Rethinking the Data Annotation Process for Multi-view 3D Pose Estimation with Active Learning and Self-Training
Supplementary Material

Qi Feng, Kun He, He Wen, Cem Keskin, and Yuting Ye

Meta Reality Labs

{fung,kunhe,hewen,cemkeskin,yuting.ye}@meta.com

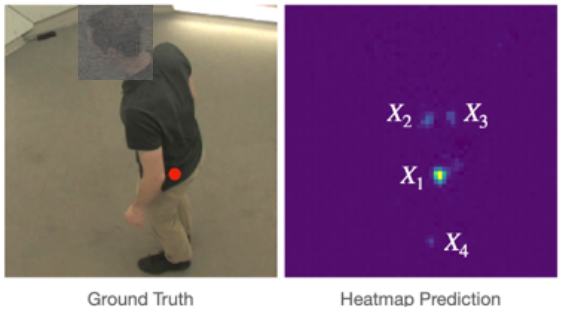


Figure 1: Illustration of entropy-based single-view AL strategies: Best vs. Second Best (BSB) and Multiple Peak Entropy (MPE). Let the normalized predicted heatmap for keypoint p be $\hat{H}^p$, and $L^p = \{l_i^p\}$ be its local peaks. $\mathcal{M}_{\text{BSB}} = \sum_p \hat{H}(l_2^p) - \hat{H}(l_1^p)$ and $\mathcal{M}_{\text{MPE}} = \sum_p \sum_{i=1}^4 -\Pr(l_i^p)\log\Pr(l_i^p)$.

This supplementary material provides details and additional ablation studies of supporting experiments that are not presented in the main paper. In the following, we first present prior works on active learning for single-view human pose estimation problem, which is discussed briefly in Sec. 3. Then, we present the visualizations of the pose clusters used in Fig. 6 in the main paper. Finally, we present our main results in the main paper (Fig. 3, 4 and 5) from a different perspective to give a complete picture of the proposed methods.

1. Prior Works on Single-view AL for Human Pose Estimation

The 2D heatmap representation in our setup for pose estimation naturally lends itself to entropy-based formulations, since a heatmap encodes uncertainty in the model's prediction, and can be normalized into a probability distribution over the 2D grid using the softmax operator. For a predicted heatmap H^k of keypoint k, let $L^k = \{l_1^k, l_2^k, \ldots\}$ be a set of 2D coordinates of local peaks obtained by applying a local maximum filter to $\hat{H}^k$, with l_1^k being the argmax, and so on. In the work of Liu and Ferrari [1], several entropy-based metrics are proposed, and a corresponding AL strategy is defined by sampling top-scoring images under each metric. We now review these metrics. A visual illustration of these metrics is shown in Fig. 1.

1.1. Best vs. Second Best (BSB)

The Best vs. Second Best metric [3] is based on a margin sampling idea, and defined as the difference between the top two local maximums in the heatmap. Intuitively, a smaller difference means larger uncertainty or a multi-modal prediction.

$$\mathcal{M}_{\text{BSB}}(V) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{H}^k(l_1^k) - \hat{H}^k(l_2^k)\right). \quad (1)$$

1.2. Multiple Peak Entropy (MPE)

Multiple Peak Entropy is also introduced by Liu and Ferrari [1] for single-view pose estimation. The idea is that, as modes on a heatmap can be spatially diffuse, simply comparing the highest and second highest would not be able to differentiate between a single wide mode and multiple tight modes. Instead, multiple peaks are considered together to better characterize the uncertainty in a predicted heatmap.

