MPJPE↓	MVE↓
SPIN [14]	60.2	102.1	130.6
+SBL [34]	58.8	100.5	128.7
+CCNet	57.8	99.7	127.5

Table 6. 3D errors on 3DPW test set. Results show that.

across all tests as expected shows confidence is really calibrated (Tab. 5). Quantitative visualization of calibrated confidence is provided in the Supp. C.

6.3. Confidence Estimation for Downstream Tasks

Better uncertainty estimation and calibration benefit downstream tasks such as 3D model fitting. We explore the application of confidence calibration in pose estimation.

3D Mesh Fitting. Mesh recovery tasks often employ 2D keypoints to optimize the output 3D pose and shape. Given 2D detection results from the off-the-shelf pose network [34], the 2D reprojection loss usually consists of a weighted sum between the reprojected 2D predictions from the 3D mesh and its corresponding 2D detection results where the weight is exactly the predicted confidence. We follow CLIFF to refine the initial predictions from SPIN. As shown in (Tab. 6), the calibrated 2D pose network better refines the 3D predictions.

6.4. Design Choices & Discussions

Surrogate Losses [1, 5, 12, 16, 25, 36] were proposed to tackle the confidence estimation task. In our exploration, surprisingly, we find these sophisticated methods capture uncertainty no better than MSE. This might be bound by post-hoc confidence estimation of a given PredNet and estimatability [37]. Note that pre-training confidence estimation is generally observed to hurt prediction accuracy as well as lead to a less satisfactory mAP [3, 23]. How to better hybridize the advantages of these two leaves a promising future work.

Input Features. Generally, confidence estimation is based on penultimate features. This promotes the lightweight of the CCNet. To verify that they contain sufficient rich information, we did the comparisons with additional input low-level features of shallow layers, prediction and original keypoint confidence roughly indicating the groundtruth range. A strategy of copying the backbone and fine-tuning similar to Corbière et al. [5], Yu et al. [36], Zhang et al. [39] is also considered. As a result, we find that penultimate feature input is sufficient [37].

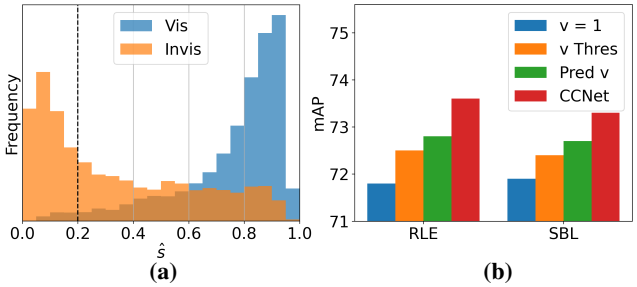

Figure 4. Visibility aggregation ablation. **(a)** Confidence distribution of visible and invisible keypoints. **(b)** Aggregations from different visibilities verify the effectiveness of visibility prediction in keypoint confidence aggregation.

Confidence Aggregation. Existing work [7, 34] empirically set STD of rendered Gaussian heatmaps and a visibility threshold based on confidence estimate. They found the network is sensitive to the choice of the hyperparameters. Furthermore, the model has a different inductive bias from the human; keypoints with low confidence are not necessarily invisible (Fig. 4 (a)). Thus, different common aggregations are studied in Fig. 4(b). The visibility classification strategy shows promising performance without much burden brought. Additional OKS calibration (Eq. (20)) further pushes up the performance.

7. Conclusion

In this paper, we are the first to tackle the pose calibration problem which requires the predicted confidence to be aligned with the accuracy metric OKS. By deriving the expected value of OKS under a general assumption, we theoretically reveal that heuristic ways of getting the confidence in current pose estimation methods cause a misalignment. Although this misalignment can be alleviated by predicting the instance size and adjusting the confidence form, it is not sufficient due to black-box nature of neural networks. Therefore, we propose a Calibrated ConfidenceNet (CCNet) to address all the issues and learn a network-aware branch to align the OKS. Our experiments demonstrate that CCNet is applicable to all pose methods on various datasets and the better calibrated confidence will also help downstream outputs like 3D keypoints.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *NeurIPS*, 2020. 2, 8

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 7, 5

[3] Lennart Bramlage, Michelle Karg, and Cristóbal Curio. Plausible uncertainties for human pose regression. In *ICCV*, 2023. 2, 7, 8, 5

[4] Rongyu Chen, Linlin Yang, and Angela Yao. MHEntropy: Entropy meets multiple hypotheses for pose and shape recovery. In *ICCV*, 2023. 4

[5] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *NeurIPS*, 2019. 8

[6] Gianni Franchi, Xuanlong Yu, Andrei Bursuc, Angel Tena, Rémi Kazmierczak, Séverine Dubuisson, Emanuel Aldea, and David Filliat. MUAD: Multiple uncertainties for autonomous driving, a benchmark for multiple uncertainty types and tasks. In *BMVC*, 2022. 7

[7] Kerui Gu, Linlin Yang, and Angela Yao. Dive deeper into integral pose regression. In *ICLR*, 2021. 5, 7, 8, 3

[8] Kerui Gu, Linlin Yang, and Angela Yao. Removing the bias of integral pose regression. In *ICCV*, 2021. 1

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 2

[10] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *ECCV*, 2018. 7

[11] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 7

[12] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017. 2, 8, 1, 6

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[14] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 7, 8

[15] Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *ICML*, 2022. 3

[16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017. 2, 8, 6

[17] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021. 1, 2, 3, 4, 5, 6, 7

[18] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *CVPR*, 2021. 3

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 3, 7, 5

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. In *SIGGRAPH Asia*, 2015. 2

[21] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *ECCV*, 2022. 1, 2, 3, 6, 7

[22] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 2

[23] Bimsara Pathiraja, Malitha Gunawardhana, and Muhammad Haris Khan. Multiclass confidence and localization calibration for object detection. In *CVPR*, 2023. 2, 8

[24] Paweł A Pierzchlewicz, R James Cotton, Mohammad Bashiri, and Fabian H Sinz. Multi-hypothesis 3D human pose estimation metrics favor miscalibrated distributions. In *arXiv*, 2022. 2, 5

[25] Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. In *NeurIPS*, 2021. 4, 8, 6

