ProbPose: A Probabilistic Approach to 2D Human Pose Estimation

Miroslav Purkrabek and Jiri Matas
Visual Recognition Group
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
purkrmir@fel.cvut.cz

Abstract

Current state-of-the-art Human Pose Estimation methods ignore out-of-image keypoints in both training and evaluation and use uncalibrated heatmaps as keypoint location representations. We propose ProbPose, which predicts for each keypoint: a calibrated probability of keypoint presence at each location in the activation window, the probability of being outside of it, and its predicted visibility. To address the lack of evaluation protocols for out-of-image keypoints, we introduce the CropCOCO dataset and the Extended OKS (Ex-OKS) metric, which extends OKS to out-of-image points. Tested on COCO, CropCOCO, and OCHuman, ProbPose shows significant gains in out-of-image keypoint localization while also improving in-image localization through data augmentation. Additionally, the model improves robustness along the edges of the bounding box and offers better flexibility in keypoint evaluation. The code and weights are available on the project website[1].

1. Introduction

Human pose estimation (HPE) has seen significant progress, achieving robust performance on standard datasets. The most successful models like ViTPose [32] typically use a top-down, heatmap-based approach, where heatmaps serve as dense representations of keypoint locations across an image.

In this paper, we address two limitations of top-down HPE methods: (1) all previous methods overlook out-of-image keypoints during training and, more importantly, ignore them in evaluation, and (2) heatmaps restrict their use to simple point estimates. While seemingly separate, these issues are interrelated and can be addressed through an appropriate choice of representation and loss function.

[1] MiraPurkrabek.github.io/ProbPose/

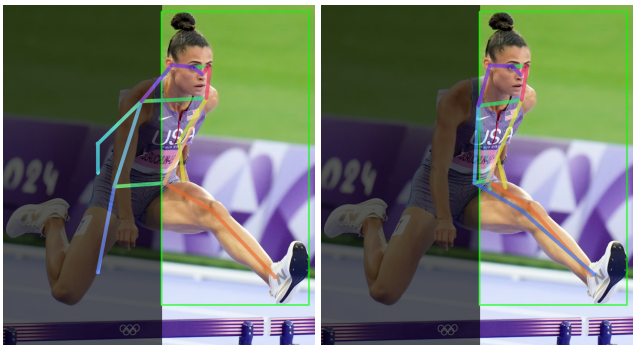


Figure 1. ProbPose (left) vs. ViTPose [32] (right) output on a cropped image with the dark part removed. The bounding box, on which the pose estimators operate, is detected by YOLOX-x [5]. ProbPose estimates keypoints outside of the image better, e.g. the right leg which is wrongly aligned with the left leg by ViTPose, for which out-of-image keypoints represent a domain shift.

Top-down methods localize keypoints within a specific region called the activation window. This activation window is the part of the image mapped to the predicted heatmap. The activation window often matches the model input and it may exceed image boundaries due to preprocessing steps, such as maintaining aspect ratios or enlarging bounding boxes.

Under challenging conditions, such as heavy occlusion or cropped images, keypoints may fall outside the activation window. Current methods do not address this issue and simply ignore out-of-window keypoints during training. Even more concerning, evaluation metrics such as OKS (Object Keypoint Similarity) and PCK (Percentage of Correct Keypoints) assess only in-image keypoints. This lack of accountability often leads models to predict keypoints inaccurately on other joints when the actual keypoint is not visible in the image. For example, as shown in Fig. 1, the left leg is mistakenly assigned to the same joints as the right leg. Furthermore, current top-down models do not provide any

indication of whether a keypoint is inside or outside the activation window, and models are not penalized for attempting to localize every keypoint.

A second challenge in existing HPE methods lies in the reliance on heatmaps as the keypoint representation. Heatmaps provide dense representation of keypoint locations. However, the standard approach of generating target heatmaps with a fixed Gaussian sigma and training with mean square error (MSE) forces the model to output heatmaps with fixed sigmas as well, limiting the accuracy of these estimates.

Decoding heatmaps to obtain a point estimate typically involves taking the argmax, often refined with UDP decoding [9]. While this approach is effective for identifying the most likely pixel, it is optimal only when the goal is to classify a single correct pixel for each keypoint. However, practical applications demand more flexible interpretations of keypoint locations. For example, an application might require an estimate of the smallest area where a keypoint is located with at least 95% probability, or, as in human-robot interaction, identify areas where the model has high uncertainty to ensure safe interactions. Heatmaps, which are prone to overconfidence, are not designed to indicate uncertainty or convey "don't know" outcomes, even when Bayes risk is high, an essential factor in safety-critical applications.

To address these limitations, we propose ProbPose, a model with multiple outputs that fully describe each keypoint. ProbPose predicts (1) calibrated *presence probabilities* indicating whether the keypoint lies within the activation window, (2) keypoint's location within the activation window, (3) quality estimate of the localization and (4) visibility. Moreover, instead of using traditional heatmaps, we employ probability maps.

Probability maps differ from heatmaps in several ways: they contain values between 0 and 1 and are normalized to sum to 1, providing calibrated probabilities. Unlike heatmaps, which assume a fixed Gaussian shape, probability maps adapt their shape based on the data, without forcing any assumptions. Additionally, instead of decoding using a simple argmax, we compute the expected OKS for each pixel and select the location that maximizes this score, an approach we term expected OKS maximization. Probability maps are trained using an OKS-based loss modified for dense predictions, formulated as an expected risk minimization problem. Probability maps do not outperform heatmaps on keypoints in the image but have comparable localization ability. Their biggest advantage is in their versatility and better modeling of the underlying distribution. Calibrated probability maps allow for more complex reasoning than point estimates and have a direct interpretation.

Since current datasets and metrics lack evaluation protocols for out-of-image keypoints, we introduce a new dataset, CropCOCO, comprising cropped images from COCO, and the Extended OKS (Ex-OKS) metric, which builds on OKS to evaluate out-of-image keypoints and presence probability.

In summary, our main contributions are:

1. A concept of **presence probability** that keypoints is in the activation window, distinct from confidence, which measures the model's trust in its own estimate.
2. **Model for out-of-image keypoints** that can localize keypoints within the activation window, even beyond the image boundaries, or predict presence probability for out-of-window keypoints.
3. **OKSLoss adapted for dense predictions** in risk minimization formulation
4. We show that **cropping-based data augmentation** improves out-of-image keypoint localization, supports presence probability estimation, and enhances in-image localization, similar to Hide&Seek [10]
5. New **CropCOCO dataset** and **Ex-OKS** evaluation metric for more realistic assesment of real-world performance

2. Related work

There are three main approaches to 2D human pose estimation: top-down, bottom-up, and single-stage. Single-stage [22, 23, 28, 31] methods directly predict individuals and their skeletons in an image. Bottom-up methods [3, 6] detect all keypoints first and then group them into individual skeletons. The top-down methods [16, 24, 29, 32, 33], despite their challenges, remain the most successful. They use human detectors to process each bounding box independently, providing the best performance on most datasets.

Among top-down methods, heatmap-based approaches are the most popular. Unlike regression-based methods [25, 27], which directly predict 2D coordinates, heatmap-based methods [16, 24, 32] predict a heatmap with the same aspect ratio as the input. The keypoints are then estimated as the maximum of the heatmap. Heatmaps are more robust and easier to train than regression-based methods, although efforts have been made to narrow the gap between the two approaches [18, 26].