To be concrete, MPE samples H^p at all the local peaks L, and computes the resulting entropy:

$$\mathcal{M}_{\text{MPE}}(V) = \frac{1}{K}\sum_{k=1}^{K}\sum_{l_i^k \in L^k} -\Pr(l_i^k)\log\Pr(l_i^k), \quad (2)$$

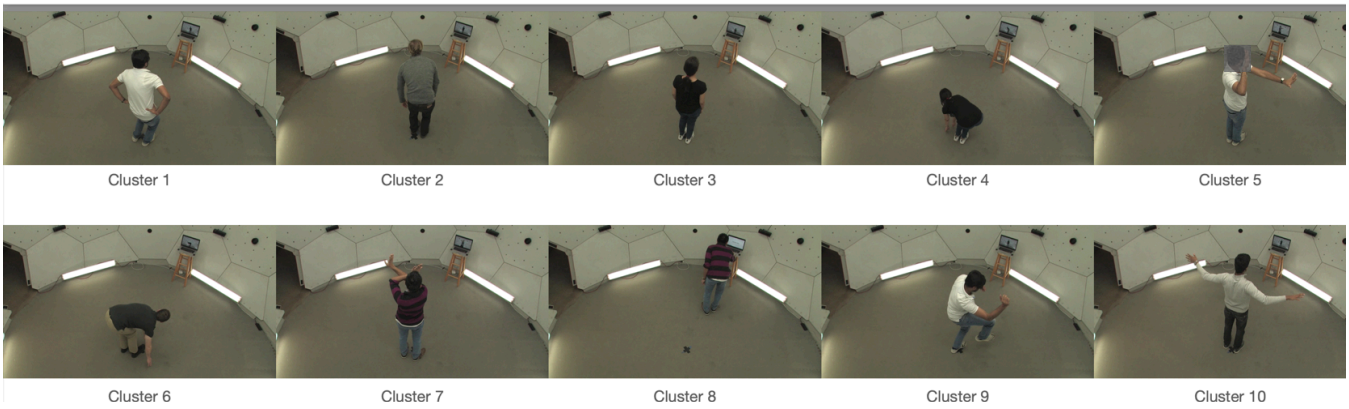
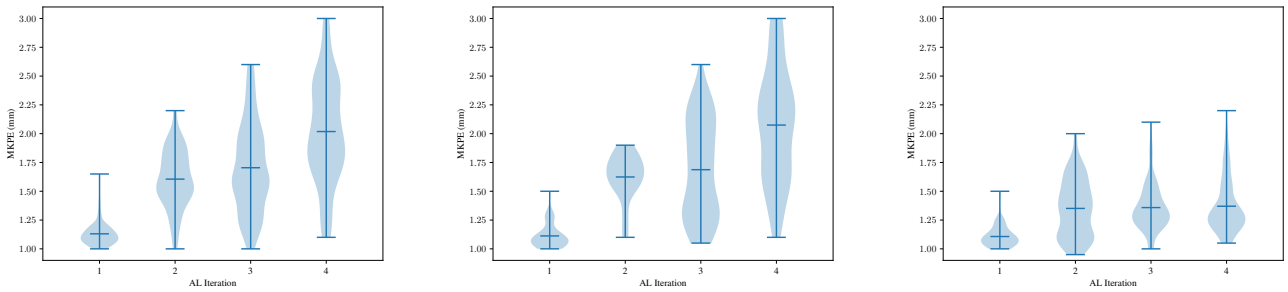


Figure 2: Sample images from each of the 10 pose clusters (Fig 6 in the main paper), obtained by K-means.



(a) Top frames, enlarge $\mathcal{P}$.

(b) Top frames, constant-sized $\mathcal{P}$.

(c) Alternating schedule.

Figure 3: Comparisons between pseudo-labeling strategies: deviation between sampled pseudo-labels and corresponding ground truth, measured in MKPE.

where

$$\Pr(l_i^k) = \frac{\exp H^k(l_i^k)}{\sum_{l_j^k \in L^k} \exp H^k(l_j^k)}. \tag{3}$$

Note that the softmax operator is applied on the sparse set of local peaks only. Liu and Ferrari [1] found that MPE performs better over the random baseline for single-view human pose estimation.

1.3. Random

Random sampling is a simple and very effective baseline strategy in active learning for all kinds of tasks [2, 4]. For pose estimation, random selection of frames from $\mathcal{U}$ ensures that the sampled poses closely follow the training distribution during the AL process.