[26] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *ICCV*, 2017. 1, 7

[27] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *CVPR*, 2022. 2, 7

[28] Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *ICML*, 2019. 2

[29] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 2, 3, 6, 7

[30] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 3, 6, 7

[31] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 1

[32] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3D human pose estimation with normalizing flows. In *ICCV*, 2021. 2

[33] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *ECCV*, 2020. 3

[34] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[35] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 1, 2, 3, 6, 7

[36] Xuanlong Yu, Gianni Franchi, and Emanuel Aldea. SLURP: Side learning uncertainty for regression problems. In *BMVC*, 2021. 8

[37] Xuanlong Yu, Gianni Franchi, Jindong Gu, and Emanuel Aldea. Discretization-induced Dirichlet posterior for robust uncertainty quantification on regression. In *arXiv*, 2023. 8

[38] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020. 3

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 8

On the Calibration of Human Pose Estimation

Supplementary Material

Figure e. An illustration of visual cues that have similar (underlying/predictive) uncertainty (red circle mean and green dashed circle range) but different per-sample annotation (blue crosses) treated as a sample from the distribution. For instance, the left one is further from the mean than the right one.

The supplementary materials include **A.** Theoretical Understanding, **B.** Implementation Details, and **C.** More Experimental Results, referred in the manuscript.

A. Theoretical Understanding

A.1. Illustration of Setting

Different from 1 image $\mathbf{x}$ corresponding to only 1 pose keypoint $\mathbf{p}$, we consider stochastics caused by annotation error and occlusion ambiguity, *etc*., by a 1-to-many distribution $p(\mathbf{p}|\mathbf{x})$ (L275-276). Specifically, for two inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$ with similar ambiguity $\sigma_1 \approx \sigma_2$, accuracy (*e.g*., OKS) may be different sample-wisely (Figure e), but they are supposed to have similar rankings regardless of uncontrollable and irreducible uncertainty [12]. Think about it from another perspective: if the person is asked to re-annotate the two images, analogous to *re-sampling* of the distribution, the accuracy of the first image has chance of being higher than that of the second. The goal is to achieve the highest mAP in the expected sense of distributions.

A.2. Expected OKS Equation (11)

Proof. It follows a Normal distribution (L321-323); the integral is also tractable to compute as shown below.

$$\mathbb{E}_{\mathbf{p}\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I})}[\text{OKS}] \tag{23}$$

$$= \int_{\mathbf{p}} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|\mathbf{p}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\mathrm{l}^2}\right) \mathbf{dp} \tag{24}$$

$$= \frac{1}{2\pi\sigma^2} \int_{\mathbf{p}} \exp\left(-\frac{\|\mathbf{p}-\boldsymbol{\mu}\|^2}{2\sigma^2} - \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\mathrm{l}^2}\right) \mathbf{dp}. \tag{25}$$

Lemma 1. *In L300 of manuscript, the form is regarded as resemblingly the random variable* $\hat{\mathbf{p}} \sim \mathcal{N}(\boldsymbol{\mu}, (\sigma^2+\mathrm{l}^2)\boldsymbol{I})$. I.e.,

$$\int_{\mathbf{p}} \mathcal{N}(\mathbf{p}|\boldsymbol{\mu}, \sigma^2\boldsymbol{I})\mathcal{N}(\hat{\mathbf{p}}|\mathbf{p}, \mathrm{l}^2\boldsymbol{I})\mathbf{dp} = \mathcal{N}(\hat{\mathbf{p}}|\boldsymbol{\mu}, (\sigma^2+\mathrm{l}^2)\boldsymbol{I}) \tag{26}$$

$$\iff \int_{\mathbf{p}} \exp\left(-\frac{\|\mathbf{p}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\hat{\mathbf{p}}-\mathbf{p}\|^2}{2\mathrm{l}^2}\right) \mathbf{dp} \tag{27}$$

$$= \frac{2\pi\sigma^2\mathrm{l}^2}{(\sigma^2+\mathrm{l}^2)} \exp\left(-\frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2(\sigma^2+\mathrm{l}^2)}\right). \tag{28}$$

Lemma 2. *In another perspective, term within* exp *of Equation (25) can be also arranged w.r.t.* $\mathbf{p}$ *as*

$$-\|\mathcal{D}\mathbf{p} - \overrightarrow{\mathcal{E}}\|^2 - \mathcal{F}, \tag{29}$$

$$\mathcal{D} = \frac{1}{\sqrt{2\frac{\sigma^2\mathrm{l}^2}{\sigma^2+\mathrm{l}^2}}}, \overrightarrow{\mathcal{E}} = \frac{\mathrm{l}^2\boldsymbol{\mu} + \sigma^2\hat{\mathbf{p}}}{\sqrt{2(\mathrm{l}^2+\sigma^2)\mathrm{l}^2\sigma^2}}, \mathcal{F} = \frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2(\sigma^2+\mathrm{l}^2)}. \tag{30}$$

Substituting back into Equation (25) obtains

$$\frac{1}{2\pi\sigma^2} \int_{\mathbf{p}} \exp(-\|\mathcal{D}\mathbf{p} - \overrightarrow{\mathcal{E}}\|^2 - \mathcal{F})\mathbf{dp} \tag{31}$$

$$= \frac{1}{2\pi\sigma^2} \int_{\mathcal{D}\mathbf{p}-\overrightarrow{\mathcal{E}}} \exp(-\|\mathcal{D}\mathbf{p} - \overrightarrow{\mathcal{E}}\|^2) \exp(-\mathcal{F}) \frac{1}{\mathcal{D}} \mathbf{d}\mathcal{D}\mathbf{p} - \overrightarrow{\mathcal{E}} \tag{32}$$

$$= \frac{\exp(-\mathcal{F})}{2\pi\sigma^2\mathcal{D}} \int_{\mathcal{D}\mathbf{p}-\overrightarrow{\mathcal{E}}} \exp(-\|\mathcal{D}\mathbf{p} - \overrightarrow{\mathcal{E}}\|^2) \mathbf{d}\mathcal{D}\mathbf{p} - \overrightarrow{\mathcal{E}} \tag{33}$$

$$= \frac{\exp(-\mathcal{F})}{2\pi\sigma^2\mathcal{D}} 2\pi\frac{1}{2} \tag{34}$$

$$= \frac{\mathrm{l}^2}{\sigma^2+\mathrm{l}^2} \exp\left(-\frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2(\sigma^2+\mathrm{l}^2)}\right), \tag{35}$$