A key development in heatmap-based methods is UDP (unbiased data processing) [9], which improves coordinate encoding and decoding. Heatmaps are typically defined as 2D Gaussians with a fixed sigma centered on the keypoint. Subpixel decoding allows for smaller heatmaps, reducing computational costs. The maximum value of the heatmap, called the "confidence," is often used for three tasks: visibility estimation, quality estimation, and in/out-of-image decisions. Visibility is sometimes outsourced to specialized heads, quality prediction could be improved by OKS prediction, as shown in [7]. In/out decisions (later defined as *presence probability* in this paper) are not explicit, as they

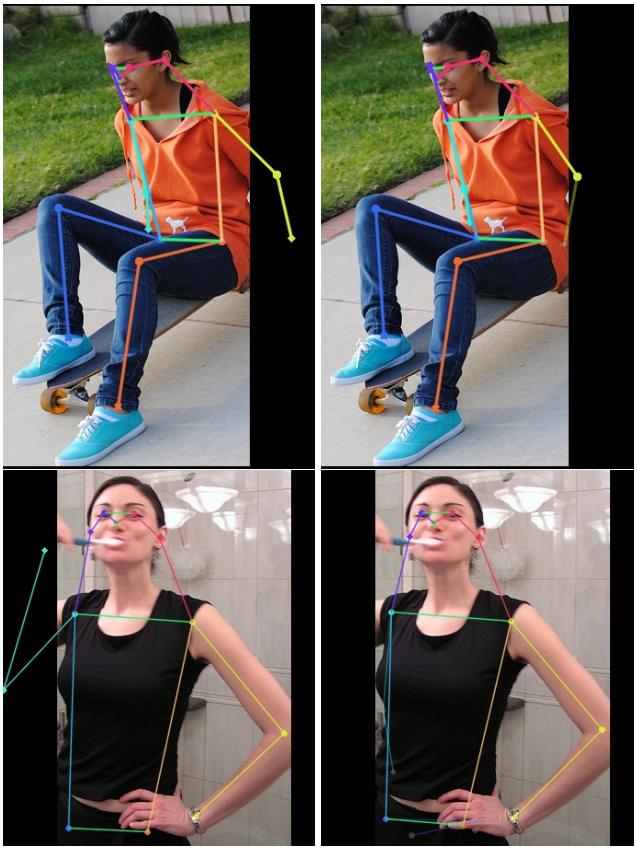


Figure 2. ProbPose-s (left), compared with ViTPose-s (right), improves localization outside the image and handles occlusion better on COCO (first row) and the proposed CropCOCO (second row). ProbPose estimates points well outside of the image.

are not needed for OKS evaluation, but are important for visualization and real-world applications.

Most heatmap-based methods are trained with MSE loss, which encourages the output heatmap to match the target Gaussian heatmap. All keypoints are treated equally, with the same Gaussian and sigma. The training focuses only on keypoints within the activation window, meaning the model never encounters an empty heatmap. Since the input of the model often extends beyond the bounding box to capture more context, the activation window can include areas outside the image. This enables a limited prediction of keypoints outside the image, although the models are not specifically trained for this task. For more details, see Appendix A in the supplementary.

Heatmap-based approaches can predict keypoints outside of the image to some degree. Regression methods like RLE [13] go even further and predict keypoints further from the image. Our goal is not to localize out-of-image keypoints per se but emphasize their importance in training and evaluation.

Alternative loss functions have also been proposed. Adaptive Wing Loss [30] improves on original Wing Loss [4]. CornerNet [11] introduces a modified Focal Loss (originally in [15]) for object detection. RTMO [17] treats pose estimation as coordinate classification, assuming a normal distribution with varying sigmas for different keypoints, and trains with a negative log likelihood loss. Distribution-based losses such as [21] or Kullback-Leibler divergence have also been used for heatmap-based pose estimation. Most importantly, [19] introduce a new OKSLoss, that is directly tied to the evaluation metric. Their OKSLoss is computed on the prediced keypoints. We modify it to use pixel-wise to optimize probability maps without any shape assumptions.

This paper also applies a cropping data augmentation technique. Similarly to information-dropping methods [10] like Hide&Seek or random bounding box masking, this technique introduces the domain shift by creating more invisible keypoints during training, encouraging the model to leverage the structure of the human body.

The most widely used dataset for 2D human pose estimation is COCO [14], with MPII [1] being a less common alternative. Datasets like OCHuman [34] and CrowdPose [12] focus on multibody problems such as occlusion and self-occlusion. For COCO and related datasets, the evaluation metric is Object Keypoint Similarity (OKS), while Percentage of Correct Keypoints (PCKh) is used for MPII.

However OKS ignores unannotated keypoints, as COCO is not fully annotated – many individuals have missing keypoints. Missing annotations can occur either because the keypoint is outside the image or because of occlusion. All annotated keypoints in current datasets are within the image, as annotating keypoints outside the image would be too costly and imprecise. Additionally, OKS does not penalize guessing, and if a model predicts a keypoint inside the image when it is not present, there is no penalty. Evaluating the presence probability of keypoints (whether they are in the activation window) is crucial for real-world applications and requires out-of-image annotations.

We address these issues by introducing a new dataset, CropCOCO, created by cropping COCO images, along with a new evaluation metric, Extended OKS (Ex-OKS), which accounts for keypoints outside the image.

3. Method

ProbPose introduces several key innovations:

1. A novel approach to keypoint localization using *probability maps*.
2. A different application of OKSLoss [19] corresponding to expected risk minimization.
3. A new attribute for each keypoint, termed *presence probability*.
4. A data augmentation technique to handle keypoints outside the image.

5. A double-probmap approach that expands the field of view, enabling localization of keypoints positioned far outside the image.

In combination with visibility and quality estimation, these elements provide a complete description of each keypoint's state in ProbPose – indicating whether it is within the image, visible, and the confidence level of the estimate. The probability map offers calibrated localization estimates. Each component is explained in detail in the following sections.

3.1. Probability maps and loss function

In conventional methods, heatmaps are trained with Gaussian targets and MSE loss, resulting in fast convergence, but a Gaussian-shaped bias in output. However, there is no theoretical basis for assuming that the posterior localization probability follows a Gaussian distribution. The normal distribution in OKS metric reflects human-induced error in the annotation process, but we assume the underlaying distribution is not Gaussian but reflects the body shape. During inference, heatmaps typically use the argmax for point localization and the peak value as a combined measure of localization quality and visibility.

In our approach, we require the model to output probability maps rather than traditional heatmaps. Unlike heatmaps, probability maps always sum to 1, achieved through Sparsemax [20] as the final activation function.

Each pixel in the probability map represents the posterior probability

$$p_L(x_i) = p(x_i | k_j \in AW, img), \tag{1}$$

where x_i is pixel coordinate, k_j is the j-th keypoint and *AW* stands for activation window. With Sparsemax, each probability map sums to 1 for each keypoint:

$$F \sum_{x_i \in AW} p(x_i | k_j \in AW, img) = 1 \tag{2}$$

We refer to this as the **localization probability** $p_L(x_i)$.

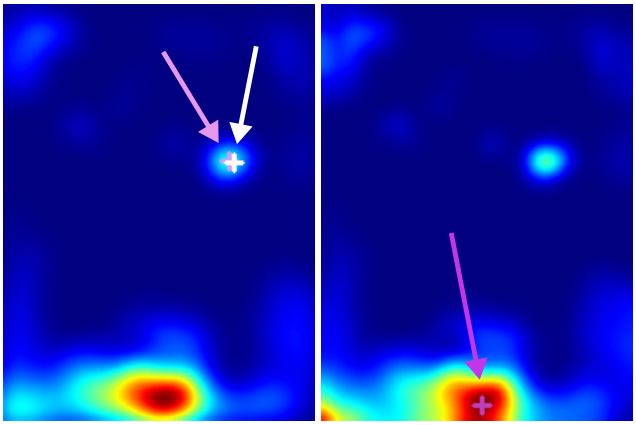
3.1.1. Loss function

We train the probability maps using a modified version of OKSLoss [19], aligned with the evaluation function. The loss is formulated as an expected risk minimization problem.

$$R_{exp}(x_i) = \big(1 - OKS(x_i)\big) \cdot p_L(x_i) \tag{3}$$

$$\mathcal{L}_{OKS}(x_i) = (1 - \alpha) R_{exp}(x_i) + \alpha g(x_i) \tag{4}$$

In Eq. (4), $R_{exp}(x_i)$ represents expected risk and $\mathcal{L}_{OKS}(x_i)$ is the loss function for pixel x_i. In contrast to [19], we apply OKSLoss on each pixel of probability map instead of on predicted keypoints. This loss encourages pixels with low OKS to have low localization probability and those with high OKS to have higher localization probability.



(a) Predicted probability map p_L (b) Expected OKS map $\mathbb{E}_{OKS}$

Figure 3. UDP [9] decoding of p_L vs. expected OKS maximization. On the left, the white cross marks the maximum value and the light purple one is the point refined by UDP. In the right image, the dark purple cross is maximum expected OKS. The probability map contains a small, one-pixel-wide peak marked by the white cross. Maximizing expected OKS is robust to sharp local peaks and prefers areas with big mass compared to simple maxima.

