On the Calibration of Human Pose Estimation

Kerui Gu [*1] **Rongyu Chen** [*1] **Xuanlong Yu** [2] **Angela Yao** [1]

Abstract

2D human pose estimation predicts keypoint locations and the corresponding confidence. Calibration-wise, the confidence should be aligned with the pose accuracy. Yet existing pose estimation methods tend to estimate confidence with heuristics such as the maximum value of heatmaps. This work shows, through theoretical analysis and empirical verification, a calibration gap in current pose estimation frameworks. Our derivations directly lead to closed-form adjustments in the confidence based on additionally inferred instance size and visibility. Given the black-box nature of deep neural networks, however, it is not possible to close the gap with only closed-form adjustments. We go one step further and propose a Calibrated ConfidenceNet (CCNet) to explicitly learn network-specific adjustments with a confidence prediction branch. The proposed CCNet, as a lightweight post-hoc addition, improves the calibration of standard off-the-shelf pose estimation frameworks. The project page is at https://comp.nus.edu.sg/~keruigu/calibrate_pose/project.html.

1. Introduction

2D human pose estimation (HPE) methods typically predict keypoint locations and corresponding confidences. The progress in developing such methods is primarily centred on improving keypoint location accuracy (Xu et al., 2022; Mao et al., 2022). The confidence, on the other hand, is estimated in an ad-hoc manner and based on heuristics such as taking the maximum value of the keypoint heatmap (Xiao et al., 2018; Sun et al., 2019) or variance of predictive distribution (Li et al., 2021a).

[*]Equal contribution [1]School of Computing, National University of Singapore [2]U2IS, ENSTA Paris, IP Paris. Correspondence to: Kerui Gu <keruigu@comp.nus.edu.sg>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

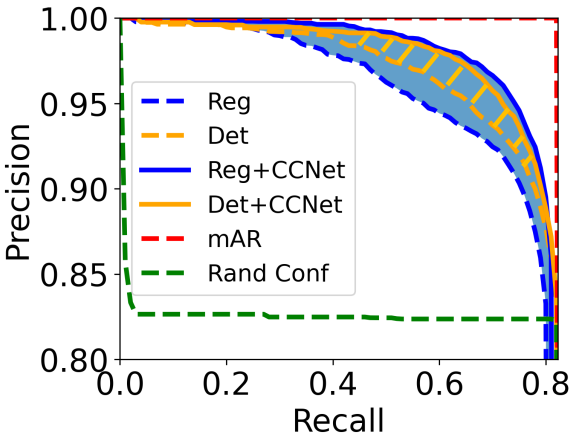


Figure 1. Adding our CCNet to detection- (Det) and regression- (Reg) based pose estimation improves confidence estimation. The area under the Precision-Recall curve measures the quality of the confidence estimate. The orange striped and blue shaded areas denote the improvement. Random assignment of confidence (Rand Conf) has a terrible calibration while mAR serve as the upper bound confidence estimation.

Having well-calibrated confidences that are aligned with the pose accuracy is important for applications that require pose estimation. In these applications, the pose can be used either as an input to a downstream task, such as 3D mesh recovery (Kolotouros et al., 2019; Li et al., 2022) or directly for reasoning and decision-making, such as robotics or autonomous driving (Abdar et al., 2021). Being able to rely on the confidence is not only useful, *e.g.* for discarding low-confidence outputs, but also safety-critical for human-machine interactions.

How well-calibrated are current pose estimation systems, and do their confidences align with the actual pose accuracy? One line of work rooted in uncertainty (Bramlage et al., 2023; Pierzchlewicz et al., 2022) introduces distribution modelling and retrains the pose-estimation network. The resulting network has more reliable confidence, but it comes at the expense of pose accuracy. Furthermore, confidence is evaluated from the perspective of distribution calibration. Such an approach ignores the alignment of confidence to the accuracy and thus may not serve as helpful indicators (Kuleshov & Deshpande, 2022).

To answer the above question, we first analyze the expectation of the predicted confidence versus the ideal confidence based on the pose accuracy metric. For example, in the popular benchmark MSCOCO (Lin et al., 2014), the accuracy is measured by object keypoint similarity (OKS). Yet the confidence, as heuristics, is wrongly formulated and is therefore systematically miscalibrated. Our analysis bridges some of the calibration gaps simply by changing the closed-form expression for confidence, *e.g*. by accounting for the instance's scale and keypoint visibilities.

However, only correcting the formulation of the confidence term is insufficient. In practice, network predictions vary depending on different backbones and datasets. In this case, we can make further network-specific adjustments to better calibrate the confidence. To that end, we propose a simple yet effective Calibrated ConfidenceNet (CCNet) to complement pose-estimation frameworks. CCNet, as an ad-hoc add-on, is framework agnostic and applicable to any existing pose estimation methods. Using the penultimate features of the original pose models, it explicitly estimates a score and visibility measure. The outputs are supervised with ground truth visibility and its OKS to directly link the predicted confidence and address the miscalibration of that network. With only a few epochs for training and minimal additional parameters, the pose estimation framework improves in calibration and mAP.

Summarizing our contributions,

- We are the first to provide a principled understanding of the calibration of 2D pose estimation. Pose calibration has been overlooked in the literature but has importance for downstream applications and safety-critical decision making.
- We mathematically formulate the ideal form of pose confidence and reveal its mismatch to the practical confidence form of current pose estimation methods. A simple solution is provided to verify and correct the misalignment.
- We propose a simple but effective method to explicitly model the calibration with minimal addition of parameters and training time. Experiments show that adding the calibration branch gives a significant improvement on the primary metric mAP and also benefits the downstream tasks.

2. Related Work

Pose Estimation. Past works in 2D top-down-based pose estimation mainly focused on improving accuracy. Few works give some heuristic or empirical understanding on the pose confidence. (Papandreou et al., 2017) proposed an effective re-scoring strategy based on the detected bounding box , which is applied in several top-down-based methods (Xiao et al., 2018; Li et al., 2021a). PETR (Shi et al., 2022) empirically found that changing the matching objective to be OKS-based improves the average precision under the same average recall, indicating a better ranking over the samples. Poseur (Mao et al., 2022), which follows a regression paradigm, noted that previous regression scoring is heuristic. They rescore the confidence into a likelihood based on the detection scores. Although there exist several works that try to change the form of confidence, they remain heuristic and purely empirical. Our paper gives a theoretical understanding of the confidence for both heatmap- and regression-based methods; it also analyzes and corrects the confidence to be better calibrated with the pose accuracy.

Confidence Estimation. Confidence estimates are essential in real-world applications (Kendall & Gal, 2017; Lakshminarayanan et al., 2017; Amini et al., 2020). Guo et al. (2017) reveal that the softmax output of modern neural networks, which typically is interpreted as a categorical distribution in classification, is poorly calibrated. The outputs do not faithfully reflect the actual accuracy and tend to be overconfident. Guo et al. (2017) study post-hoc confidence calibration, which can be plugged into any trained model. Similarly, recent work (Pathiraja et al., 2023) introduces train-time calibration to object detection.