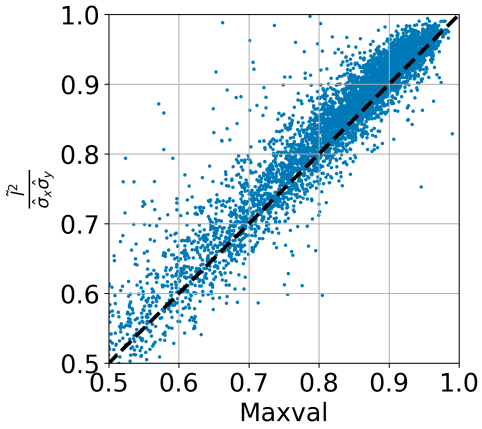

Figure f. Maximum values of the heatmap are almost coincident with our estimated scoring (peak density as Equation (17)), which verifies derivation.

where $\mathcal{D}, \mathcal{F}$ are independent of $\mathbf{p}$ conditional on the image (Equation (33)); Equation 35 is based on

$$\int_{\mathbf{x}} \frac{1}{2\pi\frac{1}{2}} \exp(-\|\mathbf{x}\|^2)\mathbf{dx} = 1. \tag{36}$$

□

A.3. Verification of Detection $\hat{\sigma}^2 = \sigma^2 + \tilde{l}^2$ (L308)

Figure f verifies Equation (17) and model distribution (or heatmap) approximates noisy ground truth distribution instead of the pure one. Following [32], sigmas are estimated by fitting heatmap with Gaussian.

A.4. Optima of Equation (18) NLL

Proof. It is well-established, but we still include it here for the convenience of readers. Formally,

$$\hat{\mathbf{p}}^*, \hat{\sigma}^* = \arg\min_{\hat{\mathbf{p}},\hat{\sigma}} \mathcal{L}_{\text{nll}} = \arg\max_{\hat{\mathbf{p}},\hat{\sigma}} \mathcal{L}_{\text{ll}}, \tag{37}$$

where in more general 2D case (1D in the manuscript for illustration), Log-Likelihood

$$\mathcal{L}_{\text{ll}} = \mathbb{E}_{\mathbf{p}\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I})}\left[\log \frac{1}{2\pi\hat{\sigma}^2}\exp\left(-\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}\right)\right] \tag{38}$$

$$= \mathbb{E}_{\mathbf{p}}\left[-\log 2\pi - \log\hat{\sigma}^2 - \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}\right] \tag{39}$$

$$\stackrel{c}{=} -\mathbb{E}_{\mathbf{p}}\left[\log\hat{\sigma}^2 + \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}\right]. \tag{40}$$

Denote

$$\mathcal{A} = \mathcal{B} + \mathcal{C}, \mathcal{B} = \log\hat{\sigma}^2, \mathcal{C} = \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}. \tag{41}$$

The following is calculated:

$$\frac{\partial\mathcal{B}}{\partial\hat{\mathbf{p}}} = \mathbf{0}, \frac{\partial\mathcal{B}}{\partial\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}, \tag{42}$$

$$\frac{\partial\mathcal{C}}{\partial\hat{\mathbf{p}}} = \frac{1}{2\hat{\sigma}^2}\frac{\partial\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{\partial\hat{\mathbf{p}}} = \frac{1}{2\hat{\sigma}^2}\frac{\partial\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{\partial\mathbf{p}-\hat{\mathbf{p}}}\frac{\partial\mathbf{p}-\hat{\mathbf{p}}}{\partial\hat{\mathbf{p}}} \tag{43}$$

$$= \frac{1}{2\hat{\sigma}^2}2(\mathbf{p}-\hat{\mathbf{p}})^T(-\boldsymbol{I}) = -\frac{\mathbf{p}-\hat{\mathbf{p}}}{\hat{\sigma}^2}, \tag{44}$$

$$\frac{\partial\mathcal{C}}{\partial\hat{\sigma}^2} = \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2}\frac{\partial\frac{1}{\hat{\sigma}^2}}{\partial\hat{\sigma}^2} = \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2}\left(-\frac{1}{\hat{\sigma}^4}\right) \tag{45}$$

$$= -\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^4}. \tag{46}$$

Optimal $\hat{\mathbf{p}}$. Taking derivative of $\mathcal{L}_{\text{ll}}$ w.r.t. $\hat{\mathbf{p}}$ and setting it to 0 give

$$\frac{\partial\mathcal{L}_{\text{ll}}}{\partial\hat{\mathbf{p}}} = \frac{\partial - \mathbb{E}_{\mathbf{p}}[\mathcal{A}]}{\partial\hat{\mathbf{p}}} = -\mathbb{E}_{\mathbf{p}}\left[\frac{\partial\mathcal{A}}{\partial\hat{\mathbf{p}}}\right] \tag{47}$$

$$= -\mathbb{E}_{\mathbf{p}}\left[\frac{\partial\mathcal{B}}{\partial\hat{\mathbf{p}}} + \frac{\partial\mathcal{C}}{\partial\hat{\mathbf{p}}}\right] = -\mathbb{E}_{\mathbf{p}}\left[\mathbf{0} - \frac{\mathbf{p}-\hat{\mathbf{p}}}{\hat{\sigma}^2}\right] \tag{48}$$

$$= \mathbb{E}_{\mathbf{p}}\left[\frac{\mathbf{p}-\hat{\mathbf{p}}}{\hat{\sigma}^2}\right] = \frac{1}{\hat{\sigma}^2}(\mathbb{E}_{\mathbf{p}}[\mathbf{p}] - \hat{\mathbf{p}}) = \frac{1}{\hat{\sigma}^2}(\boldsymbol{\mu} - \hat{\mathbf{p}}) \tag{49}$$