The term $g(x_i)$ is the heatmap gradient, computed with the Sobel operator, and serves as a regularizer controlled by the hyperparameter α. Without regularization, the probability map quickly forms a sharp peak and overfits to training data. Smoothing regularization minimizes differences between neighboring pixels, promoting a smooth distribution without assuming a specific shape, as Gaussian regularization does for MSE-trained heatmaps.

For localization, instead of directly using argmax, we compute the expected OKS (Eq. (5)) for each pixel and take its argmax, refining it with quadratic interpolation for subpixel precision. This approach, based on expected OKS rather than UDP decoding, slightly enhances localization, especially in cases of bimodal heatmaps such as in Fig. 3.

$$\mathbb{E}_{OKS}(x_i) = \sum_{x_j \in PM} p_L(x_j) \cdot OKS(x_i, x_j) \tag{5}$$

Probability maps offer improved versatility. They allow point estimates, as required by the COCO evaluation protocol, and enable probabilistic queries, such as defining the smallest area containing a keypoint with at least 90% probability. However, the peak of the probability map can no longer reliably indicate the quality of the localization or the presence of keypoints in the activation window. The former is addressed by predicting the OKS for each keypoint as in [7], and the latter is addressed by directly predicting the keypoint's *presence probability*.

3.2. Presence probability

Previous models and evaluation protocols have overlooked the question of whether a keypoint is within the activation window. Top-down approaches assume that the bounding boxes fully enclose all keypoints, but this is not always the case due to factors such as occlusion or crop (see Appendix A in the supplementary for further discussion). Thus, it is essential for the model to assess whether a keypoint is actually present in the activation window. This is usually addressed by thresholding confidence (the maximum heatmap value) with an arbitrary cut-off. However, evaluation protocols have ignored keypoint presence, with models predicting a location regardless.

Presence probability $p(k_j \in AW|img)$ complements probability maps, which inherently assume that keypoints are within the activation window. If the keypoint is outside, the presence probability reflects this; if inside, the probability map provides the location. The visual explanation is in Fig. 4. We abbreviate the *presence probability* to $p_p(k_j)$ and train it with Binary Cross Entropy loss.

To train the presence probability, we need keypoints that are certainly outside the activation window. These samples are generated through various data augmentation techniques, such as scaling and rotation, with *random cropping* as the primary source.

3.3. Calibration

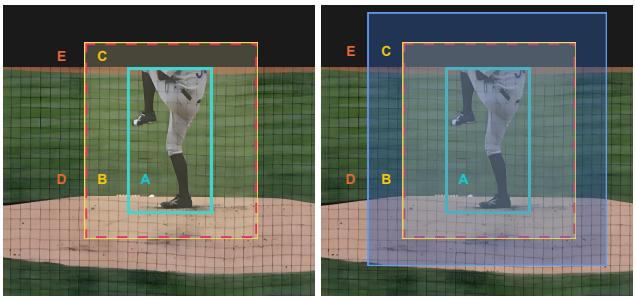
Both the probability maps and presence probability are calibrated. Probability maps are calibrated so that the top 5% of the map contains 5% of the GT keypoints, the top 10% contains 10% of the GT keypoints, and so on. We achieve this calibration through temperature scaling [8] on CropCOCO. As a result, the calibrated probabilities align with the true underlying distribution, enabling more versatile evaluations beyond simple point estimates. Details and calibration curves are in Appendix B in the supplementary.

3.4. Crop image augmentaton

As discussed in Appendix A, we consider the image border as a type of occlusion. To train the model to correctly identify keypoints outside the image, we need suitable training samples. Manual annotation of such keypoints is both costly and imprecise, so we generate these samples by cropping existing annotated keypoints out of the image.

Depending on the crop strength, the in-the-image keypoint may end up within the activation window but outside the image, or entirely outside the activation window. Keypoints within the activation window but outside the image are used to train probability maps, while those outside the activation window train the presence probability.

The cropping not only provides data for presence probability training but also enhances the model's ability to localize keypoints near the image boundary, extending even to areas beyond the visible border. This is especially useful for cases where the subject is partially cropped by the image border or in close-up pose estimation, where keypoints may lie far from the visible body. Examples of crops and result visualizations are in the Figs. 1 and 2.



(a) Traditional method (b) Double-probmap approach

Figure 4. Key regions in top-down pose estimation. Rectangles represent the bounding box, the model input and the activation window. Right: the double-probmap approach with a larger activation window. The probability map is responsible for localization of keypoints in activation windows (A, B, C) and the presence probability predicts whether the keypoint is outside (D, E). Note that double-probmap has a larger field-of-view. For detailed discussion about keypoint location, see Appendix A.

3.5. Enlarging activation window size

When predicting keypoints outside the image, a key question is how far from the image we aim to localize them. In applications where out-of-image keypoint localization is essential, it might be helpful to enlarge the activation window. However, simply increasing the input size or activation window typically reduces performance.

To address this, we propose a double-probmap approach (illustrated in Fig. 4b). In this method, an expert heatmap precisely localizes keypoints within a smaller activation window that matches the model input, while a larger activation window captures keypoints further from the image. The larger probability map has the same resolution but a wider field of view, resulting in a lower effective resolution. Both maps are trained on keypoints within their respective activation areas. If the larger map predicts a keypoint within the smaller activation area, the expert map refines the localization for greater accuracy.

This approach balances the trade-off between field-of-view and precision. The larger map offers a broader field of view, but with lower accuracy, while the expert map maintains precision within a smaller activation window. This method is particularly suitable for applications like close-up pose estimation, where a wide field of view is essential.

4. Data

To evaluate ProbPose, we introduce a new evaluation protocol based on OKS and mAP, along with a new dataset specifically designed for out-of-image keypoint localization and presence probability classification.

4.1. Evaluation metrics

Current metrics, such as OKS and PCK, evaluate only in-image keypoints, overlooking false positives. As a result, models overestimate confidence, guess keypoint locations, and are not penalized for predicting absent keypoints.

We propose a new evaluation approach that checks whether a keypoint is within the activation window and, if so, assesses its precise localization. This approach better reflects real-world scenarios, like human-machine interaction, where knowing if a keypoint cannot be localized is critical. To address this, we introduce Extended OKS (Ex-OKS), an extension of OKS that accommodates keypoints outside the activation window. For bottom-up and single-stage methods, where the activation window spans the entire image, Ex-OKS can be adapted accordingly.

In its localization-only form, Ex-OKS functions identically to the original OKS. However, Ex-OKS extends OKS by accounting for keypoints outside the activation window. The essential goal is for the model to indicate if a keypoint is absent, rather than attempting a guess.

Ex-OKS evaluates two variables for each keypoint: location x_i and presence probability p_p. Although the model outputs a continuous value for p_p, a binary decision (0 or 1) is often required in real-world applications, so we optimally threshold it for each model.

Formally, we define **Ex-OKS** as:

$$d_i = \begin{cases} d_e(x_i^*, x_i') & \text{if } p_p^* = 1 \text{ and } p_p' = 1 \\ d_e(\text{AW}, x_i') & \text{if } p_p^* = 0 \text{ and } p_p' = 1 \\ d_e(x_i^*, \text{AW}) & \text{if } p_p^* = 1 \text{ and } p_p' = 0 \\ 0 & \text{else} \end{cases} \tag{6}$$

$$\text{Ex-OKS} = \exp\left(\frac{-d_i^2}{2k^2\sigma^2}\right) \tag{7}$$

, where *AW* stands for activation window, d_e is Euclidean distance and $\cdot^*$ and $\cdot'$ are ground truth and predicted variables, respectivelly.

If both the ground truth and the model agree that $p_p = 1$, Ex-OKS defaults to the standard OKS evaluation.

When the ground truth $p_p^* = 0$ and the model predicts $p_p' = 1$, the penalty depends on the predicted location x_i' and its distance from the edge of the activation window, with lower penalties for points near the boundary. Similarly, if $p_p^* = 1$ but the model predicts $p_p' = 0$, the penalty is based on the ground truth location x_i^*, with smaller penalties for errors near the border of the activation window.