For regression tasks, there are no agreed-upon conventions. The quantile-based definition is common (Song et al., 2019), but the evaluation is nontrivial in high dimensions. Other methods directly improve and evaluate probability distribution models (Kendall & Gal, 2017; Amini et al., 2020). Finally, (Xiao et al., 2018; Yu et al., 2021; Mukhoti et al., 2023) estimate prediction errors or related metrics instead of probability values and thus adopt ranking-based evaluation such as Area Under Curves (Ilg et al., 2018; Franchi et al., 2022). However, there are few works studying calibration in pose estimation (Bramlage et al., 2023; Pierzchlewicz et al., 2022). We argue that pose confidence is useful and informative only when it aligns well with actual accuracy; otherwise, it is not beneficial (Kuleshov & Deshpande, 2022). To this end, our work mainly studies more efficient Auxillary Confidence Regression (Corbière et al., 2019; Yu et al., 2021; Shen et al., 2023) and evaluates calibration with comprehensive metrics including AP (AUPRC).

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.

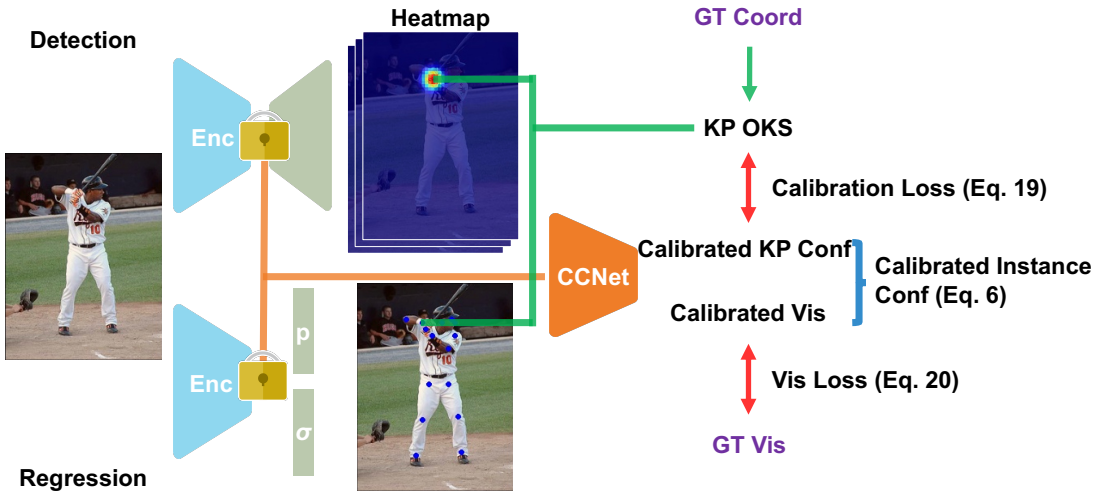


Figure 2. We introduce CCNet, a lightweight post-hoc addition to off-the-shelf pose estimation methods. CCNet directly estimates better-calibrated confidences from latent pose representations without modifying backbone parameters. The green arrows depict the accuracy (*e.g.*, OKS and PCK) calculation flow used during training. Using COCO OKS as an example, CCNet's predicted keypoint confidence and visibility are supervised by the confidence calibration loss and visibility loss (indicated by red arrows), respectively. The final instance confidence is obtained through a weighted aggregation.

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 (Xiao et al., 2018; Sun et al., 2019; Xu et al., 2022) estimate K heatmaps $\hat{\boldsymbol{H}} \in \mathbb{R}^{H\times W}$ to represent pseudo-likelihoods, *i.e.* unnormalized probabilities of each pixel being the k-th keypoint (see example in Fig. 2). The heatmap $\hat{\boldsymbol{H}}_k$ can be decoded into the joint coordinate $\hat{\boldsymbol{p}}_k$ and joint 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 (Zhang et al., 2020; Gu et al., 2021b) in place of the $\arg\max$.

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

$$\mathcal{L}_{\text{det}} = \sum_{k=1}^{K} \text{MSE}(\hat{\boldsymbol{H}}_k, \boldsymbol{H}_k). \tag{2}$$

Typically, the ground truth heatmap $\boldsymbol{H}_k$ is constructed as 2D Gaussian, with the mean at the ground truth keypoint location and a fixed standard deviation $\tilde{l}$.

Regression methods directly regress either deterministic coordinates of the keypoints or likelihood distributions of the coordinates. We focus on the state-of-the-art RLE regression (Li et al., 2021a; Mao et al., 2022), which models the likelihood as a distribution parameterized by mean and standard deviation parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}$. The keypoint prediction and its confidence are given as

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

The loss is 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 loss between $\hat{\mathbf{p}}_k$ and $\mathbf{p}_k$, along with some regularization term such as $\log \hat{\sigma}_k^2$.

Other regression-based methods use heatmap maximum (Wei et al., 2020) or keypoint classification confidence from another head (Li et al., 2021b) or simply fill the confidence as 1 (Sun et al., 2018).

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 and $\tau_{\hat{s}}$ is a manually defined threshold. The keypoint-to-instance aggregation $\text{agg}(\cdot)$ is an averaging function that selects only keypoints above the threshold $\tau_{\hat{s}}$, with $\hat{s}_k > \tau_{\hat{s}}$.

3.2. Evaluating Pose Models

Several metrics are used for evaluating keypoint accuracy. One example is the End-Point Error (EPE), defined as the

mean Euclidean distance between the estimated and ground truth keypoint. EPE, measured in pixels, cannot account for a person's scale. Another metric is the Percentage of Correct Keypoint (PCK), which tallies the fraction of keypoints within varying thresholds. PCK is normalized with respect to head size and factors in scale, but does not distinguish between different types of keypoints.

A more sophisticated evaluation measure for keypoint accuracy is Object Keypoint Similarity (OKS) (Lin et al., 2014). 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:

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

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

Above, 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 in the exponential envelope accounts for differences in scale across the different body joints and overall pose area. A person instance estimate is regarded as correct (positive) if its OKS exceeds some threshold.

3.3. mAP & mAR

Based on OKS, a ranking-independent metric mean Average Recall (mAR) and a ranking-dependent metric mean Average Precision (mAP) can be established to evaluate a given pose model. The mAR purely evaluates the pose accuracy of the model while the mAP considers confidence as well. With the same pose accuracy, a higher similarity between the rankings of the confidence and OKS brings higher mAP. Note that the formulations of mAP and mAR are the same as the Area Under (maximum) Precision-Recall Curve (AUPRC) and the Area Under the Receiver Operating Characteristic (AUROC) used in conventional classification (Qi et al., 2021). We give mathematical formulations of mAR and mAP as follows.

Over a dataset with N samples, we can tabulate the mean Average Recall (mAR) and mean Average Precision (mAP) over T thresholds $\{\tau_t\}$ as

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

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 Precision (mAP). The mAP is defined as

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

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.

4. An Analysis on Pose Calibration

4.1. Problem Formulation & Assumptions

For a well-calibrated pose model, the predicted pose confidence should follow the same ranking as the accuracy. While there are several accuracy measures for the pose, as outlined in Sec. 3.2, we center our analysis on OKS, as it is the most comprehensive, and its corresponding mAP metric.