2. Visualization of 3D Pose

As stated in our main paper, we study the distribution of sampled frames with respect to a discrete clustering of ground truth poses. In Fig. 6 of the main paper, we have visualized the distribution of frames sampled by each AL strategy, along with the entropy values computed from the discrete distributions. The ground truth 3D poses are shifted in 3D to have keypoint 2 (waist) at origin, and then we use K-means to cluster them into 10 clusters. Sample images from each cluster are visualized in Fig 2. The visualization confirms the findings that the proposed OURS-MC samples frames with better diversity in poses (higher entropy), especially in the early iterations.

3. Ablation Studies on Self-Training and Augmentation

3.1. Ablation Studies on Self-Training

In addition to the differences between the proposed self-training algorithm and the multi-view bootstrapping method [5] mentioned in the main paper, self-training produces new and more accurate pseudo-labels as the amount of human-annotated data increases with the AL iteration. Here, we detail the design choices for our specific self-training strategy.

We have considered the following three strategies. In

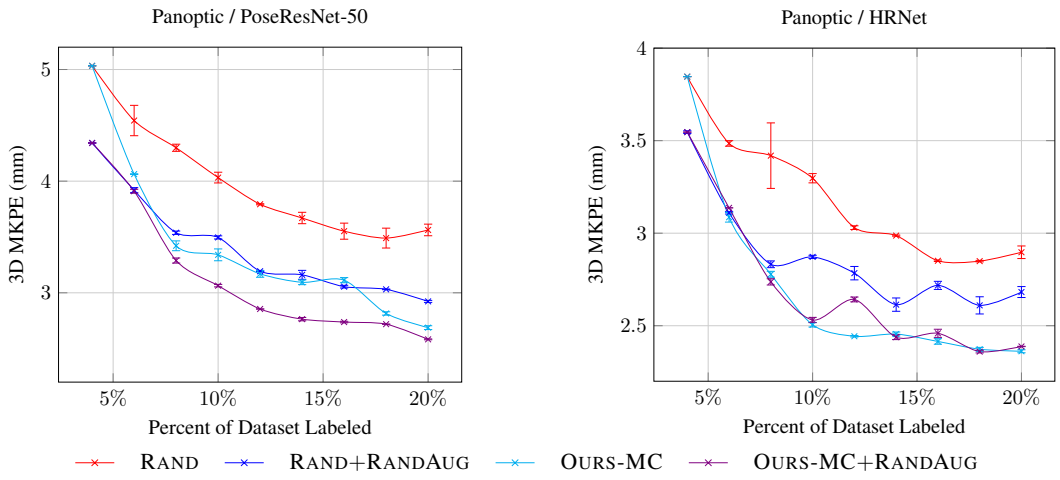


Figure 4: Effects of data augmentation using RandAugment on CMU Panoptic. OURS-MC achieves better label efficiency than RAND + AUG without data augmentation.

