Supplementary Materials: Denoising and Selecting Pseudo-Heatmaps for Semi-Supervised Human Pose Estimation

Zhuoran Yu*†, Manchen Wang*‡, Yanbei Chen, Paolo Favaro, Davide Modolo
AWS AI Labs
zhuoran.yu@wisc.com, {manchenw,yanbec,pffavaro,dmodolo}@amazon.com

1. Additional Ablation Study

1.1. Pseudo-heatmaps accuracy

As described in Section 3.2, we employ the winner-take-all strategy to compute the ensemble pseudo-heatmaps by taking the maximum score per pixel over outputs of all augmented views. In Table 1, we compare pseudo-heatmaps accuracy calculated with Percentage of Correct Keypoint (PCK) when using average or maximum operation to ensemble multi-view augmentations. We find that winner-take-all yields a substantially higher performance (66.0 VS 63.9) by considering human pose characteristics: A clean pseudo-heatmap should have one strong peak. However, average operation often results in multiple weak peaks (see Figure 1), leading to a noisier learning target and poorer performance. On the other hand, maximum operation removes spurious noise by selecting the most confident peak.

Aggregate	None	Average	Maximum
PCK	63.5	63.9	**66.0**

Table 1. **Comparison of pseudo-heatmaps accuracy when using different aggregations.** Taking the maximum response generates the most accurate pseudo-heatmaps.

1.2. Teacher-Student VS Dual-Student

As described in Section 3, we employ a dual-student framework [2] which trains two student networks jointly and uses the output from one network as the learning targets to supervise the other. In Table 2, we show the comparison between our method and the adaptation of a teacher-student network Unbiased Teacher [1] to the human pose estimation task. We follow Unbiased Teacher to evolve both Teacher and Student models via mutual learning scheme, where Teacher model generates pseudo-heatmaps to train the Student model, and the knowledge that the Student model learned is then transferred to Teacher model via exponential moving average (EMA) on network weights. Results show that dual-student achieves significantly better results than teacher-student. This justifies why we use dual-student as baseline throughout this work. Furthermore, adding Threshold-and-Refine scheme to denoise the pseudo-heatmaps always improves the model performance in the teacher-student and dual-student frameworks. This suggests our design is agnostic to the underlying semi-supervised learning frameworks.

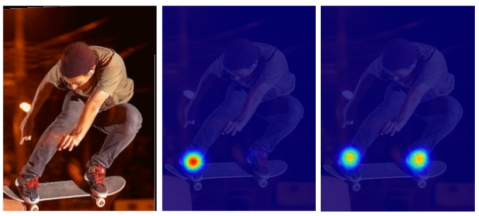

Figure 1. **Pseudo-heatmaps visualization for right ankle by using the maximum and average ensemble.** Maximum operation generate a cleaner and more accurate pseudo-heatmap.

Row	Method	Threshold-and-Refine	mAP
1	Teacher-Student [1]		32.83
2	Teacher-Student [1]	✓	36.53
3	Dual-Student [2]		42.67
4	Dual-Student [2]	✓	**44.13**

Table 2. **Comparison between Teacher-Student and Dual-Student framework.** Results on COCO-Partial protocol using 1K labeled images.

2. Pseudo Code

We provide the pseudo code of pseudo heatmap denoising and uncertainty-guided selection in Algorithm 1.

*These authors contributed equally to this work.
†Currently at The University of Wisconsin–Madison. Work conducted during an internship with AWS AI Labs.
‡Corresponding author.