For the analysis, we formulate the expected OKS and predicted confidence of both heatmap- and RLE-based methods from a statistical perspective, following two standard assumptions (Xiao et al., 2018; Li et al., 2021a). First, the K keypoints of a person are conditionally independent given the image. For clarity, we drop the k subscript in this section. Secondly, we assume that the ground truth location of each keypoint in an image follows a Gaussian distribution $\mathbf{p} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ (Li et al., 2021a; Chen et al., 2023), where $\boldsymbol{\mu}$ specifies the true underlying location. For simplicity, we consider a 2D 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.

4.2. Expected OKS

Assuming a Gaussian distribution parameterized by $(\boldsymbol{\mu}, \sigma^2)$ for the ground truth pose $\mathbf{p}$, the expected value of the OKS distribution for an estimated pose $\hat{\mathbf{p}}$ is given by

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

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

Note the above equation is a function of $\{\boldsymbol{\mu}, \sigma, l, \hat{\mathbf{p}}\}$, depending on the ground truth Gaussian and l, the exponential envelope fall-off rate given in Eq. 7. When a network is perfectly trained, $\hat{\mathbf{p}}$ will approach $\boldsymbol{\mu}$ and the exponential term simplifies to 1, which results in the following confidence

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

[1] Accounts for occluded and unoccluded keypoints.

4.3. Ad-Hoc 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

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

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

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

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

The resulting optimal spatial heatmap is $\hat{\boldsymbol{H}} = \{\hat{h}\} \overset{c}{\approx} \mathcal{N}(\boldsymbol{\mu}, \hat{\sigma}^2\boldsymbol{I})$, which approximates the synthesized ground truth heatmap, with $\hat{\sigma}^2 = \sigma^2 + \tilde{l}^2$ (see Appendix for a similar derivation for the case when $\mathbf{p}$ is not centered at $\boldsymbol{\mu}$). 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 (Gu et al., 2021a). See Appendix Sec. A for proof.

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

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

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

The two expected values from Eq. 12 and Eq. 17 are different for the same input, *i.e.* a same location $\boldsymbol{\mu}$. This difference arises because $\tilde{l}$ is constant, while l changes depending on the (person) instance size and keypoint. For example, a larger person leads to an underestimation of the OKS.

RLE-based regression methods are learned by minimizing a negative log-likelihood over the predicted distribution as shown in Eq. 4. For simplicity, we consider a normal distribution $\mathbf{p}' \sim \mathcal{N}(\hat{\boldsymbol{p}}, \hat{\sigma}^2\boldsymbol{I})$, though alternative distributions such as a Laplace or Normalizing Flow lead to the same conclusions. After training, we show that the predicted distribution approximates the optimal, $\hat{\mathbf{p}} \approx \boldsymbol{\mu}, \hat{\sigma} \approx \sigma, \mathbf{p}' \approx \mathbf{p}$. Detailed derivations are given in Appendix Sec. A.

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

$$\hat{s}_{\text{reg}} = 1-\hat{\sigma} = \mathbb{E}\left[1-\sqrt{\frac{\pi}{8}}\|\hat{\mathbf{p}}-\mathbf{p}\|_1\right]. \quad (18)$$

Comparing Eq. 12 with Eq. 18, the predicted confidence of RLE-based methods are linear in $\hat{\sigma}$ and only models the annotation variation but ignores the instance size. It also averages across all the keypoints, without excluding the occluded keypoints, leading to more inconsistencies with the expected value of OKS (Eq. 7).

Table 1. mAP (first four columns) and mAR (last column). "Orig" means applying their original ways of estimating confidence. "Mean", "Pred", and "GT" correspond to adjusting the confidence prediction using mean, predicted, and ground truth area in Eq. 12 respectively. Results show that our closed-form adjustment improves the calibration and thus increases the mAP.

Method	Type	Orig	Variables in Eq. 12			mAR↑
			Mean	Pred	GT	
SBL	heatmap	72.4	72.2	73.0	73.6	75.6
RLE	regression	72.2	71.8	73.2	73.3	75.4

4.4. Confidence Correction

Although the confidence values in these three forms (Eqs. 12, 17, and 18) all decrease as σ increases, the rankings of estimated confidence still differ from that of actual OKS accuracy. One explanation is that when it comes to a specific sample, the actual OKS will vary, but the predicted confidences of both heatmap- and RLE-based methods remain unchanged since they don't consider the instance size and keypoint falloff constants. For two similar σ's, the OKS will likely have different rankings depending on l, which becomes inconsistent with the ranking of confidences.

Motivated by the above analysis, we provide a simple confidence correction to make the pose network better calibrated to the OKS. This can serve as the empirical verification of our theoretical derivations. From Eq. 12, we can see that to match the format of expected OKS, we need the knowledge of σ and l. For σ, we obtain from the heatmaps by Pearson's chi-squared test for heatmap-based methods and directly from the predicted $\hat{\sigma}$ for RLE-based methods. For l, we estimate it and use the ground truth. Table 1 demonstrates that this adjustment from the theoretical analysis of the expected OKS improves mAP. However, this rescoring is based on the dismantling of metrics under ideal assumptions. We further address non-idealities in the next section, using the proposed ConfidenceNet.

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.

Method	Confidence	Backbone	Input Size	#Params (M)	#GFLOPs	mAP↑	AP.5↑	AP.75↑	AP (M)↑	AP (L)↑	mAR↑
Detection											
SBL	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	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	Hm	HRNet-W32	256×192	28.54	7.70	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	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	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	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	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	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	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	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	*const.*	ResNet-50	256×192	34.0	5.5	65.6	88.1	71.8	61.3	70.2	74.9
IPR	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

5. Calibrated ConfidenceNet (CCNet)

The correction in Sec. 4.4 is insufficient to fully close the calibration gap because it assumes that the network predicts the keypoint location perfectly. In practice, different models have different correlations between the prediction and $\hat{\sigma}$. As such, we propose Calibrated ConfidenceNet (CCNet) (see Fig. 2) as an efficient and effective calibration add-on to existing pose estimation methods. Denoting the previous pose network as PredNet, which estimates keypoint locations, we add the lightweight CCNet to predict confidence based on the features of PredNet. For instance, for heatmap-based methods, we detach and utilize the penultimate features after the 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 CCNet to access PredNet's rich features. Furthermore, as the PredNet is fixed, mAR remains unaffected.