$$= \mathbf{0}. \tag{50}$$

The facts that given an image, $\mathbf{p}$ in expectation is constant w.r.t. $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}, \hat{\sigma}^2$ are constant w.r.t. $\mathbf{p}$ are used in Equations (47) and (49), respectively. Thus, rearrangement gives optima

$$\hat{\mathbf{p}}^* = \boldsymbol{\mu}. \tag{51}$$

Optimal $\hat{\sigma}$. Similarly, we derive derivative of $\mathcal{L}_{\text{ll}}$ w.r.t. $\hat{\sigma}^2$ as

$$\frac{\partial\mathcal{L}_{\text{ll}}}{\partial\hat{\sigma}^2} = \frac{\partial - \mathbb{E}_{\mathbf{p}}[\mathcal{A}]}{\partial\hat{\sigma}^2} = -\mathbb{E}_{\mathbf{p}}\left[\frac{\partial\mathcal{A}}{\partial\hat{\sigma}^2}\right] \tag{52}$$

$$= -\mathbb{E}_{\mathbf{p}}\left[\frac{\partial\mathcal{B}}{\partial\hat{\sigma}^2} + \frac{\partial\mathcal{C}}{\partial\hat{\sigma}^2}\right] = -\mathbb{E}_{\mathbf{p}}\left[\frac{1}{\hat{\sigma}^2} - \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^4}\right] \tag{53}$$

$$= -\frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\mathbb{E}_{\mathbf{p}}[\|\mathbf{p}-\hat{\mathbf{p}}\|^2]. \tag{54}$$

Equation 51 optimal $\hat{\mathbf{p}}^*$ helps simplify it as

$$-\frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\mathbb{E}_{\mathbf{p}}[\|\mathbf{p}-\boldsymbol{\mu}\|^2] = -\frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}2\sigma^2 \tag{55}$$

$$= -\frac{1}{\hat{\sigma}^2} + \frac{\sigma^2}{\hat{\sigma}^4}. \tag{56}$$

For Equation (55) the variance of the Normal distribution is used. Setting it to 0 arrives at

$$\hat{\sigma}^* = \sigma. \tag{57}$$

□

A.5. The Case of Imperfect Prediction $\hat{\mathbf{p}} \neq \boldsymbol{\mu}$

Proof. TL;DR: when prediction is imperfect, confidence will decrease correspondingly.

It is a more general case and will lead to more misalignment to the ideal score (Equation (11)). For instance, the prediction deviation of easy samples is likely to be less than that of hard samples. We derive optimal $\hat{\sigma}$ in this case. It makes sense to some extent for deviation is usually easier to estimate than mean since it only requires to predict a range instead of an exact value. Denote prediction deviation as

$$\hat{\boldsymbol{\delta}} = \hat{\mathbf{p}} - \boldsymbol{\mu}, \hat{\Delta}^2 = \|\hat{\boldsymbol{\delta}}\|^2 \neq 0; \boldsymbol{\delta} = \mathbf{p} - \boldsymbol{\mu}. \tag{58}$$

For **Regression**, Equation (54)$= 0$ tells

$$\hat{\sigma}^{*2} = \frac{1}{2}\mathbb{E}_{\mathbf{p}}[\|\mathbf{p} - \hat{\mathbf{p}}\|^2] = \frac{1}{2}\mathbb{E}_{\mathbf{p}}[\|\mathbf{p} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\mathbf{p}}\|^2] \tag{59}$$

$$= \frac{1}{2}\mathbb{E}_{\mathbf{p}}[\boldsymbol{\delta}^T\boldsymbol{\delta} - 2\boldsymbol{\delta}^T\hat{\boldsymbol{\delta}} + \hat{\boldsymbol{\delta}}^T\hat{\boldsymbol{\delta}}] \tag{60}$$

$$= \frac{1}{2}(\mathbb{E}_{\mathbf{p}}[\|\boldsymbol{\delta}\|^2] - 2\mathbb{E}_{\mathbf{p}}[\boldsymbol{\delta}]^T\hat{\boldsymbol{\delta}} + \hat{\Delta}^2) \tag{61}$$

$$= \frac{1}{2}(2\sigma^2 - 2\mathbf{0}^T\hat{\boldsymbol{\delta}} + \hat{\Delta}^2) = \sigma^2 + \frac{\hat{\Delta}^2}{2}. \tag{62}$$

Equation 61 is based on $\hat{\boldsymbol{\delta}}$ is constant w.r.t. $\mathbf{p}$. The score Equation (19) becomes

$$\hat{\mathrm{s}}_{\text{reg}} = 1 - \hat{\sigma} = 1 - \sqrt{\sigma^2 + \frac{\hat{\Delta}^2}{2}} < 1 - \sigma. \tag{63}$$

For **Detection**, derivation assumes

Proposition 1. *Imperfect (but not bad) heatmap follows [7]*

$$\hat{\mathrm{h}}_{\boldsymbol{m}} = \hat{\mathrm{o}}\exp\left(-\frac{\|\boldsymbol{m} - \hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}\right), \tag{64}$$

where $\hat{\mathrm{o}}$ *is a scaling factor.*

MSE (Equation (14)) is derived as

$$\mathcal{L}_{\text{mse}} \overset{c}{=} \mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}(\hat{\mathrm{h}}_{\boldsymbol{m}} - \mathrm{h}_{\boldsymbol{m}})^2\right]. \tag{65}$$

For each location $\boldsymbol{m}$,

$$\frac{\partial\mathcal{L}}{\partial\hat{\mathrm{h}}} = \frac{\partial(\hat{\mathrm{h}} - \mathrm{h})^2}{\partial\hat{\mathrm{h}}} = 2(\hat{\mathrm{h}} - \mathrm{h}); \tag{66}$$

$$\frac{\partial\hat{\mathrm{h}}}{\partial\hat{\sigma}^2} = \hat{\mathrm{o}}\exp\left(-\frac{\|\boldsymbol{m} - \hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}\right)\frac{\partial - \frac{\|\boldsymbol{m}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}}{\partial\hat{\sigma}^2} = \hat{\mathrm{h}}\frac{\|\boldsymbol{m} - \hat{\mathbf{p}}\|^2}{2\hat{\sigma}^4}, \tag{67}$$

$$\frac{\partial\hat{\mathrm{h}}}{\partial\hat{\mathrm{o}}} = \exp\left(-\frac{\|\boldsymbol{m} - \hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}\right) = \frac{\hat{\mathrm{h}}}{\hat{\mathrm{o}}}. \tag{68}$$

Derivation of Equation (67) uses Equation (46).