When both $p_p^* = 0$ and $p_p' = 0$, there is no penalty and the similarity score is 1.

Ex-mAP, built on Ex-OKS, extends mAP by incorporating both keypoint localization and presence probability, making it suitable for evaluating models with keypoints outside the activation window.

4.2. Datasets

We evaluated our model on the COCO dataset to ensure a fair comparison with existing models. All training was conducted on COCO, with crop image augmentation applied in specified experiments.

To assess performance on out-of-image keypoints, we introduce a new dataset, CropCOCO. CropCOCO consists of randomly cropped COCO validation images, with some keypoints positioned outside the activation window. Bounding boxes are recomputed for the cropped images to evaluate the model appropriately. While other researchers can create their own cropped dataset, we release CropCOCO to support reproducibility.

During COCO evaluation, we observed that keypoints near the image boundary are often annotated further inside the image, likely due to annotators avoiding placements near the edge. This causes our model to be penalized for accurately localizing keypoints outside the image. When keypoints near the border are excluded from evaluation, ProbPose demonstrates even stronger performance compared to other state-of-the-art methods. Although this issue is minor (causing about a 0.2% performance difference) and difficult to correct, we highlight it to raise awareness of its potential impact. For further details, see Appendix D.

Additionally, we evaluated our model on the OCHuman dataset to demonstrate its generalizability across different domains.

5. Experiments

The results in Tab. 1 compare our method with other SOTA approaches on COCO, CropCOCO, and OCHuman datasets using mAP and Ex-mAP metrics. For fair comparison, we evaluate models of similar size. All experiments use ground truth bounding boxes.

ProbPose improves on its baseline model, ViTPose, by approximately 1% in localization, a gain comparable to that achieved by other information-dropping augmentations like [10]. Importantly, ProbPose demonstrates stronger performance in Ex-OKS, showing improved detection of keypoints outside the activation window due to its direct *presence probability* prediction, as discussed in Sec. 5.1. Since COCO lacks out-of-image keypoints, Ex-OKS improvement here reflects enhanced presence probability classification. Localization metrics could be slightly higher if not for incorrect annotations near the bounding box borders, discussed further in Appendix D.

Model	COCO		CropCOCO		OCHuman	
	mAP	Ex-mAP	mAP	Ex-mAP	mAP	Ex-mAP
HRFormer-s [33]	75.2	74.6	70.9	64.3	60.3	60.0
PVTv2 [29]	72.0	71.5	70.7	63.1	58.5	58.1
SWIN-t [16]	73.5	72.9	71.3	65.0	58.1	57.9
ViTPose-s [32]	75.9	75.3	72.7	66.5	60.3	60.1
ProbPose-s	**76.6**	**76.4**	**81.7**	**73.9**	60.4	60.2
ProbPose-s-DP	76.2	75.4	80.9	71.4	**61.4**	**61.2**

Table 1. Keypoint localization accuracy in terms the of the standard mAP and the novel Ex-mAP metrics. ProbPose is compared with SOTA models of similar size on the COCO, CropCOCO, and OCHuman datasets. All evaluation is done using GT bounding boxes. DH stands for the double-probmap approach with a wider field of view. To compute Ex-mAP, we select the optimal threshold for each model. The best model is in **bold**, the second-best is underlined. ProbPose brings significant improvement for out-of-image keypoints and minor improvement on standard COCO.

The most significant improvement is unsurprisingly observed on CropCOCO, where ProbPose excels at detecting out-of-image keypoints, primarily due to crop data augmentation. Again, the localization metrics could be higher without annotation errors along the bounding box borders.

ProbPose-DP, the double-probmap variant, shows a slight performance decrease on COCO and CropCOCO, reflecting the trade-off between a broader field of view and in-image precision (see Sec. 3.5). In this variant, each heatmap is predicted by one head from shared features, and the larger heatmap decides when to defer to the more precise heatmap. If the larger heatmap estimates the wrong localization, the specialized heatmap cannot correct it.

ProbPose-DP has worse Ex-mAP as it solves a more difficult task. Ex-mAP evaluates keypoint presence in the AW so the larger AW in ProbPose-DP makes presence classification harder as the AW border is further from the image.

ProbPose-DP remains competitive with other SOTA methods, providing a better alternative to simply enlarging the input size or activation window. Interestingly, ProbPose-DP shows a notable improvement in OCHuman, likely due to its extended field of view, which enhances individual differentiation in multi-body scenes.

ProbPose variants perform well on OCHuman (unseen during the training), highlighting their strong generalization across domains.

5.1. Presence probability vs. confidence

Tab. 1 highlights the advantages of using presence probability over confidence, reflected in the increase in Ex-mAP. For clarity, Fig. 5 shows an additional classification accuracy for CropCOCO and its balanced subset. The task is to classify whether keypoints are inside or outside the activation window, indicating if they should be drawn or used in other applications. CropCOCO is unbalanced as there is much more in-AW than out-of-AW keypoints. To get a balanced subset, we randomly selected keypoints from both classes to have the same amount of each.

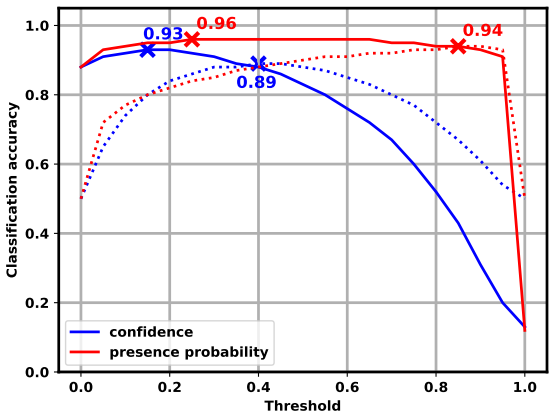


Figure 5. Keypoint presence in the activation window. Accuracy, inside and outside of the AW, of the classification based on thresholding of the ProbPose's presence probability and ViTPose's confidence. The accuracy is measured for CropCOCO dataset (full line) and its balanced subset (dotted line). The proposed presence probability outperforms ViTPose's confidence in both cases, reducing the error by 30% and 45% respectively.

In both balanced and unbalanced datasets, the presence probability outperforms the confidence score from the ViTPose model. This improvement is due to confidence not being trained to reflect presence probability, and the standard COCO dataset, without crop data augmentation, lacks sufficient training samples for this purpose.

Fig. 5 also provides insights into confidence thresholding. Current models often use an arbitrary threshold of 0.3 to decide whether to trust predictions without theoretical justification. The evaluation on CropCOCO reveals that the optimal confidence threshold is between 0.15 (for unbalanced datasets) and 0.4 (for balanced datasets). Thus, the commonly used threshold of 0.3 is close to the optimum,

crop	p-maps	COCO	CropCOCO
✗	✗	76.0	73.7
✗	✓	76.4	72.4
✓	✗	76.6	81.7
✓	✓	76.6	81.7

Table 2. ProbPose ablation; mAP on COCO and CropCOCO for versions without the probability maps and crop data augmentation. Crop augmentation clearly helps for cropped images. Probability maps bring negligible performance improvement as their main strength is in versatility.

which aligns with its generally good results. This study offers guidance on setting an appropriate threshold based on the characteristics of a specific dataset.

5.2. Ablation study

Tab. 2 evaluates the core components of our approach, probability maps and crop data augmentation. The analysis is conducted on the ProbPose-s model using ground truth bounding boxes on the COCO and CropCOCO datasets.

Using probability maps without crop data augmentation yields the lowest performance. This setup slightly improves localization on COCO but reduces the accuracy of out-of-image keypoints in CropCOCO. Without crop data augmentation, ProbPose encounters few out-of-image training samples, and probability map normalization forces predictions within the activation window, making probability maps less effective than heatmaps for out-of-image keypoints.

Adding crop data augmentation improves localization accuracy on both COCO and CropCOCO. Training with out-of-image keypoints also improves in-image localization, as the model relies more on the structure of the human body, similar to [10]. Compared to Hide&Seek, crop augmentation brings the same in-image performance improvement, but also improve out-of-image localization and generate samples for training *presence probability*.

Combining crop augmentation and probmaps achieves the same localization accuracy as using crop data augmentation alone. However, as discussed in Tab. 1, the probabilistic approach is more important for presence classification. While COCO-like evaluation protocols focus on point estimates, probability maps are more versatile and provide a direct mathematical interpretation. Heatmaps are optimized for point estimates, but probmaps support broader applications without compromising point-estimate accuracy.