Formally, CCNet outputs a calibrated confidence $\hat{s}_k \in [0, 1]$ for each keypoint given the input $\mathbf{x}$. It additionally predicts a visibility $\hat{v}_k \in [0, 1]$ to correct the bias caused by the thresholding operation in existing practice (Eq. 5). 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 keypoints as

$$\mathcal{L}_{\text{conf}} = \sum_{k=1}^{K} (\hat{s}_k - s_k)^2, \tag{19}$$

where s_k is the OKS for this keypoint. For visibility, we commonly treat it as a binary classification and use a Binary Cross-Entropy loss

$$\mathcal{L}_{\text{vis}} = -\sum_{k=1}^{K} (v_k \log \hat{v}_k + (1 - v_k) \log(1 - \hat{v}_k)). \tag{20}$$

The total loss, which updates only CCNet, is the following weighted sum

$$\mathcal{L} = \mathcal{L}_{\text{conf}} + \lambda \mathcal{L}_{\text{vis}}, \tag{21}$$

where λ serves as a weighting hyperparameter. Following the OKS form (Eq. 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 pose estimation tasks on three benchmarks: MSCOCO (Lin et al.,

Table 3. mAP evaluation on the COCO-WholeBody validation set based on Poseur (Mao et al., 2022). We base our CCNet on the part confidence and improve the whole body AP by 2.3.

		Body	Foot	Face	Hand	Whole
	Whole	67.2	63.6	84.6	58.3	61.0
mAP↑	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 4. The proposed CCNet improves all mAP and AUSE-PCK evaluations on the MPII validation set.

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

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

	mAP↑	mAR↑	Pearson Corr↑		AUSE↓	
	KP		Ins	KP	Ins	KP
RLE	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	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

2014), MPII (Andriluka et al., 2014), and MSCOCO-WholeBody (Jin et al., 2020). For the downstream tasks, we evaluate the 3D fitting task on 3DPW (Von Marcard et al., 2018).

MSCOCO consists of 250k person instances annotated with 17 keypoints. We evaluate the model with mAP over the standard 10 OKS thresholds. We also evaluate on **MPII** with the Percentage of Correct Keypoints (PCK) and on **MSCOCO-WholeBody**, which includes face and hand keypoints. 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 task, **3DPW** is a more challenging outdoor benchmark with around 3k SMPL annotations for testing. We follow the convention (Kolotouros et al., 2019) and use MPJPE, PA-MPJPE, and MVE as the evaluation metric. Additional implementation details and pseudo-code are provided in Appendix Sec. B.

6.2. Comparisons with SOTA

MSCOCO is the most challenging dataset to evaluate pose models on. Since our method is a plug-and-play module after the training of pose models, we evaluate our method on several baselines, including SBL (Xiao et al., 2018), HRNet (Sun et al., 2019), ViTPose (Xu et al., 2022) for heatmap-based pipelines, RLE (Li et al., 2021a), IPR (Sun et al., 2018), Poseur (Mao et al., 2022) for regression-based pipelines, using their officially released checkpoints. Results in Tab. 2 show that our simple yet effective method gives improvements across varying backbones, learning pipelines, and scoring functions. It is model-agnostic and is applicable even when the uncertainty estimation capabilities of different networks vary.

We further posit that pose estimation methods should be aware of confidence estimation and report improved mAP and corresponding mAR even though many methods do not compare their mARs. The gap between the two reflects how well- (or rather, poorly-) calibrated a pose model is. Qualitative visualizations of the calibrated confidence are provided in Appendix Sec. C.

COCO-WholeBody dataset evaluates the task of whole-body pose estimation, which includes body, face and hand keypoints. The convention is to assign the whole instance confidence to each part, which is unreasonable for evaluating the AP of the corresponding part. By simply changing the confidence of each part to the aggregation of the predicted part (Tab. 3 third row) instead of all the keypoints (Tab. 3 second row), the AP is significantly improved. After applying the proposed CCNet, we further improve the AP on every part and the whole body.

MPII is a single-person dataset and another commonly used benchmark. We use both OKS and PCK as the accuracy metric, which reflects on the mAP and AUSE (Ilg et al., 2018), respectively, as the final evaluation that considers both accuracy and calibration. Table 4 demonstrates the proposed CCNet is better on all metrics and therefore is metric and benchmark agnostic.

6.3. Confidence Evaluation

We are among the first to systematically explore confidence estimation for human pose estimation. The additional studies verify how CCNet will benefit confidence estimation beyond the AP measure.

Pearson Correlation between instance/keypoint accuracy and its confidence estimate is another measure of the quality of the confidence forecaster (Li et al., 2021a; Gu et al., 2021a; Bramlage et al., 2023). A well-estimated confidence estimate is proportional to the expected accuracy given the input condition. Our model gives stronger correlations between confidence and accuracy (4-5th col in Tab. 5).

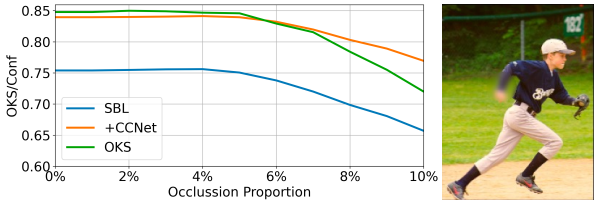


Figure 3. The (average) OKS and the decrease in estimated confidence correlate well with the circular occluder size (proportion to the input size). The right panel illustrates a Gaussian blur placed on the right wrist.

Table 6. 3D errors on 3DPW test set. Results show that better calibrated 2D pose net further improves the 3D results.

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

Area Under Sparsification Error (AUSE) (Ilg et al., 2018; Franchi et al., 2022) is plotted by gradually removing the most uncertain samples and computing the remaining error. Such a metric reveals how closely the estimated confidence matches the factual accuracy. The best confidence ranking is based on the actual coincidence between the prediction and ground truth. The results in the last two columns of Tab. 5 show that our method is qualified to pick out more accurately predicted poses and filter out predictions with larger errors, which is helpful for real-world deployments.

Occlusion Robustness (Bramlage et al., 2023) tests the confidence estimate based on simulating object occlusions with synthesis patches added. As the size of the added occlusion patch increases (such as blur at the wrist), the annotation ambiguity caused by blur occlusions also increases. Predicted coordinates at multiple positions behind the occluder are considered feasible and possible (Chen et al., 2023), so the distance (error) between the (mean) prediction and a single annotation coordinate increases as the size of the occluder increases. Figure 3 shows that confidence shrinks along with accuracy (the green curve) as the occlusion patch expands, but it better matches accuracy after calibration. Once occlusion exceeds a certain level, humans can no longer estimate the keypoint and simply label it as invisible.

3D Mesh Recovery is a challenging task, especially on the in-the-wild data, such as 3DPW test set (Von Marcard et al., 2018). A common way to improve the 3D predictions is to align the projected 3D poses with the predicted 2D poses from off-the-shelf 2D pose estimators. Confidence, therefore, is a critical indicator of whether the predicted

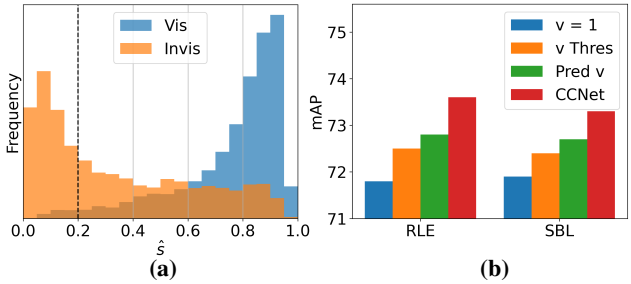


Figure 4. Visibility aggregation ablation. **(a)** Confidence distribution of visible and invisible keypoints. **(b)** The effectiveness of additional visibility prediction in keypoint-to-instance confidence aggregation.