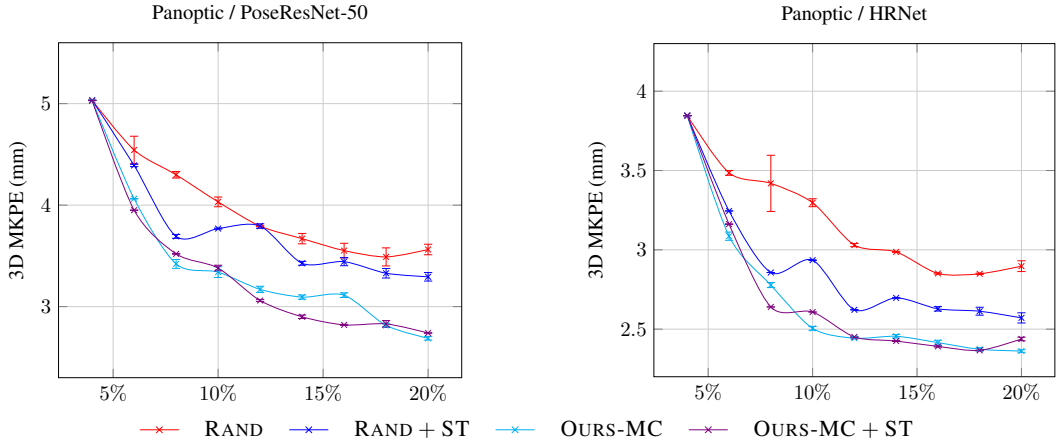


Figure 5: AL + self-training (ST) on CMU Panoptic *without* data augmentation. X-axis: percent of dataset labeled. When combined with AL, our automated self-training strategy enables additional label efficiency gains at no extra computational cost, especially for RAND and during the early stages of training. Best viewed in color.

Fig. 3, we plot the distributions of MKPE between pseudo-labels in $\mathcal{P}$ and their corresponding ground truth, over the first four AL iterations.

1. Fig. 3a. Enlarge $\mathcal{P}$ in each AL iteration with the top pseudo-labeled frames. (Selection criterion is discussed in the main paper.)

2. Fig. 3b. Keep the size of $\mathcal{P}$ constant, and pick the top pseudo-labeled frames in each AL iteration.

3. Fig. 3c. Alternating schedule (described in the paper): in each AL iteration i, pick top frames that are not already in $\mathcal{P}_{i-1}$ from the last iteration, to form $\mathcal{P}_i$ for the current iteration.

To begin with, the first strategy is easily susceptible to label drifting, as more and less accurate pseudo-labels would enter $\mathcal{P}$ and pollute the training set over time. Somewhat surprisingly, the second strategy of keeping the size of $\mathcal{P}$ constant does not work either. We have empirically verified that, in this scenario, the set of frames selected to form $\mathcal{P}$ is very stable across iterations. Then, with every passing iteration, this strategy essentially re-labels a same set of frames, using a new model trained on a training set containing them, and the errors would accumulate. Note that in this case, the model needs to achieve zero training error on $\mathcal{P}_i$ in every AL iteration i for the pseudo-labels to remain the same, let alone improve.

Lastly, we found the alternating schedule to be robust against label drifting. In each iteration i, all frames in $\mathcal{P}_{i-1}$

are evicted, and prevented from re-entering until the next iteration. This effectively avoids the above error accumulation problem, as a model trained on frames from $\mathcal{P}_i$ (among others) is never used to infer pseudo-labels on the same set of frames.

3.2. Ablation Studies on Data Augmentation

We present the effect of RandAugment on CMU Panoptic in Fig. 4 that is presented separately in Fig. 4 and Fig. 5 in the main paper. For each training image, we randomly apply two of the following augmentation operations:

- `Rotate` (within $\pm 30°$)
- `AutoContrast`
- `Equalize`
- `Invert`
- `Posterize`
- `Solarize`
- `Color`
- `Contrast`
- `Brightness`
- `Sharpness`

In the case of image rotation, we also rotate the target heatmap by the same amount. All other operations are label-preserving and do not alter the heatmap.

Overall, we find that the choice of AL strategy outweighs data augmentation. For example, OURS-MC without data augmentation even outperforms RAND + RANDAUG with PoseResNet-50 backbone. For OURS-MC with the HRNet backbone, the performance gain from data augmentation gets smaller as it saturates more quickly towards the fully-supervised baseline.

Additionally, we compare performances of self-training on RAND and OURS-MC without RandAugment and present the comparison in Fig. 5 to complement Fig. 5 in the main paper. Self-training suffers from the fact that no augmentation is used in these experiments and only provides marginal gains, with one exception to RAND with HRNet, where self-training shows a slightly larger gain.

References

[1] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 4363–4372, 2017. 1, 2

[2] Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with illusions about deep active learning, 2019. 2

[3] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer, 2006. 1

[4] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[5] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1145–1153, 2017. 2