Computational complexity. ViTPose-s (ViT-s backbone + HeatmapHead) performs 24M FLOPs. The full ProbPose-s (visibility, OKS, and p_p estimation) used for all experiments performs 35M FLOPs. The minimal version shrinks to 27M FLOPs (16% more than ViTPose-s). The difference in runtime is smaller than our measurement precision and both models run a $\sim$50 fps.

6. Conclusions

We present ProbPose, a novel approach to 2D human pose estimation that uses probability maps instead of heatmaps and introduces presence within-the-activation-map probability prediction. Unlike previous methods, we do not assume a Gaussian distribution for the localization probability and employ the expected OKS for both training and decoding, thus aligning the training loss and the evaluation metric. The key findings are:

1. Probability maps with OKSLoss improve versatility without sacrificing performance, offering more flexibility than point estimates.
2. Modified OKSLoss adapts to each keypoint type, enforcing sharper distributions for high-precision keypoints like eyes or nose, while broader distributions are used for less precise keypoints like hips. These distributions are more reflective of the image and are not constrained by a Gaussian assumption.
3. Predicting presence probability is more effective than thresholding confidence, reducing error by 45%. Presence probability is a better indicator for in/out decisions than the peak of a predicted Gaussian.
4. Cropping images during training improves both out-of-image localization and in-image performance by increasing data variability, similar to other information-dropping augmentations. Unlike Hide&Seek [10], crop augmentation also enables out-of-image localization.
5. Increasing the activation window size comes with some performance trade-offs, but the impact can be minimized. Our double-probmap model can localize keypoints up to 25% outside the bounding box with only an 0.4% decrease in AP performance.

Future work could explore the behavior of probability maps in multibody scenarios. Decoding with expected OKS maximization has proven especially useful for bimodal heatmaps, which are common in multibody images.

The issue of COCO keypoints near the edge of the bounding box remains unresolved. Ignoring these keypoints during training reduces performance as valuable data is lost. However, training with them teaches the model to place keypoints further inside the image. Cropping as a data augmentation strategy helps mitigate this issue during training but not during evaluation. Re-annotating the dataset would be too costly and likely imprecise, as it is difficult for humans to accurately place keypoints outside the image.

Acknowledgements. This work was supported by the Ministry of the Interior of the Czech Republic project No. VJ02010041 and the Czech Technical University student grant SGS23/173/OHK3/3T/13.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3

[2] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang. Adversarial semantic data augmentation for human pose estimation. In *European Conference on Computer Vision*, 2020. 12

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[4] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2245, 2018. 3

[5] Z Ge. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1

[6] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14676–14686, 2021. 2

[7] Kerui Gu, Rongyu Chen, and Angela Yao. On the calibration of human pose estimation. *arXiv preprint arXiv:2311.17105*, 2023. 2, 4

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 5, 12

[9] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 12

[10] Junjie Huang, Zheng Zhu, Guan Huang, and Dalong Du. Aid: Pushing the performance boundary of human pose estimation with information dropping augmentation. *arXiv preprint arXiv:2008.07139*, 2020. 2, 3, 6, 8

[11] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 3

[12] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 3

[13] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11025–11034, 2021. 3

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 3

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 7

[17] Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1491–1500, 2024. 3

[18] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019. 2

[19] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022. 3, 4

[20] Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR, 2016. 4

[21] Haoxuan Qu, Li Xu, Yujun Cai, Lin Geng Foo, and Jun Liu. Heatmap distribution matching for human pose estimation. *Advances in Neural Information Processing Systems*, 35:24327–24339, 2022. 3

[22] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022. 2

[23] Lucas Stoffl, Maxime Vidal, and Alexander Mathis. End-to-end trainable multi-instance pose estimation with transformers. *arXiv preprint arXiv:2103.12115*, 2021. 2

[24] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[25] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2602–2611, 2017. 2

[26] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018. 2

[27] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2013. 2

[28] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11060–11068, 2022. 2

[29] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 2, 7

[30] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6971–6981, 2019. 3

[31] Yabo Xiao, Xiaojuan Wang, Dongdong Yu, Kai Su, Lei Jin, Mei Song, Shuicheng Yan, and Jian Zhao. Adaptivepose++: A powerful single-stage network for multi-person pose regression. *arXiv preprint arXiv:2210.04014*, 2022. 2

[32] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 7

[33] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in neural information processing systems*, 34:7281–7293, 2021. 2, 7

[34] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul L. Rosin, Zixi Cai, Han Xi, Dingcheng Yang, Hao-Zhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation, 2019. 3, 16

ProbPose: A Probabilistic Approach to 2D Human Pose Estimation

Supplementary Material

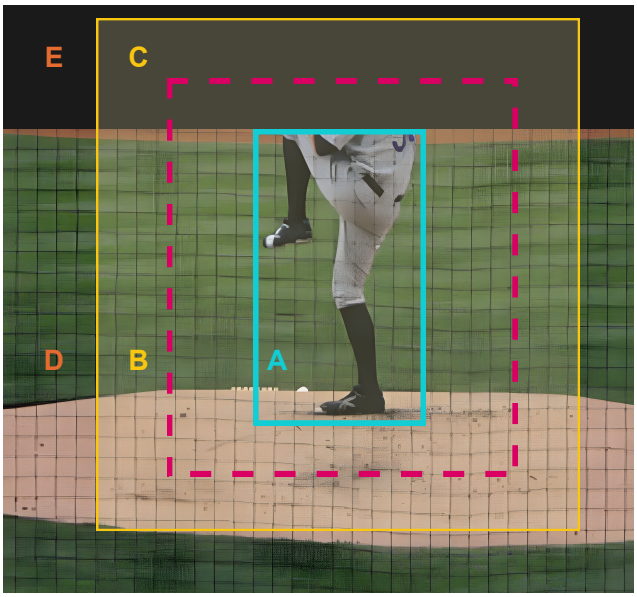


Figure 6. A scheme explaining where keypoints could be in the image. Rectangles represent the bounding box, the model input and the activation window (usually coincides with the model input). Image taken from the COCO val dataset.

A. On the types of keypoint

Fig. 6 illustrates possible keypoint locations, assuming that all keypoints of an individual are present within or outside the image. The image edge is treated as another form of occlusion, much like an object blocking part of the body. Whether a person is occluded by an object (e.g., a wardrobe) or cropped by the image, invisible keypoints must be estimated from the visible ones and the structure of the human body.

In the top-down approach to human pose estimation, the image is divided into three main areas, which are depicted in Fig. 6 and defined below.

Bounding box (bbox) – the tightest rectangle enclosing all visible parts of the individual. A perfect human detector outputs this kind of bounding box.

Model input – the part of the image cropped and fed into the top-down model. Due to aspect ratio constraints and the need for contextual information, the model input is usually larger than the bounding box and often includes areas outside the image.

Activation window (amap, AM) – the area where the top-down model localizes keypoints. This typically coincides with the model input but can be larger or smaller. Like the model input, the activation window often contains regions outside the image.

Dataset	A	B	C	D	E
COCO train	96.2	3.5	0.0	0.2	0.0
COCO val	95.8	3.9	0.0	0.2	0.0
CropCOCO val	68.8	2.2	23.5	0.1	5.3
OCHuman	99.2	0.8	0.0	0.0	0.0

Table 3. Domain shift between used datasets. Percenatages of keypoint types. For definitions, see text.

The bounding box, activation window, and image edge divide the space into five subareas, each behaving differently in the context of top-down human pose estimation. These subareas (A-E) are visualized in the Fig. 6.

A – inside the bbox. Visible keypoints can only exist within the bounding box.

B – inside both the activation window and the image. The vast majority of COCO keypoints fall into areas A and B. No visible keypoints are located outside the bounding box.

C – inside the activation window but outside of the image. Previous methods could theoretically predict keypoints in area C, but they lack the necessary training data to do so.

D – outside of the activation window but inside the image. Prior top-down methods cannot localize keypoints in this area or describe them in any way. Approximately 0.2% of keypoints in the COCO dataset fall into this category, meaning top-down methods are always penalized by OKS for these points. However, ProbPose marks these keypoints as "out" by predicting low presence probability and won't get penalized by Ex-OKS.