2D poses are trustworthy. Mathematically, it can be treated as the weight of the distance between the projected 2D location and its predicted 2D location. In this way, A better-calibrated model will better distinguish the quality of the predicted 2D keypoints and help reduce the downstream error for mesh recovery. Empirically, the 2D detection results are given by the off-the-shelf pose network (Xiao et al., 2018); we update the 3D mesh with a 2D reprojection loss. The initial 3D predictions are from SPIN (Kolotouros et al., 2019). Table 6 shows that the calibrated 2D pose network better refines the 3D predictions.

6.4. Design Choices & Discussions

Surrogate Losses (Kendall & Gal, 2017; Lakshminarayanan et al., 2017; Corbière et al., 2019; Amini et al., 2020; Qi et al., 2021; Yu et al., 2021) propose various methods to estimate confidence. In our empirical explorations, we surprisingly find these sophisticated methods capture uncertainty no better than MSE (Appendix Tab. C). This might be bound by post-hoc confidence estimation of a frozen PredNet and estimatability (Yu et al., 2024). Note that training-time adaptation of confidence (Bramlage et al., 2023; Pathiraja et al., 2023) which require adjusting the model architecture and training from the beginning with the proposed losses, though well-calibrated, generally hurts prediction accuracy (Oh & Shin, 2022) and leads to a less satisfactory mAP. How to better hybridize the advantages of these two approaches is a promising direction for future work.

Input Features are the basis of our confidence estimation and we treat them as frozen penultimate features to preserve the lightweight nature of CCNet. To verify that they contain sufficiently rich information, we compared them with input features from shallower layers, prediction, and original keypoint confidence estimates which roughly indicate the ground-truth range. A strategy of copying the backbone and fine-tuning similar to Corbière et al. (2019); Yu et al. (2021); Zhang et al. (2023) is also considered. As results in Appendix Tab. g, we find that the penultimate feature input

is sufficient (Yu et al., 2024). In particular, keeping spatial information and predicting confidence using 1x1 channel-wise convolution is crucial for detection-based methods with spatial penultimate features.

Confidence Aggregation studies how to convert the keypoint confidences into their corresponding instance confident. Existing works (Xiao et al., 2018; Gu et al., 2021a; 2023) empirically set a visibility threshold based on the confidence estimate ($\tau_{\hat{s}}$ in Eq. 5). They found that the AP is sensitive to the choice of this thresholding hyperparameter (0.2 as default). Furthermore, the model has a different inductive bias from the human; keypoints with high confidence are not necessarily visible (Fig. 4(a)). This indicates that, human annotators would mark these keypoints as invisible, while the model remains confident in guessing the occluded keypoints' positions. The calculation of instance confidence includes unnecessary keypoints that are not considered in the accuracy evaluation of only visible keypoints, leading to further misalignment. Thus, different common aggregations are studied in Fig. 4(b). The visibility classification strategy (the green bars) of the CCNet shows more consistency with human-perceived visibility without much computational burden. Additional confidence calibration (Eq. 19 shown with the red bars) further increases performance.

Limitations. While post-hoc methods share the merit of less training time and computation, they are also limited by the penultimate features given by the frozen pose estimator. Additionally, to extend our work to 3D pose estimation may need nontrivial changes since it has a different learning paradigm or output representation (3D coordinates or pose and shape parameters). The other bottom-up paradigm for multi-person pose estimation needs to further consider the association of keypoints with each individual. Their calibration problems are important and challenging for future work.

7. Conclusion

This work is the first to study the pose calibration problem of aligning the predicted confidences with the OKS accuracy metric. We show theoretically how current methods are miscalibrated and empirically verify the derivation with a closed-form solution to close the gap. We further propose a Calibrated ConfidenceNet (CCNet) to learn a network-aware branch to align the OKS. Our experiments demonstrate that CCNet applies to various pose methods on various datasets. The improved confidence is thoroughly evaluated and also shows promise to help downstream tasks.

Acknowledgments

This research/project is supported by A*STAR under its National Robotics Programme (NRP) (Award M23NBK0053). We would also like to thank the ACs and reviewers for their valuable suggestions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 2021.

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. In *NeurIPS*, 2020.

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

Bramlage, L., Karg, M., and Curio, C. Plausible uncertainties for human pose regression. In *ICCV*, 2023.

Chen, R., Yang, L., and Yao, A. MHEntropy: Entropy meets multiple hypotheses for pose and shape recovery. In *ICCV*, 2023.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. Addressing failure prediction by learning model confidence. In *NeurIPS*, 2019.

Franchi, G., Yu, X., Bursuc, A., Tena, A., Kazmierczak, R., Dubuisson, S., Aldea, E., and Filliat, D. MUAD: Multiple uncertainties for autonomous driving, a benchmark for multiple uncertainty types and tasks. In *BMVC*, 2022.

Gu, K., Yang, L., and Yao, A. Dive deeper into integral pose regression. In *ICLR*, 2021a.

Gu, K., Yang, L., and Yao, A. Removing the bias of integral pose regression. In *ICCV*, 2021b.

Gu, K., Yang, L., Mi, M. B., and Yao, A. Bias-compensated integral regression for human pose estimation. *TPAMI*, 2023.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, 2017.

Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., and Brox, T. Uncertainty estimates and multi-hypotheses networks for optical flow. In *ECCV*, 2018.

Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., and Luo, P. Whole-body human pose estimation in the wild. In *ECCV*, 2020.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2014.

Kolotouros, N., Pavlakos, G., Black, M. J., and Daniilidis, K. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

Kuleshov, V. and Deshpande, S. Calibrated and sharp uncertainties in deep learning via density estimation. In *ICML*, 2022.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., and Lu, C. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021a.

Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., and Tu, Z. Pose recognition with cascade transformers. In *CVPR*, 2021b.

Li, Z., Liu, J., Zhang, Z., Xu, S., and Yan, Y. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z., and den Hengel, A. v. Poseur: Direct human pose regression with transformers. In *ECCV*, 2022.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., and Gal, Y. Deep deterministic uncertainty: A simple baseline. *CVPR*, 2023.

Oh, D. and Shin, B. Improving evidential deep learning via multi-task learning. In *AAAI*, 2022.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., and Murphy, K. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017.

Pathiraja, B., Gunawardhana, M., and Khan, M. H. Multiclass confidence and localization calibration for object detection. In *CVPR*, 2023.

Pierzchlewicz, P. A., Cotton, R. J., Bashiri, M., and Sinz, F. H. Multi-hypothesis 3D human pose estimation metrics favor miscalibrated distributions. In *arXiv*, 2022.

Qi, Q., Luo, Y., Xu, Z., Ji, S., and Yang, T. Stochastic optimization of areas under precision-recall curves with provable convergence. In *NeurIPS*, 2021.

