ProbPose: A Probabilistic Approach to 2D Human Pose Estimation

Supplementary Material

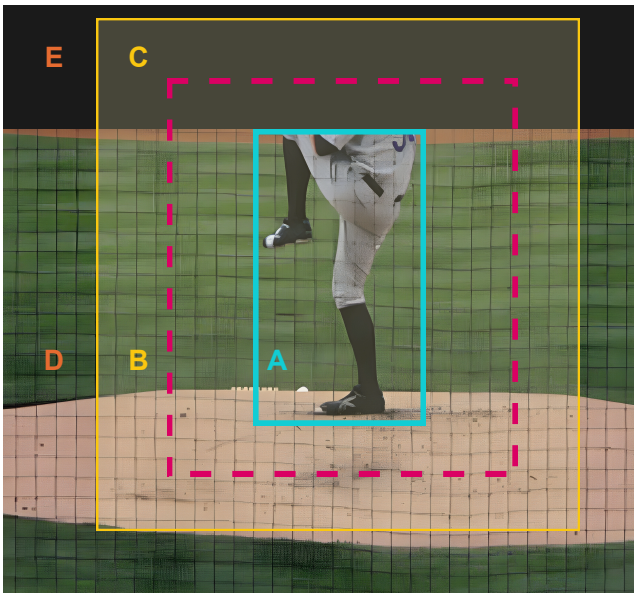


Figure 6. A scheme explaining where keypoints could be in the image. Rectangles represent the bounding box, the model input and the activation window (usually coincides with the model input). Image taken from the COCO val dataset.

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

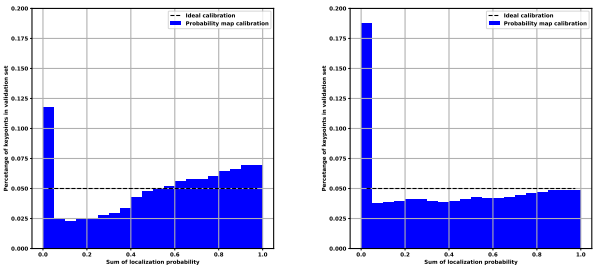
Training probability maps with OKSLoss results in uncalibrated probability maps. Calibration ensures that a probability map accurately refelcts the likelihood of finding a point in a specific area. For example, among all predictions where the model assigns probability 80%, approximately 80% of these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

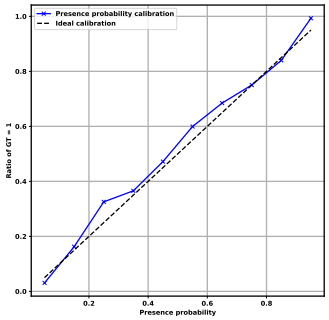
Calibration is performed on a validation set using temperature scaling [8], optimizing a single parameter T. The goal is to create an evenly distributed histogram as shown in Fig. 7a, which shows the calibration curves before and after the temperature scaling. Before calibration, the model was underconfident 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.