E – outside of both the image and the activation window. Like points in area D, keypoints in area E have been ignored by previous methods in both estimation and evaluation. ProbPose, along with Ex-OKS, addresses this issue using presence probability and a novel evaluation metric Ex-OKS.

The proportion of annotated keypoints in each area defines the domain of a dataset. For example, the domain of the COCO-val dataset is represented by the vector (95.8, 3.9, 0, 0.2, 0), where each value indicates the percentage of points in the corresponding subarea. In particular, there are no annotated keypoints outside the image, and approximately 99.8% of the keypoints are within the activation window. Traditional top-down methods assume that 100% of keypoints lie within the activation window.

Tab. 3 compares the domains of the datasets used for the ProbPose evaluation. Before this paper, no dataset included annotations outside the image, specifically in areas C and D. Therefore, no evaluation protocol worked with these areas. Thus, previous evaluation protocols did not account for these areas. Area D becomes critical under heavy occlusion, where the detected bounding box is much smaller than the individual. Likewise, areas C and E become important when the image is heavily cropped or in close-view pose estimation. The CropCOCO dataset tests the model under domain shift, where keypoints were moved from area A to areas C and E.

The visibility of the keypoint is only loosely related to areas A-E. Although visible keypoints are always within the bounding box (area A), invisible keypoints can be located in any of the areas. Importantly, classifying keypoint visibility is a different task from determining whether a keypoint is present in the activation window.

B. Model calibration

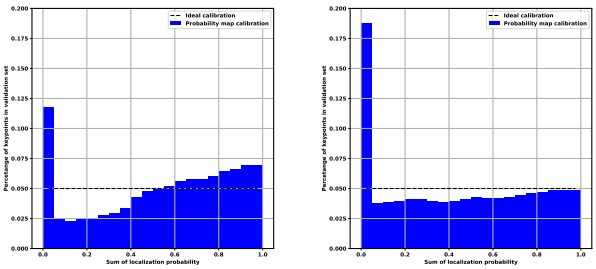
Training probability maps with OKSLoss results in uncalibrated probability maps. Calibration ensures that a probability map accurately refelcts the likelihood of finding a point in a specific area. For example, among all predictions where the model assigns probability 80%, approximately 80% of these predictions should be correct.

To achieve this, probabilities are summed starting from the largest values, prioritizing the most likely regions first. For instance, area with the top 5% of probabilities should contain exactly 5% of the ground truth points. Because the summation starts from the highest probabilities, the area corresponding to the top 5% is always a subset of the top 20% etc. As a single pixel often has a probability greater than 10%, calibration is done on sub-pixel precision.

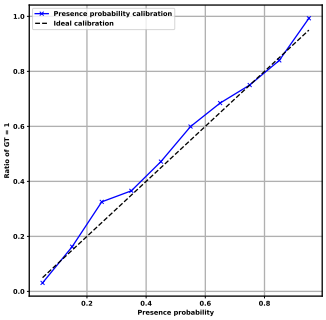
Calibration is performed on a validation set using temperature scaling [8], optimizing a single parameter T. The goal is to create an evenly distributed histogram as shown in Fig. 7a, which shows the calibration curves before and after the temperature scaling. Before calibration, the model was underconfident with more than 5% of GT points in the top-5% area. Notice the large peak in the 0% to 5% range. These correspond to keypoints where the model failed completely, predicting very low probability for the correct area, or incorrectly detecting the keypoint elsewhere ("keypoint stealing" in overlapping individuals). More research on overlapping individuals or data augmentation techniques such as [2] can help address this issue.

The Fig. 7b presents the calibration curve of the presence probability. Unlike keypoint localization, presence probability is naturally calibrated during training due to the use of binary cross-entropy loss and the similar distribution between the training and testing sets.

Calibrating probability maps allows for estimating calibrated posterior probability. By multiplying all elements in the probability map with presence probability, we get a posterior probability map, the probability that the keypoint is in the given pixel. The posterior probability maps are shown in Fig. 13 where ich color corresponds to area where a keypoint is with 10% probability. This leads to probabilistic statements such as shown in Figs. 11 and 12.



(a) Probability map calibration curves



(b) Presence probability calibration curves

Figure 7. Calibration curves of probability maps (a) before and after temperature scaling and presence probability (b). While presence probability is calibrated in a standard way, for probability maps we require that top 5% of pixels contains 5% of points etc. The peak in the calibration of the probability maps between 0% and 5% are hard errors where the model fails to locate the keypoint. These are usually caused by occlusion and multibody situations.

C. Expected OKS maximization vs. UDP

UDP decoding [9] estimates the global maximum of the predicted heatmap and then refines it to subpixel precision. To refine the localization, the heatmap is blurred using a Gaussian with fixed variance, and the maximum value is shifted toward the estimated peak of the local Gaussian. However, if the initial estimate (heatmap maximum) is incorrect, UDP refining cannot correct it. UDP decoding assumes the predicted heatmap follows a Gaussian distribution and estimates the peak of the predicted Gaussian.

Expected OKS maximization convolves the predicted probability map with an OKS kernel and calculates the expected OKS for each pixel. The OKS kernel varies for each keypoint type. The initial estimate is the global maximum of the expected OKS map. To achieve subpixel precision, we fine-tune the estimate using quadratic interpolation in the neighboring pixels. Expected OKS decoding makes no assumptions about the shape of the distribution and takes a global approach, favoring areas with larger mass over sharp peaks, aligning more closely with probability maps.

The difference between the UDP and OKS maximization decodings is shown in Figs. 9 and 10. When the predicted probability map is unimodal (which is true for most predicted heatmaps), the difference is negligible. However, if the probability map is multimodal or lacks a clear peak, expected OKS favors areas with greater mass.

D. Points on the bounding box border

When we evaluated ProbPose qualitatively on the standard COCO dataset, we observed that the ground truth annotations were not always where we expected, particularly in cases where OKS scores worsened the most. Specifically, we noticed that ground truth keypoints near the bounding box border were annotated inside the box but should have been placed outside. It appears that human annotators for the COCO dataset prioritized annotating as many keypoints as possible, even at the cost of accuracy.

Examples of such misannotations are shown in Fig. 8. As illustrated in the last row, this issue is not limited to the image border but also occurs along the bounding box border. This supports our hypothesis that the image border behaves as another form of occlusion, as discussed in Appendix A.

ProbPose demonstrates that training with crop data augmentation can help mitigate the impact of these incorrect annotations in COCO. However, we did not find an easy and automated solution to fix this issue in the evaluation set. Ignoring points near the bounding box border during the evaluation showed that ProbPose performs even better, but this approach also excludes many correctly annotated and presumably challenging keypoints. Manual reannotation may be necessary to address these errors.