Detection's Optimal $\hat{\mathrm{o}}$ (entangling with $\hat{\sigma}^2$). Further,

$$\frac{\partial\mathcal{L}_{\text{mse}}}{\partial\hat{\mathrm{o}}} = \frac{\partial\mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}(\hat{\mathrm{h}}_{\boldsymbol{m}} - \mathrm{h}_{\boldsymbol{m}})^2\right]}{\partial\hat{\mathrm{o}}} = \mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}\frac{\partial\mathcal{L}_{\boldsymbol{m}}}{\partial\hat{\mathrm{o}}}\right] \tag{69}$$

$$= \mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}\frac{\partial\mathcal{L}_{\boldsymbol{m}}}{\partial\hat{\mathrm{h}}_{\boldsymbol{m}}}\frac{\partial\hat{\mathrm{h}}_{\boldsymbol{m}}}{\partial\hat{\mathrm{o}}}\right] \tag{70}$$

$$= \mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}2(\hat{\mathrm{h}}_{\boldsymbol{m}} - \mathrm{h}_{\boldsymbol{m}})\frac{\hat{\mathrm{h}}_{\boldsymbol{m}}}{\hat{\mathrm{o}}}\right] \tag{71}$$

$$= \frac{2}{\hat{\mathrm{o}}}\sum_{\boldsymbol{m}}(\hat{\mathrm{h}}_{\boldsymbol{m}}^2 - \hat{\mathrm{h}}_{\boldsymbol{m}}\mathbb{E}_{\mathbf{p}}[\mathrm{h}_{\boldsymbol{m}}]) = 0. \tag{72}$$

The last step makes use of that given the image, only $\mathrm{h}_{\boldsymbol{m}}$ depends on $\mathbf{p}$.

Denote

$$\hat{\mathrm{h}}_{\boldsymbol{m}}^2 = \hat{\mathrm{o}}^2\mathcal{G}_{\boldsymbol{m}}, \mathcal{G}_{\boldsymbol{m}} = \exp\left(-\frac{\|\boldsymbol{m} - \hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2}\cdot 2\right), \tag{73}$$

$$\hat{\mathrm{h}}_{\boldsymbol{m}}\mathbb{E}_{\mathbf{p}}[\mathrm{h}_{\boldsymbol{m}}] = \hat{\mathrm{o}}\mathcal{H}_{\boldsymbol{m}}, \tag{74}$$

$$\mathcal{H}_{\boldsymbol{m}} = \frac{\tilde{\mathrm{l}}^2}{\tilde{\sigma}^2}\exp\left(-\frac{\|\boldsymbol{m} - \hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2} - \frac{\|\boldsymbol{m} - \boldsymbol{\mu}\|^2}{2\tilde{\sigma}^2}\right). \tag{75}$$

$\mathbb{E}_{\mathbf{p}}[\mathrm{h}_{\boldsymbol{m}}]$ comes from Equation (17), and

$$\tilde{\sigma}^2 \triangleq \sigma^2 + \tilde{\mathrm{l}}^2. \tag{76}$$

Substituting them back to Equation (72), we obtain

$$\frac{2}{\hat{\mathrm{o}}}\sum_{\boldsymbol{m}}(\hat{\mathrm{o}}^2\mathcal{G}_{\boldsymbol{m}} - \hat{\mathrm{o}}\mathcal{H}_{\boldsymbol{m}}) = 0 \tag{77}$$

$$\left(\sum_{\boldsymbol{m}}\mathcal{G}_{\boldsymbol{m}}\right)\hat{\mathrm{o}} - \sum_{\boldsymbol{m}}\mathcal{H}_{\boldsymbol{m}} = 0 \tag{78}$$

$$\hat{\mathrm{o}}^* = \frac{\sum_{\boldsymbol{m}}\mathcal{H}_{\boldsymbol{m}}}{\sum_{\boldsymbol{m}}\mathcal{G}_{\boldsymbol{m}}}. \tag{79}$$

Consider the limit as $\Delta\boldsymbol{m} \to \mathbf{0}$ and almost full support of nonnegligible $\mathcal{H}_{\boldsymbol{m}}, \mathcal{G}_{\boldsymbol{m}}$ is within heatmap –

$$\Delta\boldsymbol{m}\sum_{\boldsymbol{m}}\mathcal{G}_{\boldsymbol{m}} \to \int_{\boldsymbol{m}}\mathcal{G}_{\boldsymbol{m}}\mathrm{d}\boldsymbol{m} \tag{80}$$

$$= \int_{\boldsymbol{m}}\exp\left(-\frac{\|\boldsymbol{m} - \hat{\mathbf{p}}\|^2}{2\left(\frac{\hat{\sigma}}{\sqrt{2}}\right)^2}\right)\mathrm{d}\boldsymbol{m} = \pi\hat{\sigma}^2, \tag{81}$$

$$\Delta\boldsymbol{m}\sum_{\boldsymbol{m}}\mathcal{H}_{\boldsymbol{m}} \tag{82}$$

$$\to \int_{\boldsymbol{m}}\frac{\tilde{\mathrm{l}}^2}{\tilde{\sigma}^2}\exp\left(-\frac{\|\boldsymbol{m} - \hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2} - \frac{\|\boldsymbol{m} - \boldsymbol{\mu}\|^2}{2\tilde{\sigma}^2}\right)\mathrm{d}\boldsymbol{m}. \tag{83}$$