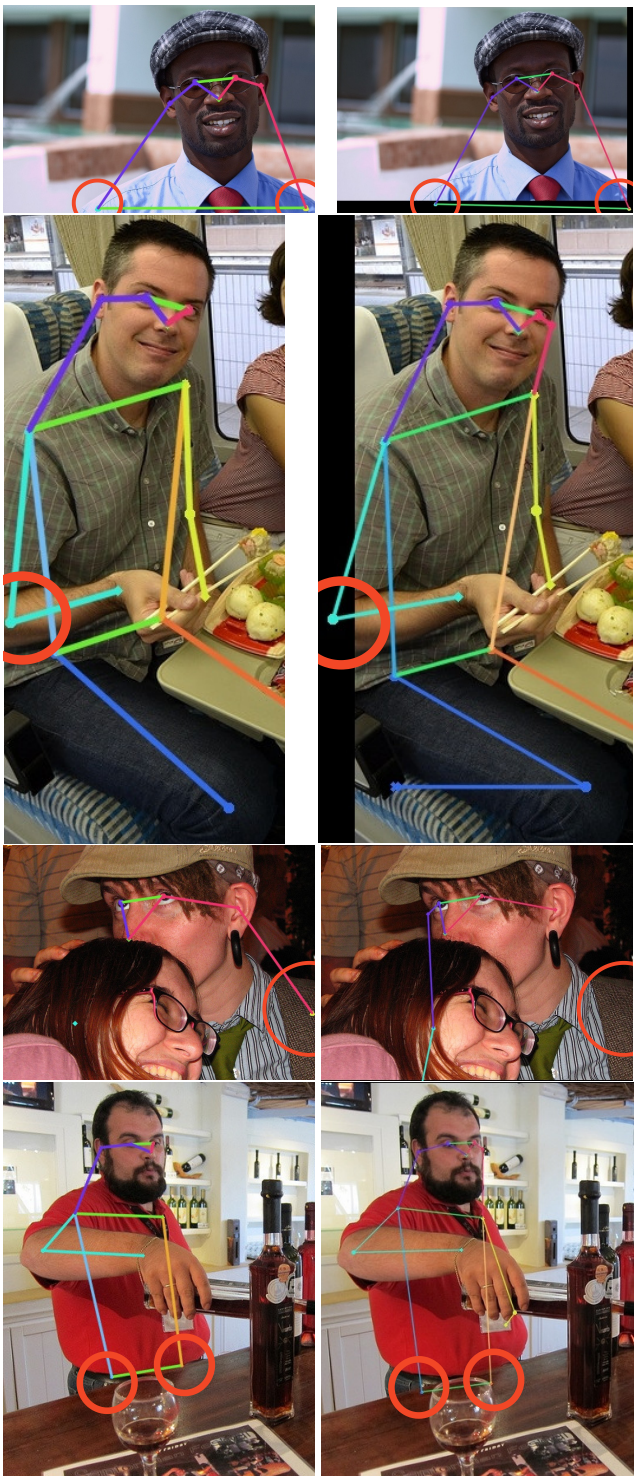


Figure 8. Ground truth annotation (left) vs. ProbPose-s (right) on COCO. Images showcasing dubious annotations along the bounding box border where ProbPose gets penalized even though its input seems better than ground truth. In the third row, our estimate is missing as we correctly predict it outside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