Shen, M., Bu, Y., Sattigeri, P., Ghosh, S., Das, S., and Wornell, G. Post-hoc uncertainty learning using a Dirichlet meta-model. In *AAAI*, 2023.

Shi, D., Wei, X., Li, L., Ren, Y., and Tan, W. End-to-end multi-person pose estimation with transformers. In *CVPR*, 2022.

Song, H., Diethe, T., Kull, M., and Flach, P. Distribution calibration for regression. In *ICML*, 2019.

Sun, K., Xiao, B., Liu, D., and Wang, J. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

Sun, X., Xiao, B., Wei, F., Liang, S., and Wei, Y. Integral human pose regression. In *ECCV*, 2018.

Von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., and Pons-Moll, G. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018.

Wehrbein, T., Rudolph, M., Rosenhahn, B., and Wandt, B. Probabilistic monocular 3D human pose estimation with normalizing flows. In *ICCV*, 2021.

Wei, F., Sun, X., Li, H., Wang, J., and Lin, S. Point-set anchors for object detection, instance segmentation and pose estimation. In *ECCV*, 2020.

Xiao, B., Wu, H., and Wei, Y. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.

Xu, Y., Zhang, J., Zhang, Q., and Tao, D. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022.

Yu, X., Franchi, G., and Aldea, E. SLURP: Side learning uncertainty for regression problems. In *BMVC*, 2021.

Yu, X., Franchi, G., Gu, J., and Aldea, E. Discretization-induced Dirichlet posterior for robust uncertainty quantification on regression. In *AAAI*, 2024.

Zhang, F., Zhu, X., Dai, H., Ye, M., and Zhu, C. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

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.

This appendix includes **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 (Fig. e), but they are supposed to have similar rankings regardless of uncontrollable and irreducible uncertainty (Kendall & Gal, 2017). 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 Eq. 11

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

$$\mathbb{E}_{\mathbf{p}\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I})}[\text{OKS}] \tag{22}$$

$$= \int_{\mathbf{p}} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|\mathbf{p}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\mathbf{p}-\hat{\mathbf{p}}\|^2}{2\mathrm{l}^2}\right) \mathbf{dp} \tag{23}$$

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

Lemma A.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.,

$$\int_{\mathbf{p}} \mathcal{N}(\mathbf{p}|\boldsymbol{\mu}, \sigma^2\boldsymbol{I})\mathcal{N}(\hat{\mathbf{p}}|\mathbf{p}, \mathrm{l}^2\boldsymbol{I})\mathbf{dp} = \mathcal{N}(\hat{\mathbf{p}}|\boldsymbol{\mu}, (\sigma^2+\mathrm{l}^2)\boldsymbol{I}) \tag{25}$$

$$\iff \int_{\mathbf{p}} \exp\left(-\frac{\|\mathbf{p}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\hat{\mathbf{p}}-\mathbf{p}\|^2}{2\mathrm{l}^2}\right) \mathbf{dp} \tag{26}$$

$$= \frac{2\pi\sigma^2\mathrm{l}^2}{(\sigma^2+\mathrm{l}^2)} \exp\left(-\frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2(\sigma^2+\mathrm{l}^2)}\right). \tag{27}$$

Lemma A.2. *In another perspective, term within* exp *of Eq. 24 can be also arranged w.r.t.* $\mathbf{p}$ *as*

$$-\|\mathcal{D}\mathbf{p} - \vec{\mathcal{E}}\|^2 - \mathcal{F}, \tag{28}$$

$$\mathcal{D} = \frac{1}{\sqrt{2\frac{\sigma^2\mathrm{l}^2}{\sigma^2+\mathrm{l}^2}}}, \vec{\mathcal{E}} = \frac{\mathrm{l}^2\boldsymbol{\mu} + \sigma^2\hat{\mathbf{p}}}{\sqrt{2(\mathrm{l}^2+\sigma^2)\mathrm{l}^2\sigma^2}}, \mathcal{F} = \frac{\|\hat{\mathbf{p}}-\boldsymbol{\mu}\|^2}{2(\sigma^2+\mathrm{l}^2)}. \tag{29}$$

Substituting back into Eq. 24 obtains

$$\frac{1}{2\pi\sigma^2} \int_{\mathbf{p}} \exp(-\|\mathcal{D}\mathbf{p} - \vec{\mathcal{E}}\|^2 - \mathcal{F})\mathbf{dp} \tag{30}$$

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

$$- \vec{\mathcal{E}} \tag{32}$$

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

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

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

where $\mathcal{D}, \mathcal{F}$ are independent of $\mathbf{p}$ conditional on the image (Eq. 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 Eq. 17 and model distribution (or heatmap) approximates noisy ground truth distribution instead of the pure one. Following (Wehrbein et al., 2021), sigmas are estimated by fitting heatmap with Gaussian.

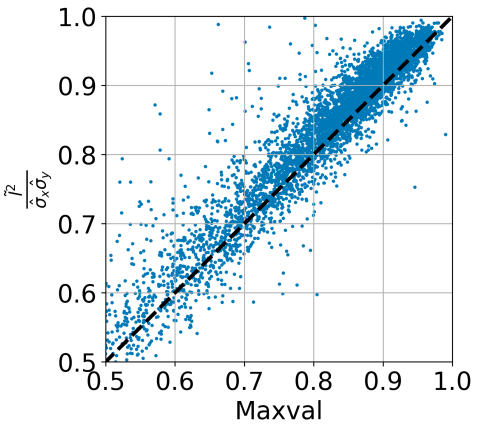


Figure f. Maximum values of the heatmap are almost coincident with our estimated scoring (peak density as Eq. 17), which verifies derivation.

A.4. Optima of Eq. 4 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 Eq. 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 (Eq. 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 confidence 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**, Eq. 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 Eq. 18 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 A.3. *Imperfect (but not bad) heatmap follows (Gu et al., 2021a)*

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

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

MSE (Eq. 14) is derived as

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

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 Eq. 67 uses Eq. 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 Eq. 17, and

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

Substituting them back to Eq. 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 –

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

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

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

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

Denoting

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

with Lemma A.1, Eq. 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 Eq. 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 A.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 Eq. 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