Denoting

$$\bar{\sigma}^2 \triangleq \tilde{\sigma}^2 + \hat{\sigma}^2, \tag{84}$$

with Lemma 1, Equation (83) is calculated as

$$\frac{\tilde{\mathrm{l}}^2}{\tilde{\sigma}^2}\frac{2\pi\tilde{\sigma}^2\hat{\sigma}^2}{\bar{\sigma}^2}\exp\left(-\frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2\bar{\sigma}^2}\right) \tag{85}$$

$$=\frac{2\pi\tilde{\mathrm{l}}^2\hat{\sigma}^2}{\bar{\sigma}^2}\exp\left(-\frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2\bar{\sigma}^2}\right), \tag{86}$$

$$\hat{\mathrm{o}}^* \approx \frac{\frac{2\pi\tilde{\mathrm{l}}^2\hat{\sigma}^2}{\bar{\sigma}^2}\exp\left(-\frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2\bar{\sigma}^2}\right)}{\pi\hat{\sigma}^2} = \frac{2\tilde{\mathrm{l}}^2}{\bar{\sigma}^2}\exp\left(-\frac{\hat{\Delta}^2}{2\bar{\sigma}^2}\right). \tag{87}$$

Remarks of $\hat{\mathrm{o}}, \hat{\sigma}^2$. We can further compute

$$\frac{\partial\hat{\mathrm{o}}^*}{\partial\hat{\sigma}^2} = \frac{\partial\hat{\mathrm{o}}^*}{\partial\bar{\sigma}^2} = -\frac{2\tilde{\mathrm{l}}^2}{\bar{\sigma}^4}\exp + \frac{2\tilde{\mathrm{l}}^2}{\bar{\sigma}^2}\exp\cdot\frac{\hat{\Delta}^2}{2\bar{\sigma}^4} = 0 \tag{88}$$

$$\text{root } \bar{\sigma}^2 = \frac{\hat{\Delta}^2}{2}. \tag{89}$$

Since the derivative is monotonical w.r.t. $\hat{\sigma}^2$, it is concluded that when $\hat{\sigma}^2 > \frac{\hat{\Delta}}{2} - \tilde{\sigma}^2$, the scale factor $\hat{\mathrm{o}}^*$ decreases with $\hat{\sigma}^2$ (; increases, otherwise).

For **Detection's Optimal** $\hat{\sigma}$,

$$\frac{\partial\mathcal{L}_{\text{mse}}}{\partial\hat{\sigma}^2} = \frac{\partial\mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}(\hat{\mathrm{h}}_{\boldsymbol{m}}-\mathrm{h}_{\boldsymbol{m}})^2\right]}{\partial\hat{\sigma}^2} = \mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}\frac{\partial\mathcal{L}_{\boldsymbol{m}}}{\partial\hat{\sigma}^2}\right] \tag{90}$$

$$= \mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}\frac{\partial\mathcal{L}_{\boldsymbol{m}}}{\partial\hat{\mathrm{h}}_{\boldsymbol{m}}}\frac{\partial\hat{\mathrm{h}}_{\boldsymbol{m}}}{\partial\hat{\sigma}^2}\right] \tag{91}$$

$$= \mathbb{E}_{\mathbf{p}}\left[\sum_{\boldsymbol{m}}2(\hat{\mathrm{h}}_{\boldsymbol{m}}-\mathrm{h}_{\boldsymbol{m}})\frac{\hat{\mathrm{h}}_{\boldsymbol{m}}\|\boldsymbol{m}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^4}\right] \tag{92}$$

$$= \frac{1}{\hat{\sigma}^4}\sum_{\boldsymbol{m}}(\hat{\mathrm{h}}_{\boldsymbol{m}}^2\|\boldsymbol{m}-\hat{\mathbf{p}}\|^2 - \hat{\mathrm{h}}_{\boldsymbol{m}}\mathbb{E}_{\mathbf{p}}[\mathrm{h}_{\boldsymbol{m}}]\|\boldsymbol{m}-\hat{\mathbf{p}}\|^2) \tag{93}$$

$$= 0. \tag{94}$$

Similarly, we introduce $\Delta\boldsymbol{m}$ as Equation (80), and the first term becomes

$$\sum_{\boldsymbol{m}}\hat{\mathrm{h}}_{\boldsymbol{m}}^2\|\boldsymbol{m}-\hat{\mathbf{p}}\|^2\Delta\boldsymbol{m} \to \hat{\mathrm{o}}^2\pi\hat{\sigma}^2\mathbb{E}_{\mathbf{m}\sim g}[\|\mathbf{m}-\hat{\mathbf{p}}\|^2] \tag{95}$$

$$= \hat{\mathrm{o}}^2\pi\hat{\sigma}^2\hat{\sigma}^2 = \pi\hat{\mathrm{o}}^2\hat{\sigma}^4, \tag{96}$$

as $g(\mathbf{m}) = \mathcal{N}(\hat{\mathbf{p}}, \frac{\hat{\sigma}^2}{2}\boldsymbol{I})$.

Following Lemma 2, $\mathcal{H}$ can also be expressed as a Normal w.r.t. $\mathbf{m}$ for

$$\exp\left(-\frac{\|\boldsymbol{m}-\hat{\mathbf{p}}\|^2}{2\hat{\sigma}^2} - \frac{\|\boldsymbol{m}-\boldsymbol{\mu}\|^2}{2\tilde{\sigma}^2}\right) = \mathcal{K}2\pi\mathcal{J}\mathcal{N}(\overrightarrow{\mathcal{I}}, \mathcal{J}\boldsymbol{I}), \tag{97}$$

$$\overrightarrow{\mathcal{I}} = \frac{\tilde{\sigma}^2\hat{\mathbf{p}}+\hat{\sigma}^2\boldsymbol{\mu}}{\bar{\sigma}^2}, \mathcal{J} = \frac{\tilde{\sigma}^2\hat{\sigma}^2}{\bar{\sigma}^2}, \mathcal{K} = \exp\left(-\frac{\hat{\Delta}^2}{2\bar{\sigma}^2}\right). \tag{98}$$