Algorithm 1 Pseudo Heatmap Denoising and Uncertainty-guided Selection

Input: Batch of unlabeled examples $U = (u_b; b \in (1, ..., B))$, two student networks θ and ξ, number of multi-view augmentations K, number of joints J, uncertainty margin Δ, weak augmentation strength S_w, strong augmentation strength S_s, multi-view augmentation strength S_m

for $b \leftarrow 1$ to B **do**
  $T_w, I_w \leftarrow$ Augment(u_b, S_w) ▷ Apply weak augmentation to u_b
  $T_s, I_s \leftarrow$ Augment(u_b, S_s) ▷ Apply strong augmentation to u_b
  $H_{\theta,w} \leftarrow \theta(I_w)$ ▷ Predict heatmaps for weakly and strongly augmented images
  $H_{\xi,w} \leftarrow \xi(I_w)$
  $H_{\theta,s} \leftarrow \theta(I_s)$
  $H_{\xi,s} \leftarrow \xi(I_s)$
  $H_{\theta,w} \leftarrow$ Transform($H_{\theta,w}$, $(T_w)^{-1}$) ▷ Transform weakly augmented pseudo heatmap back to original view
  $H_{\xi,w} \leftarrow$ Transform($H_{\xi,w}$, $(T_w)^{-1}$)
  for $k \leftarrow 1$ to K **do**
    $T_{s,k}, I_{s,k} \leftarrow$ Augment(u_b, S_m) ▷ Apply k^{th} round of multi-view augmentation to u_b
    $H^k_{\theta,s} \leftarrow \theta(I_{s,k})$ ▷ Predict heatmaps on k^{th} multi-view augmented image I^k_s
    $H^k_{\xi,s} \leftarrow \xi(I_{s,k})$
    $H^k_{\theta,s} \leftarrow$ Transform($H^k_{\theta,s}$, $(T_{s,k})^{-1}$) ▷ Transform multi-view augmented pseudo heatmap back to original view
    $H^k_{\xi,s} \leftarrow$ Transform($H^k_{\xi,s}$, $(T_{s,k})^{-1}$)
  end for
  $P_\theta \leftarrow \max\{H_{\theta,w}, \{H^k_{\theta,s}\}^K_{k=1}\}$ ▷ Ensemble pseudo-heatmaps
  $P_\xi \leftarrow \max\{H_{\xi,w}, \{H^k_{\xi,s}\}^K_{k=1}\}$
  $U_\theta \leftarrow$ stdev $\{H_{\theta,w}, \{H^k_{\theta,s}\}^K_{k=1}\}$ ▷ Estimate uncertainty as pixel-wise standard deviation
  $U_\xi \leftarrow$ stdev $\{H_{\xi,w}, \{H^k_{\xi,s}\}^K_{k=1}\}$
  $u_\theta \leftarrow \{\max U^j_\theta\}^J_{j=1}$ ▷ Take the maximum value on uncertainty maps for every joint
  $u_\xi \leftarrow \{\max U^j_\xi\}^J_{j=1}$
  $\hat{P}_{\theta,j} \leftarrow P_{\theta,j}$ **if** $u_{\theta,j} + \Delta < u_{\xi,j}$ **else** $P_{\xi,j}$ ▷ Learning targets selected by uncertainty
  $\hat{P}_{\xi,j} \leftarrow P_{\xi,j}$ **if** $u_{\xi,j} + \Delta < u_{\theta,j}$ **else** $P_{\theta,j}$
  $\mathcal{L}^U_\theta = ||H_{\theta,s} -$ Transform$(\hat{P}_\xi, T_s)||^2$ ▷ Compute unsupervised loss on unlabeled data
  $\mathcal{L}^U_\xi = ||H_{\xi,s} -$ Transform$(\hat{P}_\theta, T_s)||^2$
end for

3. Additional qualitative results

We provide more qualitative results of predicted human pose on COCO val2017 dataset in Figure 2 to demonstrate the strong performance of our model.

References

[1] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 1

[2] Rongchang Xie, Chunyu Wang, Wenjun Zeng, and Yizhou Wang. An empirical study of the collapsing problem in semi-supervised 2d human pose estimation. In *ICCV*, 2021. 1

Figure 2. **Qualitative results** of some example images in COCO dataset.