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

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

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

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

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
   ...
```

The score is unnormalized density at $\hat{\mathbf{p}}$ Eq. 64 as

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

□

B. Implementation Details

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

Architectures part supplements Para 1 in Sec. 5. For regression-based methods, the architecture is one fully connected layer with 2048D flattened feature input, 17D keypoint confidence and 17D visibility output; for heatmap-based method, we apply a 2D 1x1 convolution with 256D-channel input and 34D-channel output before a spatial global average pooling. Different numbers of layers and widths of the network are experimented, and only one FC/-Conv layer is found to work very well with the rich penultimate features. Negligible additional inference latency is brought by this lightweight head. Sigmoid is used for normalized confidence.

Training. The parameters are initialized with the default Kaiming initialization. The Adam (Kingma & Ba, 2014) optimizer is used for training without weight decay. The initial learning rate is $1e-3$, multiplied by 0.1 in the 9K-th step, and results are reported for 12K steps. Ground truth bounding boxes are provided as input to top-down pose estimation methods.

OKS on MPII (Andriluka et al., 2014). Since annotated keypoint sets are different between COCO (Lin et al., 2014) and MPII (Andriluka et al., 2014) 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.

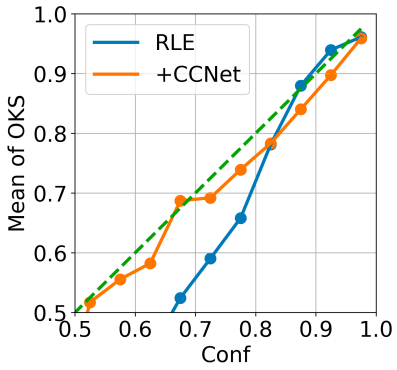


Figure g. Calibration plot. Estimated confidence well reflects the expected OKS value after pose calibration in our context.

	RLE	Alea	CE	DeepEns	SOAP
mAP	72.2	73.4	73.5	73.5	73.4

Method		**+CCNet**	**+CCNet+Add1**	**+CCNet+Add2**
RLE	72.2	73.6	73.6	73.5
SBL	72.4	73.3	73.2	73.3

Method		**Pool+FC**	**Conv+Pool**
SBL	72.4	72.6	73.3

Table g. mAP results of input feature studies on the COCO: the use of (**Add1**) information from keypoints and their original under-calibrated confidence, where for detection-based methods, the penultimate 2D feature map is concatenated with the interpolated final keypoint heatmap; (**Add2**) lower-level feature input (ResNet's "`layer3.5`"); (**3**) keeping spatial information is important for detected-based methods.

Pose Calibration variants in Pierzchlewicz et al. (2022); Bramlage et al. (2023) are mainly based on keypoint EPE instead of instance OKS and also do not focus on mAP. Instead, in our context, calibrated pose confidence is expected to well predict pose accuracy (Fig. g).

C. More Experimental Results