Thus,

$$\sum_{\boldsymbol{m}}\hat{\mathrm{h}}_{\boldsymbol{m}}\mathbb{E}_{\mathbf{p}}[\mathrm{h}_{\boldsymbol{m}}]\|\boldsymbol{m}-\hat{\mathbf{p}}\|^2\Delta\boldsymbol{m} \tag{99}$$

$$\to \left(\hat{\mathrm{o}}\frac{\tilde{\mathrm{l}}^2}{\tilde{\sigma}^2}\right)2\pi\mathcal{J}\mathcal{K}\mathbb{E}_{\mathbf{m}\sim h}[\|\mathbf{m}-\hat{\mathbf{p}}\|^2] \tag{100}$$

$$\overset{c}{=}\mathbb{E}[\|\mathbf{m}-\overrightarrow{\mathcal{I}}+\overrightarrow{\mathcal{I}}-\hat{\mathbf{p}}\|^2] \tag{101}$$

$$=\mathbb{E}[\|\mathbf{m}-\overrightarrow{\mathcal{I}}\|^2] + \mathbb{E}[\|\hat{\mathbf{p}}-\overrightarrow{\mathcal{I}}\|^2] = 2\frac{\tilde{\sigma}^2\hat{\sigma}^2}{\bar{\sigma}^2} + \frac{\hat{\sigma}^4\hat{\Delta}^2}{\bar{\sigma}^4} \tag{102}$$

$$\iff \textit{Equation (99)} = 2\pi\hat{\mathrm{o}}\tilde{\mathrm{l}}^2\hat{\sigma}^4\left(\frac{2\tilde{\sigma}^2}{\bar{\sigma}^4}+\frac{\hat{\sigma}^2\hat{\Delta}^2}{\bar{\sigma}^6}\right)\mathcal{K}, \tag{103}$$

where $h(\mathbf{m}) = \mathcal{N}(\overrightarrow{\mathcal{I}}, \mathcal{J}\boldsymbol{I})$.

Therefore, substituting Equations (96) and (103) back into Equation (93) gives

$$\pi\hat{\mathrm{o}}^2 - 2\pi\hat{\mathrm{o}}\tilde{\mathrm{l}}^2\left(\frac{2\tilde{\sigma}^2}{\bar{\sigma}^4}+\frac{\hat{\sigma}^2\hat{\Delta}^2}{\bar{\sigma}^6}\right)\mathcal{K} = 0 \tag{104}$$

$$\hat{\mathrm{o}} = 2\tilde{\mathrm{l}}^2\frac{2\tilde{\sigma}^2\bar{\sigma}^2+\hat{\sigma}^2\hat{\Delta}^2}{\bar{\sigma}^6}\exp\left(-\frac{\hat{\Delta}^2}{2\bar{\sigma}^2}\right). \tag{105}$$

Combining Equations (87) and (105) gets

$$\frac{1}{\bar{\sigma}^2} = \frac{2\tilde{\sigma}^2\bar{\sigma}^2+\hat{\sigma}^2\hat{\Delta}^2}{\bar{\sigma}^6} \tag{106}$$

$$(\tilde{\sigma}^2+\hat{\sigma}^2)^2 - 2\tilde{\sigma}^2(\tilde{\sigma}^2+\hat{\sigma}^2) - \hat{\Delta}^2\hat{\sigma}^2 = 0 \tag{107}$$

$$\hat{\sigma}^4 - \hat{\Delta}^2\hat{\sigma}^2 - \tilde{\sigma}^4 = 0 \tag{108}$$

$$\hat{\sigma}^{*2} = \sqrt{\tilde{\sigma}^4+\frac{\hat{\Delta}^4}{4}} + \frac{\hat{\Delta}^2}{2}. \tag{109}$$

The score is unnormalized density at $\hat{\mathbf{p}}$ Equation (64) as

$$\hat{\mathrm{s}}_{\text{det}} = \hat{\mathrm{o}}^* = \frac{2\tilde{\mathrm{l}}^2}{\tilde{\sigma}^2+\sqrt{\tilde{\sigma}^4+\frac{\hat{\Delta}^4}{4}}+\frac{\hat{\Delta}^2}{2}} < \frac{\tilde{\mathrm{l}}^2}{\tilde{\sigma}^2}. \tag{110}$$

□

A.6. A Misspecification Case of $p(\mathbf{p}'|\hat{\mathbf{p}}) \neq \mathcal{N}$

Proof. Model misspecification is a common problem in practice. Here, an example of Laplace predictive distribution $\mathbf{p}' \sim$ Laplace different from underlying Normal noise distribution is shown. As a result, it still has a perfect prediction $\hat{\mathbf{p}}^* = \boldsymbol{\mu}$, but the score deviates from the ideal score (Equation (11)).

With Laplace model, Equation (18) is in form of

$$\mathcal{L}_{\text{lap}} = \mathbb{E}_{\mathbf{p}\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2)}\left[\log \frac{1}{4\hat{\mathrm{b}}^2}\exp\left(-\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|_1}{\hat{\mathrm{b}}}\right)\right] \tag{111}$$

$$\stackrel{c}{=} -\mathbb{E}_{\mathbf{p}}\left[2\log\hat{\mathrm{b}} + \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|_1}{\hat{\mathrm{b}}}\right]. \tag{112}$$

With

$$\frac{\partial 2\log\hat{\mathrm{b}}}{\partial\hat{\mathbf{p}}} = \mathbf{0}, \frac{\partial 2\log\hat{\mathrm{b}}}{\partial\hat{\mathrm{b}}} = \frac{2}{\hat{\mathrm{b}}}, \tag{113}$$

$$\frac{\partial\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|_1}{\hat{\mathrm{b}}}}{\partial\hat{\mathbf{p}}} = -\frac{\operatorname{sgn}(\mathbf{p}-\hat{\mathbf{p}})}{\hat{\mathrm{b}}}, \frac{\partial\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|_1}{\hat{\mathrm{b}}}}{\partial\hat{\mathrm{b}}} = -\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|_1}{\hat{\mathrm{b}}^2}, \tag{114}$$

derivative is computed and set to 0 as

$$\frac{\partial\mathcal{L}_{\text{lap}}}{\partial\hat{\mathbf{p}}} = -\mathbb{E}_{\mathbf{p}}\left[\frac{\partial 2\log\hat{\mathrm{b}}}{\partial\hat{\mathbf{p}}} + \frac{\partial\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|_1}{\hat{\mathrm{b}}}}{\partial\hat{\mathbf{p}}}\right] \tag{115}$$

$$= -\mathbb{E}_{\mathbf{p}}\left[\mathbf{0} - \frac{\operatorname{sgn}(\mathbf{p}-\hat{\mathbf{p}})}{\hat{\mathrm{b}}}\right] = \mathbb{E}_{\mathbf{p}}\left[\frac{\operatorname{sgn}(\mathbf{p}-\hat{\mathbf{p}})}{\hat{\mathrm{b}}}\right] \tag{116}$$

$$= \frac{1}{\hat{\mathrm{b}}}\mathbb{E}_{\mathbf{p}}[\operatorname{sgn}(\mathbf{p}-\hat{\mathbf{p}})] = \mathbf{0}. \tag{117}$$

