Removing the Bias of Integral Pose Regression Supplementary Material

Kerui Gu[1] Linlin Yang[1,2] Angela Yao[1]
[1]National University of Singapore, Singapore
[2]University of Bonn, Germany
{keruigu, yangll, ayao}@comp.nus.edu.sg

We present the detailed derivation of the bias and some details about experiments. Note that all the notation and abbreviations here are consistent with the main paper.

A. Derivation of Bias

As defined in the main paper, we obtain the normalized heatmap by soft-argmax function as follows:

$$\tilde{\mathbf{H}}(\mathbf{p}) = \frac{\exp(\beta \cdot \mathbf{H}(\mathbf{p}))}{\sum_{\mathbf{p}' \in \Omega} \exp(\beta \cdot \mathbf{H}(\mathbf{p}'))}, \quad \beta > 0, \tag{1}$$

where $\mathbf{H}(\mathbf{p})$ is the heatmap output of the network and indexed by pixel $\mathbf{p}$ over the range of pixels Ω. For convenience, we further define a variable C as the denominator of Eq. (1):

$$C = \sum_{\mathbf{p}' \in \Omega} \exp(\beta \cdot \mathbf{H}(\mathbf{p}')). \tag{2}$$

We can further partition the heatmap's pixels Ω into four sections $\{\Omega_1, \Omega_2, \Omega_3, \Omega_4\}$ as visualized in Fig. 1. Ω_1 is defined such that the true joint location (x_o, y_o) is the expected value and center of the section. The key assumption that we make in our work is that the heatmap support for the true joint location (x_o, y_o) is well localized and fully contained within Ω_1 in $\mathbf{H}$. As such, the sections Ω_2 to Ω_4 contain only zero or near-zero elements so we can approximate Eq. (1) for the four sections as follows:

$$\tilde{\mathbf{H}}(\mathbf{p}) \approx \begin{cases} \frac{1}{C} \cdot \exp(\beta_k \mathbf{H}(\mathbf{p})) & \text{for } \mathbf{p} \in \Omega_1 \\ \frac{1}{C} & \text{for } \mathbf{p} \in \{\Omega_2, \Omega_3, \Omega_4\} \end{cases} \tag{3}$$

where the normalized heat map value approximates to $1/C$ for $\mathbf{p} \in \{\Omega_2, \Omega_3, \Omega_4\}$, since the exponential of a zero in the numerator is simply 1.

The (biased) joint location $\mathbf{J}^r(x_r, y_r)$ is defined as the expected value of the entire heatmap, which can be further decomposed into the four sections:

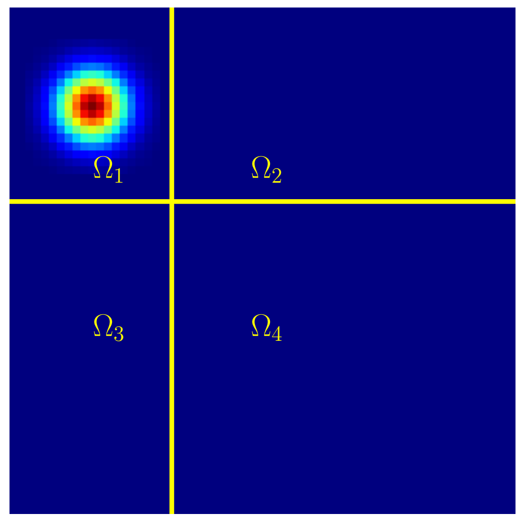

Figure 1: Partitioning of heatmap into four sections to estimate the bias. Ω_1 is assumed to contain the full support and is centered at the true joint location (x_o, y_o), while Ω_2, Ω_3 and Ω_4 are assumed to be (near)-zero values. We illustrate the heatmap with a Gaussian for visualization purposes, but our method does not make any assumption on the form or symmetry of the heatmap density.

$$\mathbf{J}^r = \sum_{\mathbf{p} \in \Omega} \tilde{\mathbf{H}}(\mathbf{p}) \cdot \mathbf{p} \tag{4}$$

$$= \sum_{\mathbf{p} \in \Omega_1} \tilde{\mathbf{H}}(\mathbf{p}) \cdot \mathbf{p} + \sum_{\mathbf{p} \in \Omega_2, \Omega_3, \Omega_4} \tilde{\mathbf{H}}(\mathbf{p}) \cdot \mathbf{p}. \tag{5}$$

We can also view $\mathbf{J}^r = (x_r, y_r)$ as a weighted sum of the expected location of each section:

$$\mathbf{J}^r = w_1 \mathbf{J}_1 + w_2 \mathbf{J}_2 + w_3 \mathbf{J}_3 + w_4 \mathbf{J}_4, \quad \text{where } w_k = \sum_{\mathbf{p} \in \Omega_k} \tilde{\mathbf{H}}(\mathbf{p}), \quad \text{for } k = 1, 2, 3, 4 \tag{6}$$

De/Re	Few kpts	some kpts	many kpts	all
Many occ	28.5/**26.6**	14.1/**13.3**	14.2/**14.7**	16.8/**16.6**
some occ	22.2/**20.5**	**6.72**/6.93	**7.18**/7.20	8.27/**8.20**
few occ	26.9/**26.4**	**7.19**/7.38	**5.12**/5.29	**5.78**/5.95
all	25.8/**24.1**	**7.98**/8.10	**6.74**/6.93	**8.09**/8.14

Table 1: HRNet - Comparisons about EPE on COCO validation set. De and Re refers to detection and regression method respectively.