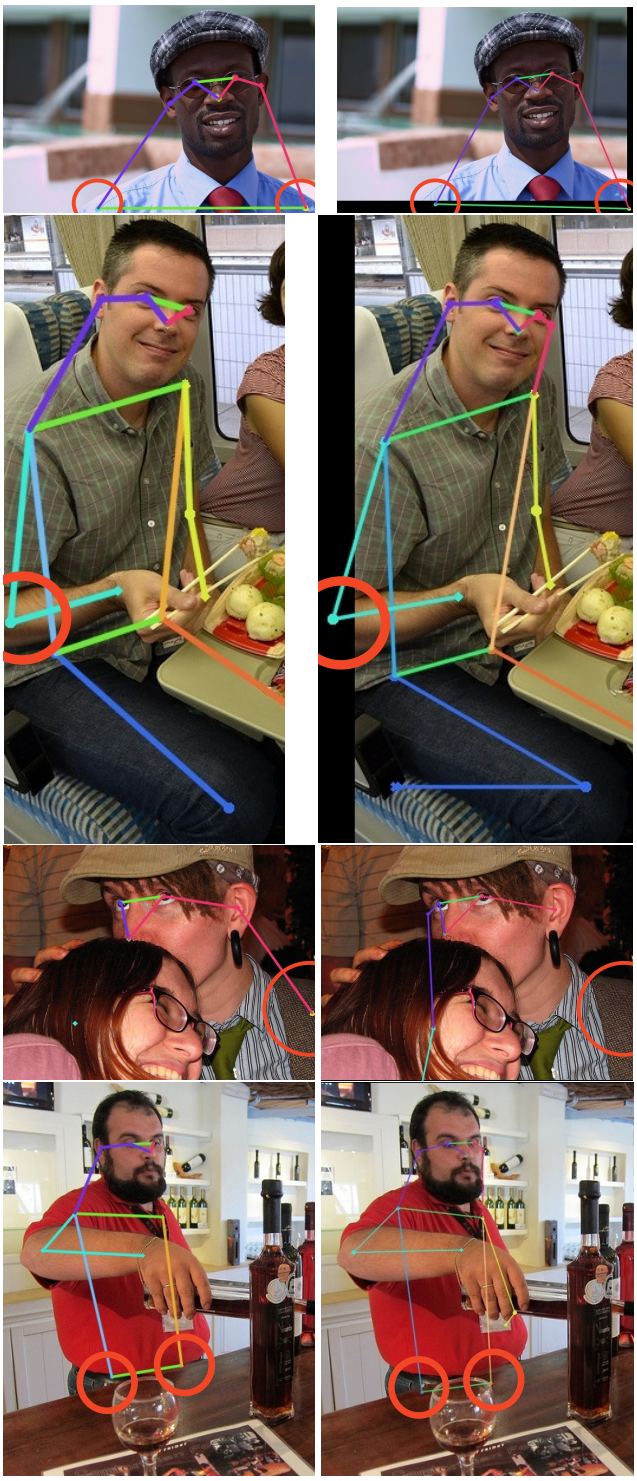


Figure 8. Ground truth annotation (left) vs. ProbPose-s (right) on COCO. Images showcasing dubious annotations along the bounding box border where ProbPose gets penalized even though its input seems better than ground truth. In the third row, our estimate is missing as we correctly predict it outside of the activation window. The problem is not only along the image border as shown in the last row.

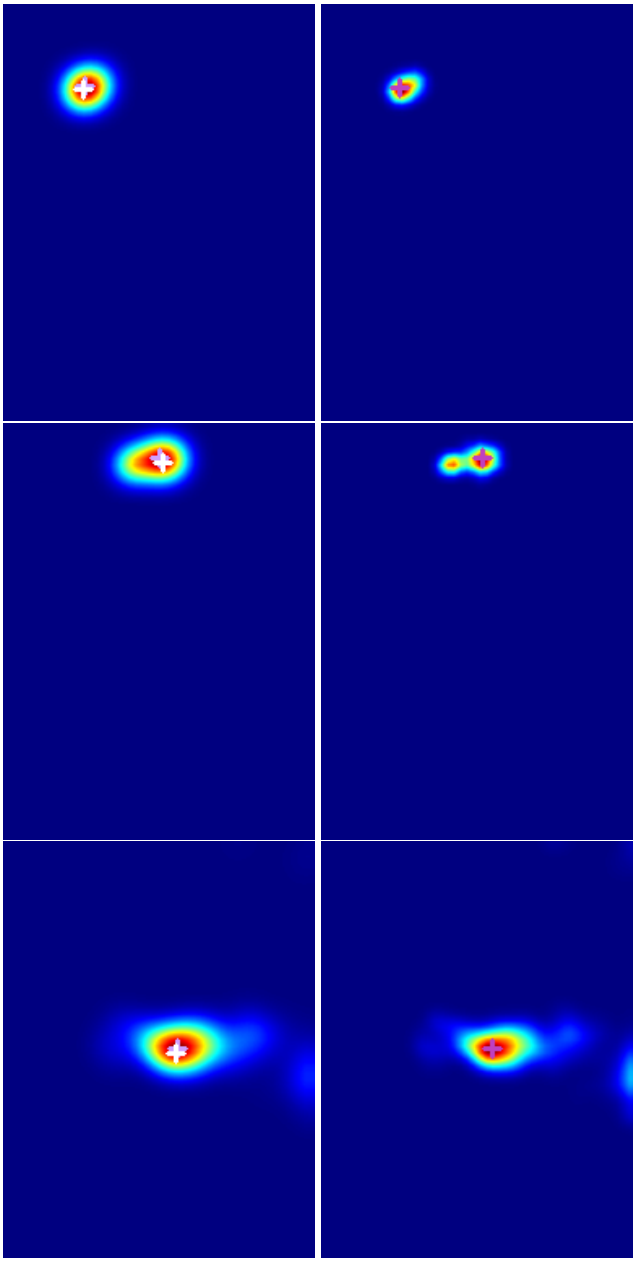


Figure 9. Decoding does not matter as the predicted heatmap is unimodal. Decoding predicted probability maps throught UDP (left) and expected OKS maximization (right). Probability map maximum in white, UDP-refined point in light purple and maximal expected OKS in dark purple. Majority of keypoints have such heatmaps so the difference in performance is not big. Notice the non-Gaussian shape of predicted probability maps and sharper peaks for expected OKS as opposed to UDP.

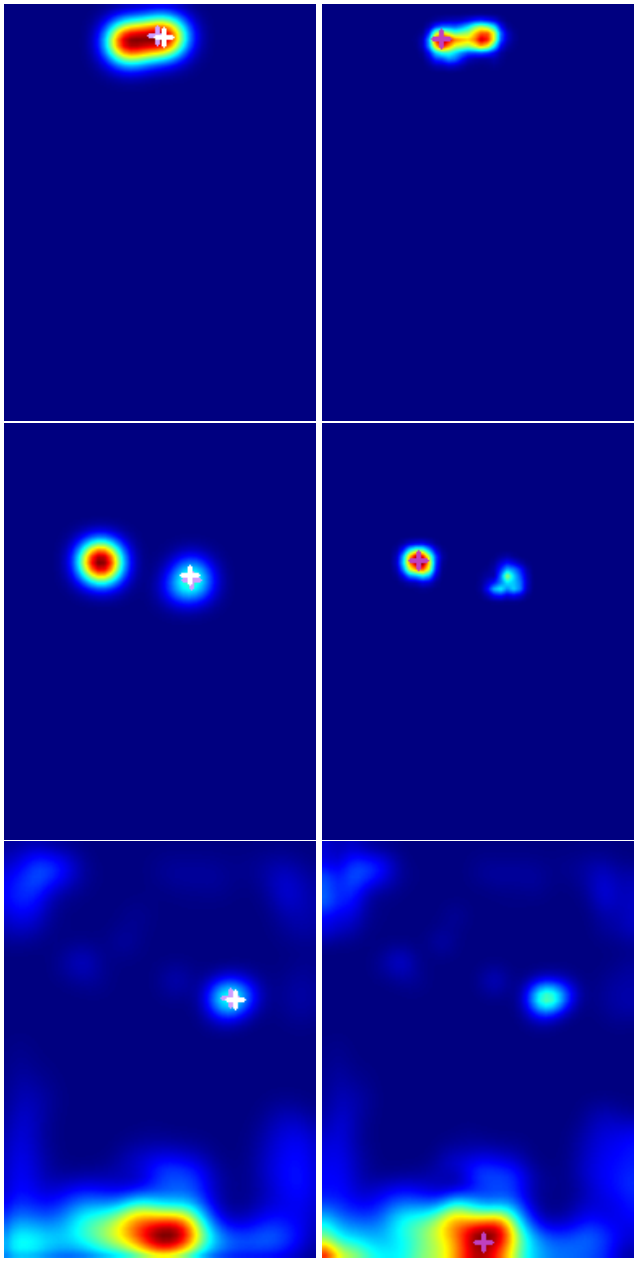


Figure 10. Decoding does matter when distributiona are multimodal. Decoding predicted probability maps throught UDP (left) and expected OKS maximization (right). Probability map maximum in white, UDP-refined point in light purple and maximal expected OKS in dark purple. The predicted heatmaps contain a small, one-pixel-wide peak marked by the white cross. Expected OKS have sharper peaks but predict optimal location globaly in areas with biggest ”mass” even though the maximal value could be elsewhere.