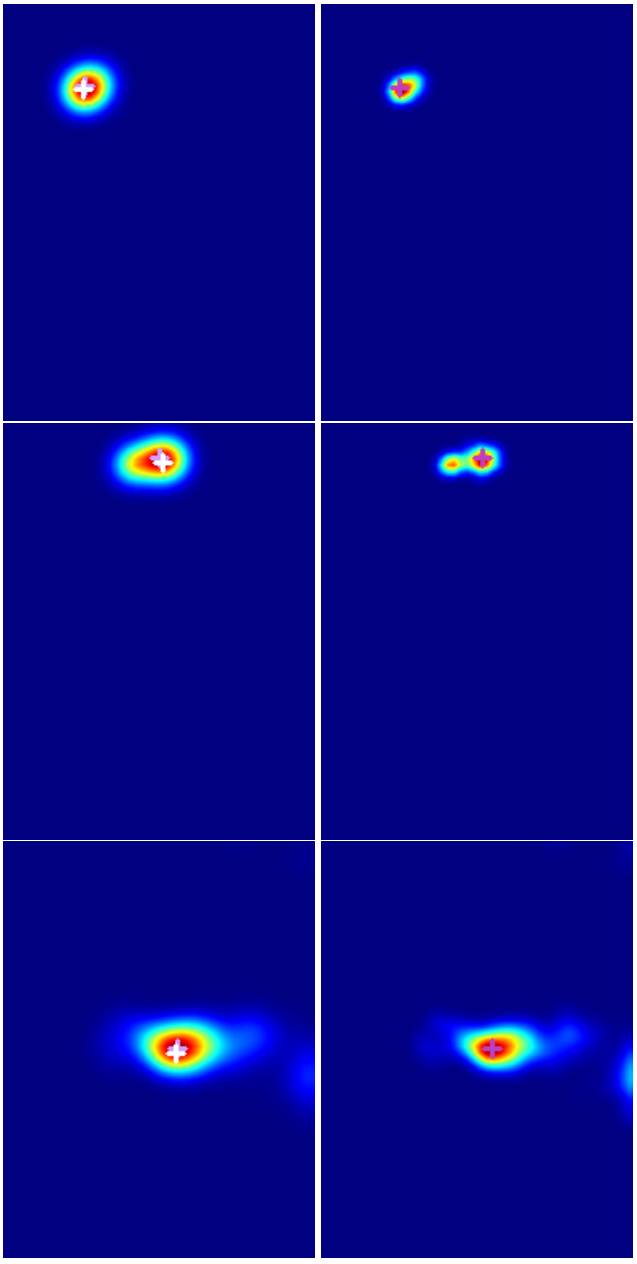


Figure 9. Decoding does not matter as the predicted heatmap is unimodal. Decoding predicted probability maps throught UDP (left) and expected OKS maximization (right). Probability map maximum in white, UDP-refined point in light purple and maximal expected OKS in dark purple. Majority of keypoints have 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.

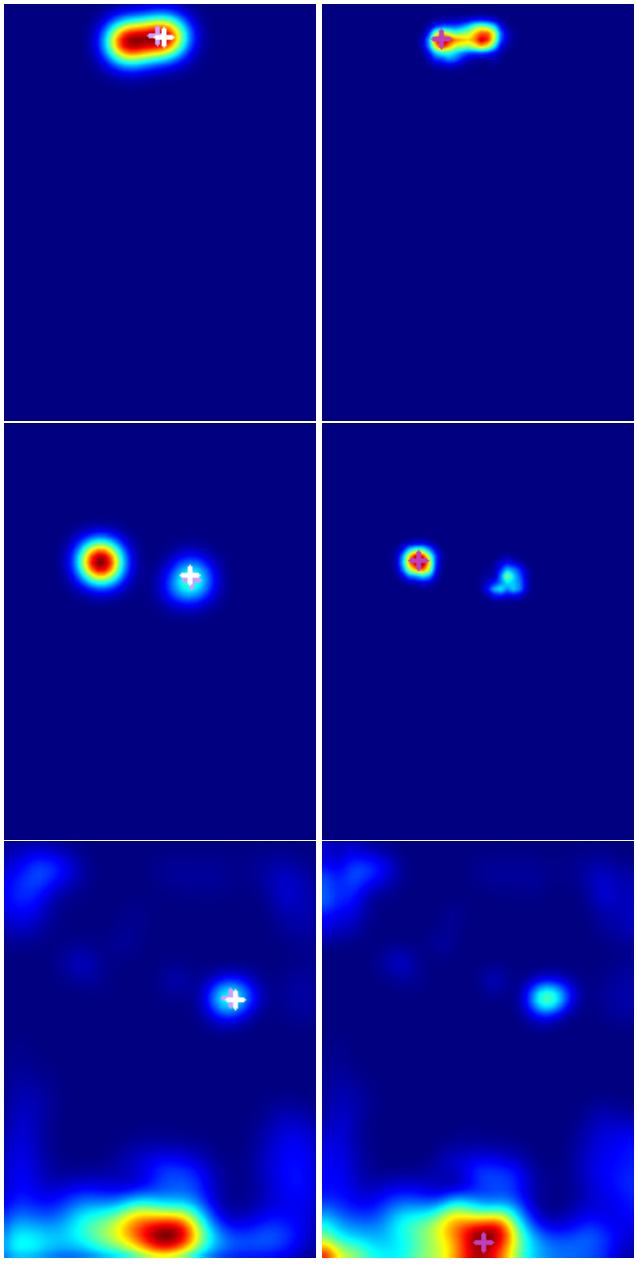


Figure 10. Decoding does matter when distributiona are multimodal. Decoding predicted probability maps throught UDP (left) and expected OKS maximization (right). Probability map maximum in white, UDP-refined point in light purple and maximal expected OKS in dark purple. The predicted heatmaps contain 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.