Surrogate Losses. Different losses including Bayesian weight posterior (Lakshminarayanan et al., 2017) and surrogate optimization (Qi et al., 2021), perform similarly well (Tab. C).

Input Features. The additional input original prediction and confidence estimate, along with lower-level features, do not result in an mAP improvement (Tab. g). Our design exploration also found maintaining the 2D spatial layout

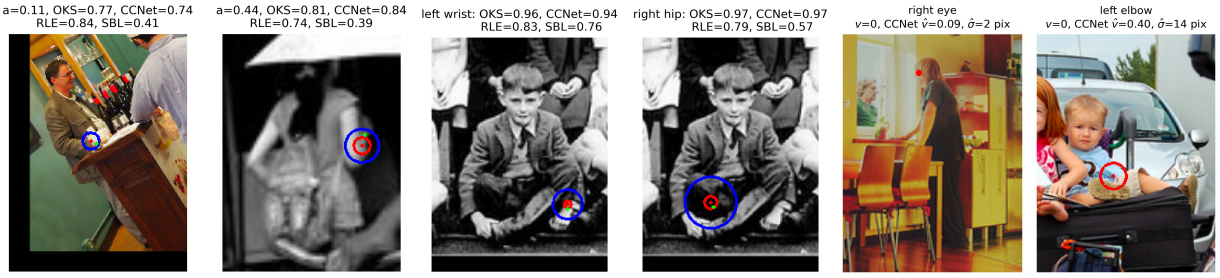
Method		$2e-3$	$2e-1$	$2e1$
RLE	72.2	73.4	73.6	73.4
SBL	72.4	73.2	73.3	73.1

Table h. The loss weighting hyperparameter λ tolerates multiple magnitudes and is not sensitive to select. The numbers are mAP.

and applying 1x1 channel-wise convolution proved to be crucial for detection-based methods.

Loss Weighting Hyperparameter λ balances the OKS loss and visibility loss. $\lambda = 2e - 1$ is used in our paper. Varying λ over several magnitudes has a limited impact on the mAP (Tab. h). We speculate that the confidence and visibility prediction tasks are relevant and not contradictory; thus, the losses will not conflict with each other.

Visualizations. Figure h shows the effects of area, per-keypoint falloff, and visibility, respectively. Our CCNet better aligns with OKS and human perception. For the 2D pose models themselves, CCNet can calibrate both under-confident and overconfident samples (Fig. i & j). From the visualizations on the downstream task in Fig. k, we can see that the better-calibrated 2D confidences can better instruct the optimization of 3D mesh.



(a) **(b)** **(c)**

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.

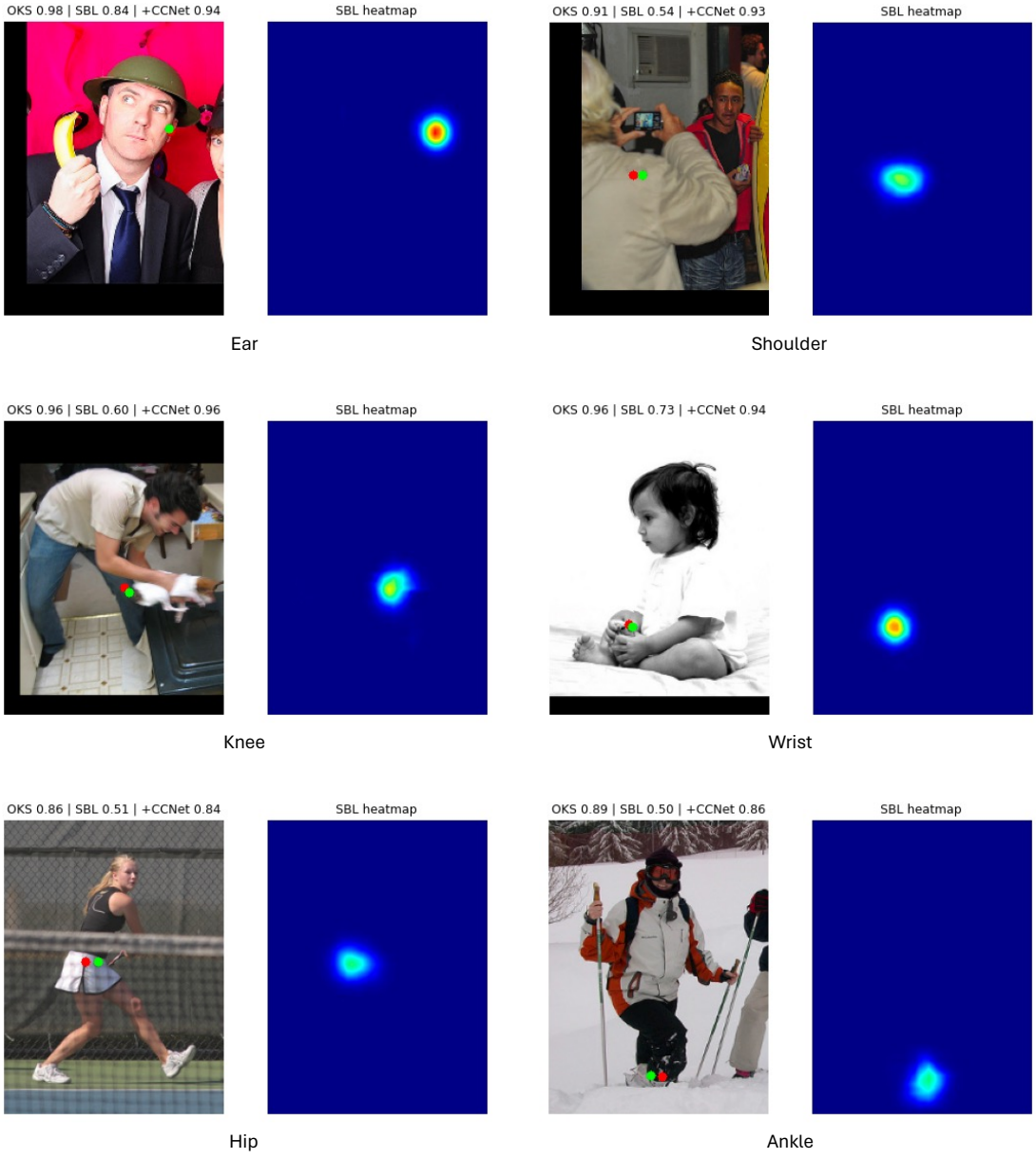


Figure i. Our CCNet also helps calibrate underconfident pose estimation – the keypoint detection is not far from the ground truth but the confidence cannot reflect the accuracy well.

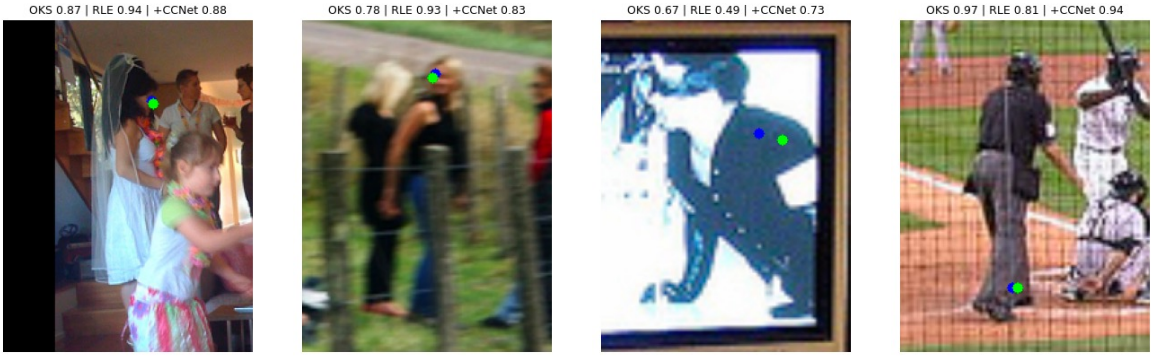


Figure j. Some qualitative visualizations of better-calibrated confidence estimation achieved by the RLE incorporated with our CCNet. Green dots are for ground truth while blue dots represent predictions.

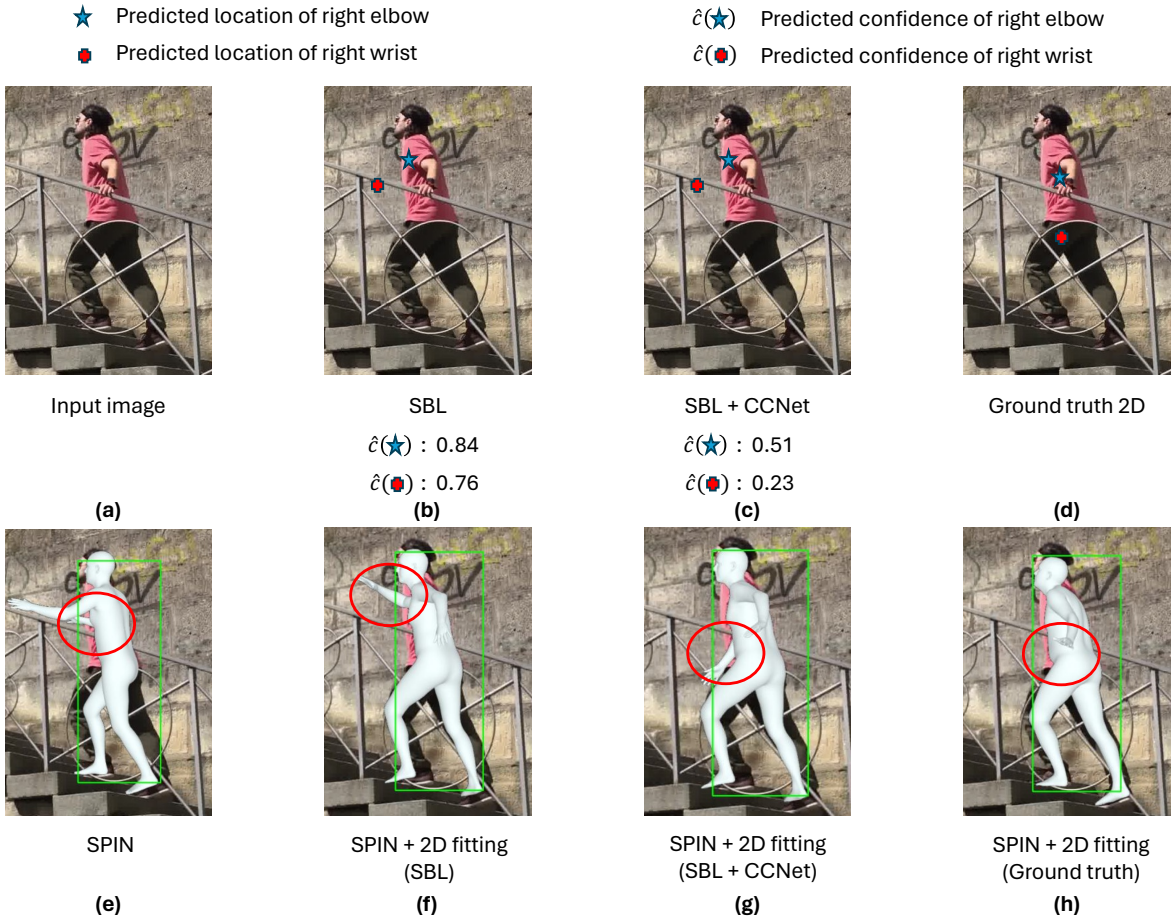


Figure k. Illustration of benefits of pose confidence calibration in downstream 3D model fitting. We demonstrate a case where the right arm (highlighted in a red circle) is heavily occluded. While the uncalibrated SBL and calibrated SBL+CCNet predict the same (wrong) 2D keypoint location (**b** & **c**), the poorly calibrated SBL estimates a *high* confidence which misleads the mesh fitting to a wrong position for the right arm (from **e** to **f**). In contrast, after adding the CCNet, the confidence for the occluded right elbow and right wrist are *lower*, which allows the mesh to maintain its original prediction for the right arm and spare more efforts on optimizing other keypoints with higher correctly estimated confidences, e.g., left arm and legs (**g**).