	Few kpts	some kpts	many kpts	all
M	401	794	1218	2413
L	174	327	747	1248
XL	103	187	492	782
XXL	256	484	1169	1909
all	934	1792	3626	6352

Table 2: number of person instances when separating the benchmarks according to number of present joints and input size.

where $\mathbf{J}_1 = (x_o, y_o)$, $\mathbf{J}_2 = (x_o, y_o + \frac{w}{2})$, $\mathbf{J}_3 = (x_o + \frac{h}{2}, w/2)$, and $\mathbf{J}_4 = (x_o + \frac{h}{2}, y_o + \frac{w}{2})$. Due to the symmetry of each region, we can also represent the weights w_2 to w_4 as an expression of w, h and C.

$$\begin{aligned} w_2 &= \frac{1}{C} \cdot 2x_o(w - 2y_o), \\ w_3 &= \frac{1}{C} \cdot 2(h - 2x_o)y_o, \\ w_4 &= \frac{1}{C} \cdot (h - 2x_o)(w - 2y_o). \end{aligned} \tag{7}$$

We can reformulate Eq. (6) in matrix format:

$$\begin{bmatrix} x_r \\ y_r \end{bmatrix} = \begin{bmatrix} w_1 x_o + w_2 x_o + w_3(x_o + \frac{h}{2}) + w_4(x_o + \frac{h}{2}) \\ w_1 y_o + w_2(y_o + \frac{w}{2}) + w_3 y_o + w_4(y_o + \frac{w}{2}) \end{bmatrix}. \tag{8}$$

Substituting the weights from Eq. (7) into Eq. (8) and with the knowledge that $w_1 = 1 - w_2 - w_3 - w_4$, we arrive at the following linear equation:

$$\mathbf{J}^r = \begin{bmatrix} x_r \\ y_r \end{bmatrix} = \begin{bmatrix} (1 - \frac{hw}{C})x_o + \frac{hw}{C}\frac{h}{2} \\ (1 - \frac{hw}{C})y_o + \frac{hw}{C}\frac{w}{2} \end{bmatrix}. \tag{9}$$

Even though we began our derivation with (x_o, y_o) being located in Ω_1 which is in the upper left quadrant, Eq. (9) is equally applicable when (x_o, y_o) is located in the other three quadrants. If we look at the Eq. (9), if $x_o < \frac{h}{2}$, then $x_r > x_o$ which pushes the coordinate to move towards the center. If $x_o > \frac{h}{2}$, then $x_r < x_o$ which also make the prediction to be closer to the center. y_o is same as x_o. Therefore, this equation is applicable to all quadrants.

Therefore, we can predict $\mathbf{J}^o$ from $\mathbf{J}^r$ in closed form as follows:

$$\mathbf{J}^{ro} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} \frac{C}{C-hw}x_r - \frac{hw^2}{2(C-hw)} \\ \frac{C}{C-hw}y_r - \frac{h^2 w}{2(C-hw)} \end{bmatrix}, \tag{10}$$

which is the result in the main paper.

	Few kpts	some kpts	many kpts	all
Many occ	167	149	52	368
some occ	182	584	602	1368
few occ	585	1059	2972	4616
all	934	1792	3626	6352

Table 3: number of person instances when separating the benchmarks according to number of present joints and percentage of occlusions.

D/R/S	Few kpts	some kpts	many kpts
Many occ	32.0 / **28.1**/28.2	23.7 / 22.8/**22.2**	27.1 / 24.3/**23.8**
some occ	16.0 / 14.8/ **14.4**	6.88/ 7.00/**6.86**	6.78 / 7.18/**6.62**
few occ	16.6 / 15.2/**14.3**	8.36 / 8.53/**7.08**	4.91 / 5.22/**4.84**

Table 4: Comparison of EPE of our method with detection and regression based method on sub benchmarks divided by our proposed method on COCO validation set.

B. Experiment Details

We report the experiment results of HRNet [2]'s performance on different sub-benchmarks in Table 1.

We report the number of person instances in each sub benchmarks divided by the proposed factors on COCO [1] validation set in Table 2 and Table 3.

We also report the detailed EPE of our method on the divided sub benchmarks in Table 4 to support the Fig. 5 in the main paper.

References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2

[2] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 2