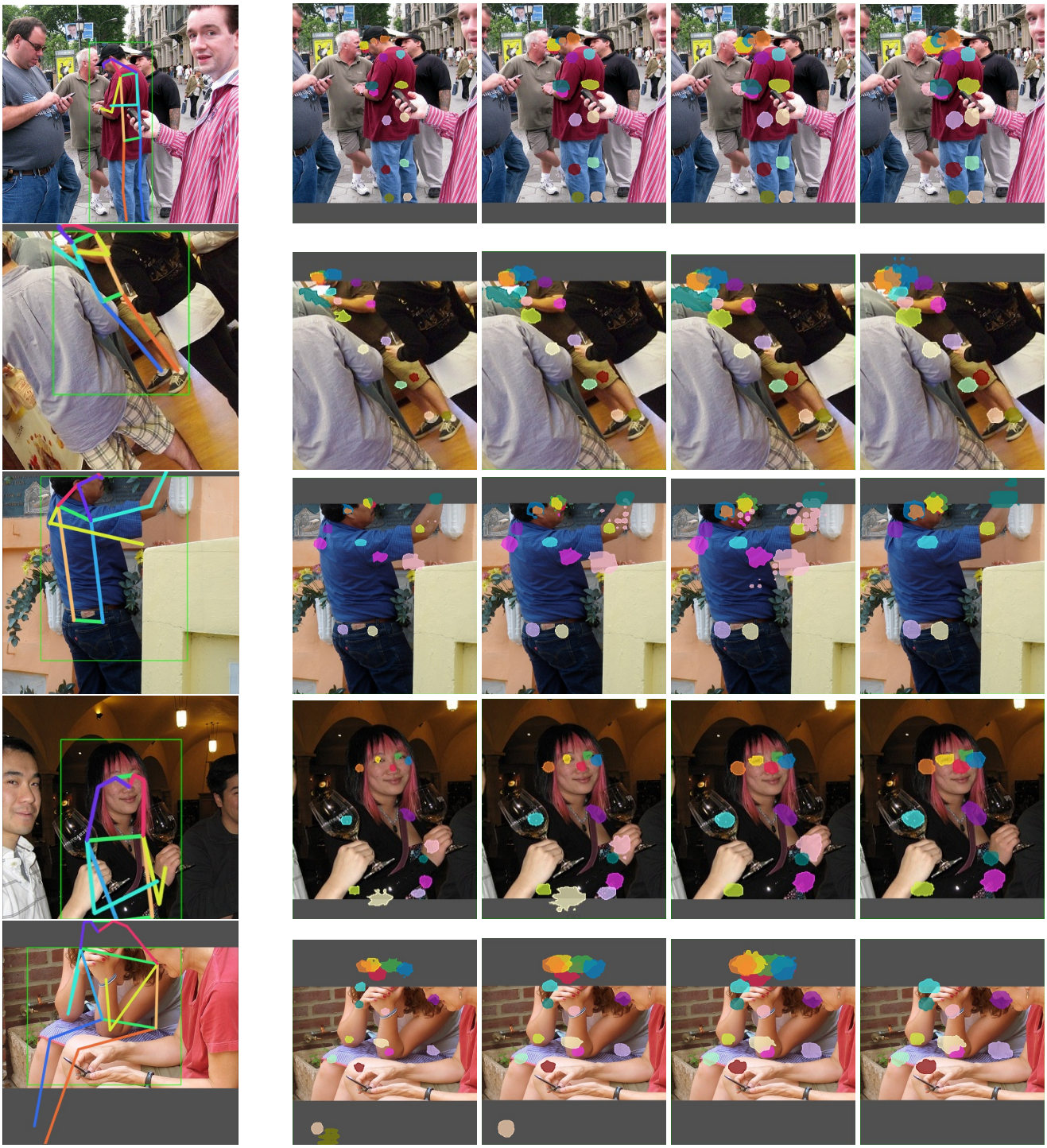


Figure 11. Images from CropCOCO; probability maps thresholded at different posterior probabilities. First column are input images with GT bboxes and estimated poses. Second to fifth columns are probability maps for all keypoints (each keypoint different color) thresholded at 50%, 75%, 90% and 95% respectivelly. 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.