According to the symmetry of Normal distributions,

$$\hat{\mathbf{p}}^* = \mathbb{E}_{\mathbf{p}}[\mathbf{p}] = \boldsymbol{\mu}. \tag{118}$$

Besides,

$$\frac{\partial\mathcal{L}_{\text{lap}}}{\partial\hat{\mathrm{b}}} = -\mathbb{E}_{\mathbf{p}}\left[\frac{2}{\hat{\mathrm{b}}} - \frac{\|\mathbf{p}-\hat{\mathbf{p}}\|_1}{\hat{\mathrm{b}}^2}\right] = -\frac{2}{\hat{\mathrm{b}}} + \frac{\mathbb{E}_{\mathbf{p}}[\|\mathbf{p}-\hat{\mathbf{p}}\|_1]}{\hat{\mathrm{b}}^2} = 0 \tag{119}$$

$$\Rightarrow \hat{\mathrm{b}}^* = \frac{\mathbb{E}_{\mathbf{p}}[\|\mathbf{p}-\boldsymbol{\mu}\|_1]}{2} = \sqrt{\frac{2}{\pi}}\sigma, \tag{120}$$

where the expectation is calculated according to $\mathbf{p} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$.

The score in this case is

$$\hat{\mathrm{s}}_{\text{reg}} = 1 - \hat{\mathrm{b}} = 1 - \sqrt{\frac{2}{\pi}}\sigma > 1 - \sigma. \tag{121}$$

Results may differ from case to case. □

Algorithm 1 CCNet Pseudocode, PyTorch-like

```
enc, predhead = freeze(prednet) # the locks in
 Fig. 2

def forward(x):
    f = enc(x) # penultimate features
    phat = predhead(f)
    shat, vhat = ccnet(f)
    return phat, [shat, vhat]

def train_step(data, kp_metric=kp_oks, cal_loss=mse,
      w_vis=2e-2):
    x, p, l, v = data
    phat, [shat, vhat] = forward(x)
    s = kp_metric(phat, p, l)

    loss_kp_conf = cal_loss(shat, s, weight=v) #
        calibration loss in Eq. 20
    loss_vis = bce(vhat, v) # Eq. 21
    loss = loss_kp_oks + w_vis * loss_vis # Eq. 22
    ...
```

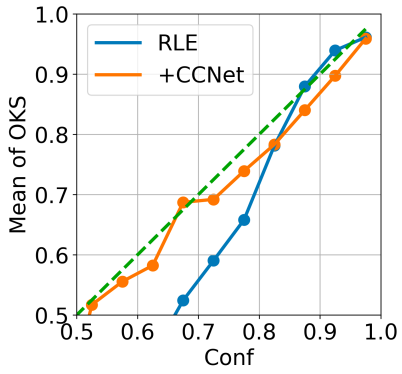

Figure g. Calibration plot. Estimated confidence well reflects the expected OKS value after pose calibration in our context.

B. Implementation Details

Pseudocode is attached in Algorithm 1, facilitating reproducibility for the community.

Training cost is minor – one Fully Connected layer is trained for a few epochs (usually 2 already give good results). Adam [13] optimizer is used with step Learning Rate decay. Ground truth bounding boxes are provided as input to top-down pose estimation methods.

OKS on MPII [2]. Since annotated keypoint sets are different between COCO [19] and MPII [2] but images are similar, per-keypoint falloff coefficients of the neighboring hip, shoulder, and nose are applied to that of the pelvis, thorax, upper neck, and head top, respectively.

Pose Calibration 's variants in Bramlage et al. [3], Pierzchlewicz et al. [24] are mainly based on keypoint EPE instead of instance OKS for computability and also do not focus on mAP (Figure g).

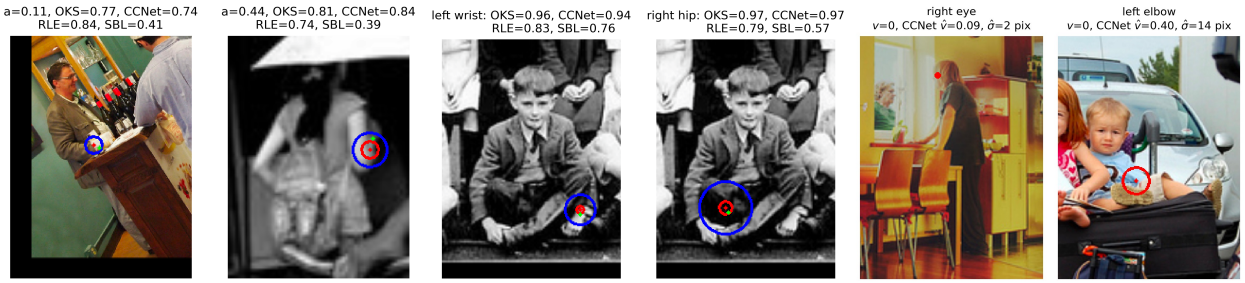

Figure h. Visualizations of **(a)** area, **(b)** per-keypoint falloff (*e.g.* hip's>wrist's), and **(c)** visibility effects to OKS, respectively. Red dot and circle represent predicted keypoints and sigma confidence; green dot indicates ground truth keypoint location; blue circle depicts OKS l range.

	RLE [17]	Alea [12]	CE	DeepEns [16]	SOAP [25]
mAP	72.2	73.4	73.5	73.5	73.4

Table g. Different loss study other than MSE.

C. More Experimental Results

Visualizations (Figure h) show effects of area, per-keypoint falloff, and visibility, respectively. Our CCNet better aligns with OKS.

Design Choices (Sec. 6.4). Different losses including Bayesian weight posterior [16] and surrogate optimization [25], perform similarly well (Table g).