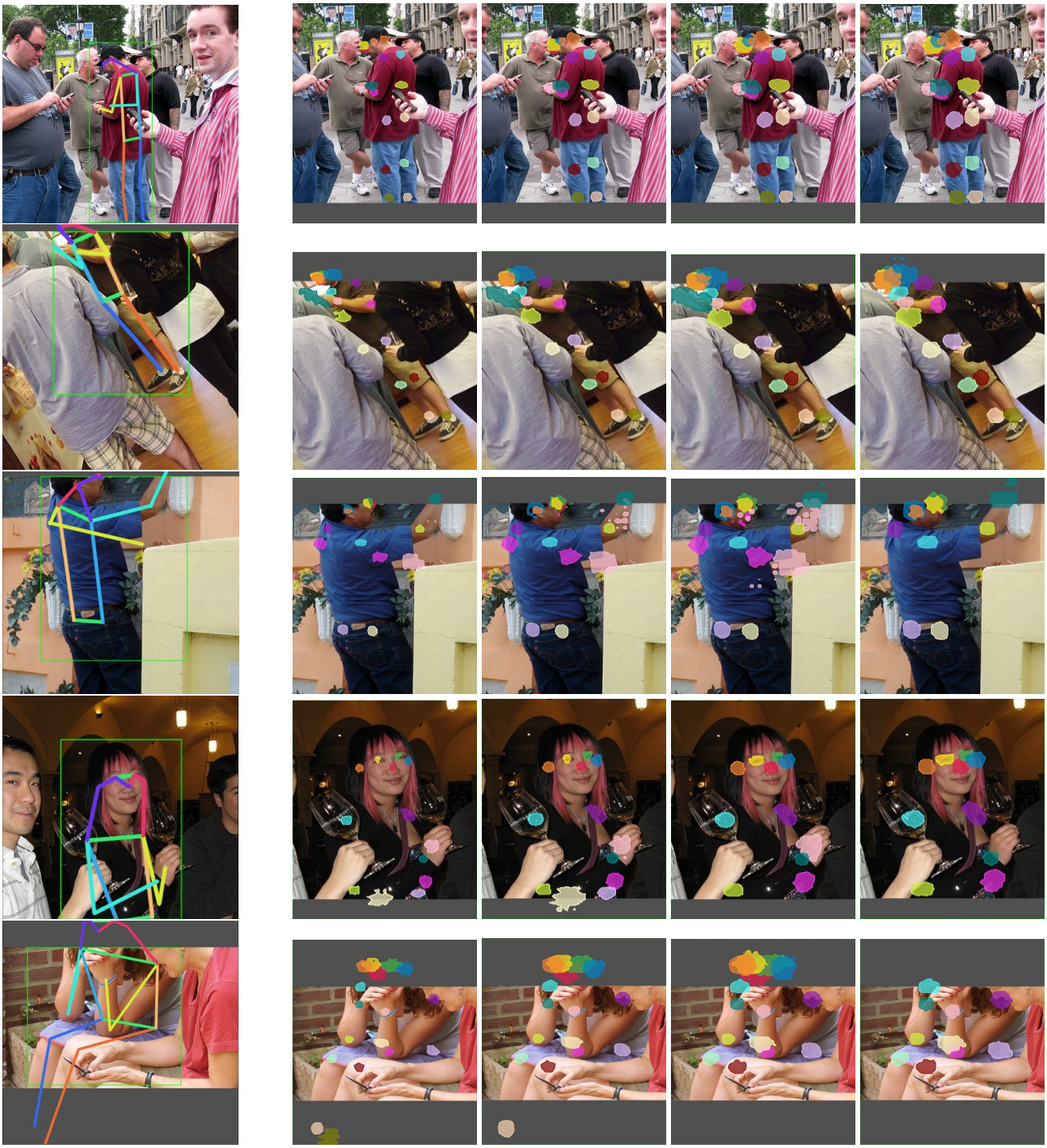


Figure 11. Images from CropCOCO; probability maps thresholded at different posterior probabilities. First column are input images with GT bboxes and estimated poses. Second to fifth columns are probability maps for all keypoints (each keypoint different color) thresholded at 50%, 75%, 90% and 95% respectivelly. Therefore, second column shows areas where keypoints are with 50% posterior probability (probmap multiplied by presence probability.) Areas expand with higher required confidence until they disapear when presence probability is below required threshold. Occuded keypoints have much larger areas as their confidence is lower.

Figure 12. Images from OCHuman [34]; probability maps thresholded at different posterior probabilities. First column are input images with GT bboxes and estimated poses. Second to fifth columns are probability maps for all keypoints (each keypoint different color) thresholded at 50%, 75%, 90% and 95% respectivelly. Therefore, second column shows areas where keypoints are with 50% posterior probability (probmap multiplied by presence probability.) Areas expand with higher required confidence until they disapear when presence probability is below required threshold. Occuded keypoints have much larger areas as their confidence is lower.

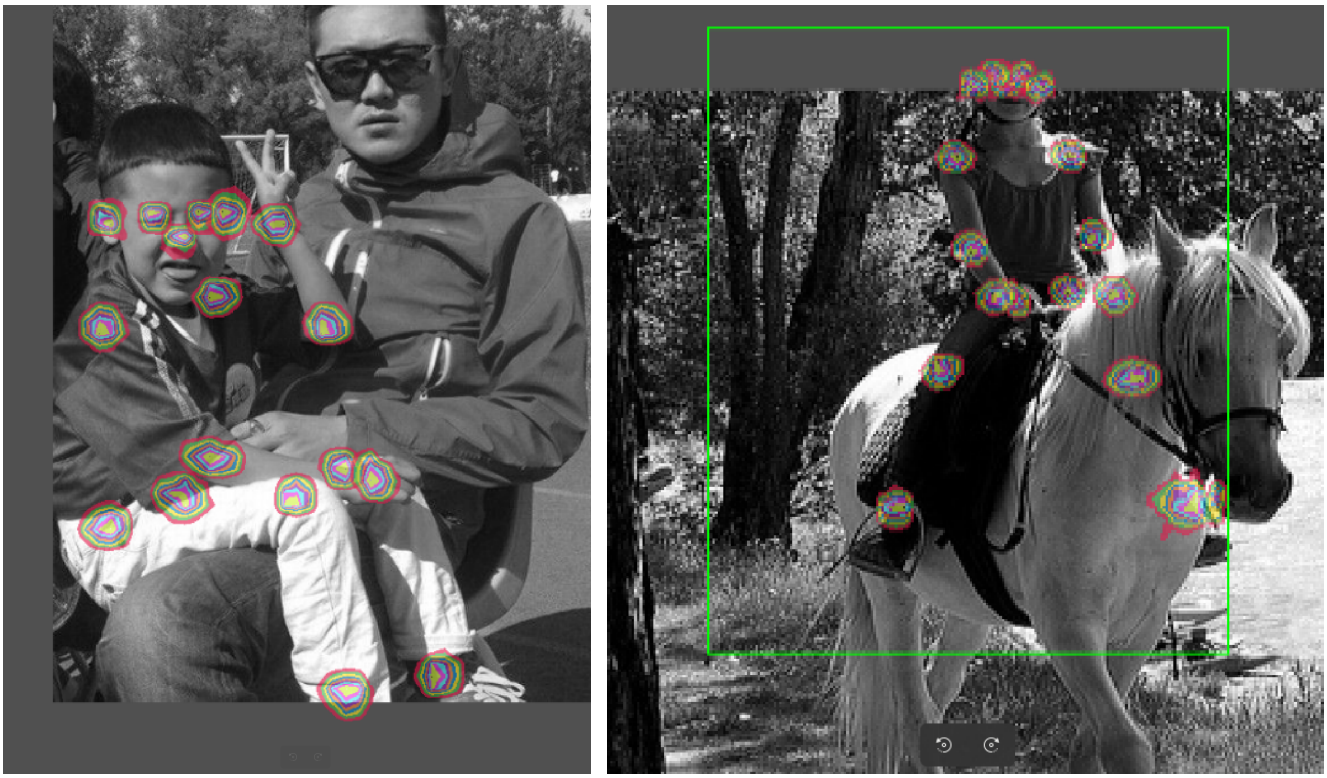


Figure 13. Calibrated probability maps with probability levels thresholded in 10% steps. Each color represents area with 10% probability that the keypoint is in that area. The green area in the right image shows activation window (AW) area. Notice that the probability map is more precise for smaller (face in the left image) and visible (left side of teh right image) keypoints.