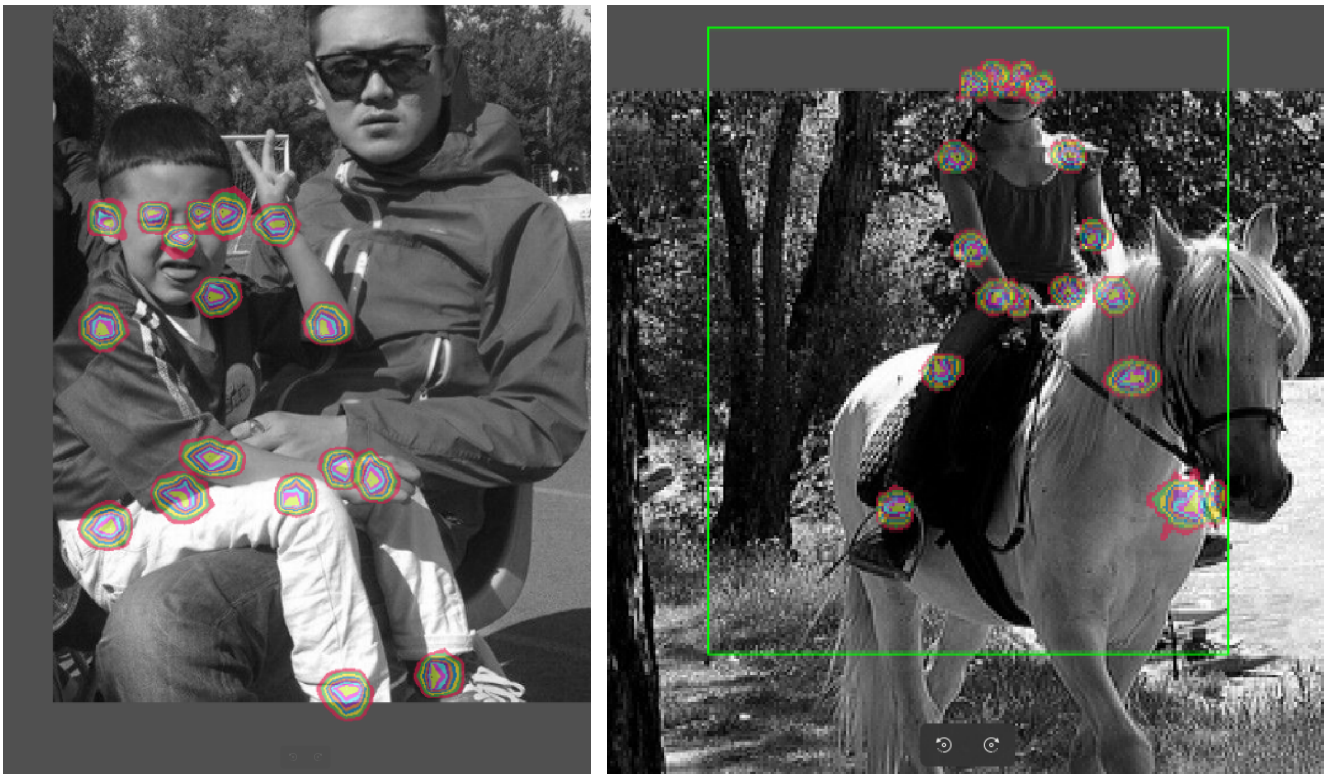


Figure 13. Calibrated probability maps with probability levels thresholded in 10% steps. Each color represents area with 10% probability that the keypoint is in that area. The green area in the right image shows activation window (AW) area. Notice that the probability map is more precise for smaller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