Supplementary Materials: Towards Labeling-free Fine-grained Animal Pose Estimation

Anonymous Author(s)
Submission Id: 2405

This appendix can be divided into these parts:

- Section 1 introduces the details of augmentations in FreeNet.
- Section 2 shows the confidence scores of unannotated joints under different feedback thresholds $\alpha_{feedback}$.
- Section 3 gives the comparing results between the Small-loss strategy and FreeNet.
- Section 4 discusses the limitations of FreeNet and gives some failure cases.
- Section 5 presents more visual analysis for FreeNet.

1 THE AUGMENTATION DETAILS

All animal images are first preprocessed to obtain two versions: the original and a color-augmented version. For color augmentations, we apply the RandAugment method [1], including translation, shearing, and rotation. Next, both versions are randomly subjected to further augmentations, including half-body cropping (30% probability), scaling (0.65-1.35), rotation (from -45° to 45°), and horizontal flipping (50% probability).

2 ANALYSIS OF THRESHOLD ($\alpha_{feedback}$)

We propose feedback learning to ensure the unannotated joints are learned effectively, which is crucial for enhancing the fine-grained APE. Figure 7 in the main paper demonstrates the confidence score for different animal body parts (i.e., head joints, frontal body joints, back body joints, and unannotated joints) across various feedback learning thresholds. Here we delve into the impact of $\alpha_{feedback}$ threshold on five specific unannotated joints in Figure 1 and 2. These joints include the “Neck” from AP-10k, “Left ear”, “Right ear”, “Throat”, and “Wither” from AnimalPose. We can observe that, compared to threshold $\alpha_{feedback}$ (0,80%), the (20%,80%) threshold more effectively facilitates the learning of unannotated joints. Specifically, “Wither”, a hard-to-detect joint, is primarily selected, while “Left ear” and “Right ear” which are easier to detect, are chosen less frequently (see Figure 2). During training, FreeNet prioritizes label quality (fewer numbers) for easily detectable unannotated joints and focuses on label quantity for those more challenging to detect.

3 COMPARED WITH SMALL-LOSS CRITERION

We also compare our method on the 10% combined dataset (AP-10k and AnimalPose Dataset) with the small-loss strategy used in semi-supervised learning tasks [2]. For small-loss, we choose to keep the 50% of pseudo labels (Gradually decrease from 100% to 50%) in the current batch with the minimum loss. The results Table 1 shows that our method performs better. The statistical results Table 2 on the numbers of joints in each part of the pseudo-labels also confirm the effectiveness of the body part-aware sampling method, while the small-loss strategy exacerbates the imbalance of the number of joints in each part of the pseudo labels.

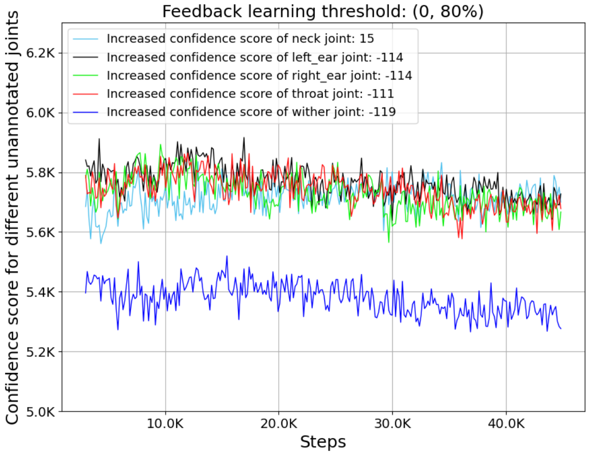

Figure 1: The confidence scores of unannotated joints under feedback learning with $\alpha_{feedback}$ (0%,80%).

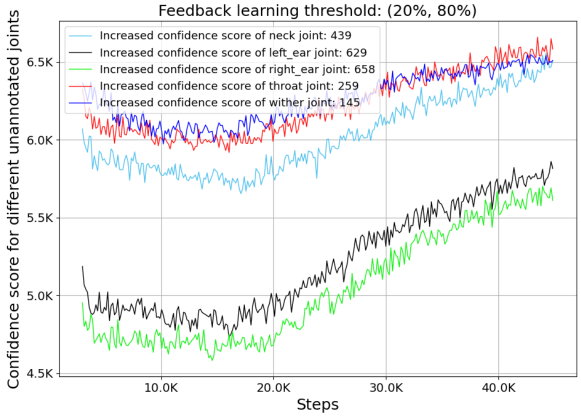

Figure 2: The confidence scores of unannotated joints under feedback learning with $\alpha_{feedback}$ (20%,80%), which largely improves the prediction confidence for unannotated joints.

Table 1: Performance of small-loss method and ours on 10% combined dataset (AP-10k and AnimalPose).

Method	mAP(%)	PCK@0.05(%)
Small-loss	56.62	70.68
Ours	**57.26**	**71.36**

Table 2: Joints statistical results of pseudo labels during the training of two compared methods.

Methods	Stage	Joint Percentage (%)			std (%)
		head	front	back	
Small-loss	start	38.23	34.33	27.42	4.470
	middle	41.37	34.17	24.45	6.961
	end	40.72	34.27	24.99	6.444
Ours	start	38.24	34.33	27.41	4.477
	middle	34.16	34.47	31.35	1.403
	end	34.78	35.09	30.12	2.273

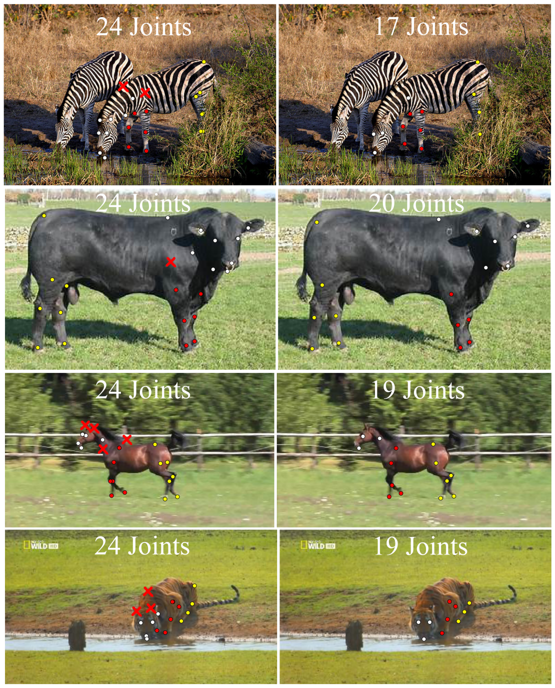

Figure 3: Visualisation of predicted results.We use red crosses to mark the unannotated joints that are not detected.

4 VISUALIZATION OF FAILURE CASES

Our method effectively identifies denser joints from a few non-standard annotations at no additional cost. However, we also found some bias in our model's predictions. We found that the model for images from AP-10k and AnimalPose sometimes misses TigDog's unannotated joints. Also, the model sometimes fails to detect AnimalPose's unannotated joints for images from TigDog.We give some examples in Figure3 with the prediction result on the left and the ground truth on the right.

There are two reasons for this phenomenon. The first is that the resolution of the images in different datasets varies greatly. For example, images in TigDog have relatively small resolutions. The second is the significant species difference between different datasets. TigDog only has tigers and horses, which means we cannot learn the features of unannotated joints from the other datasets.

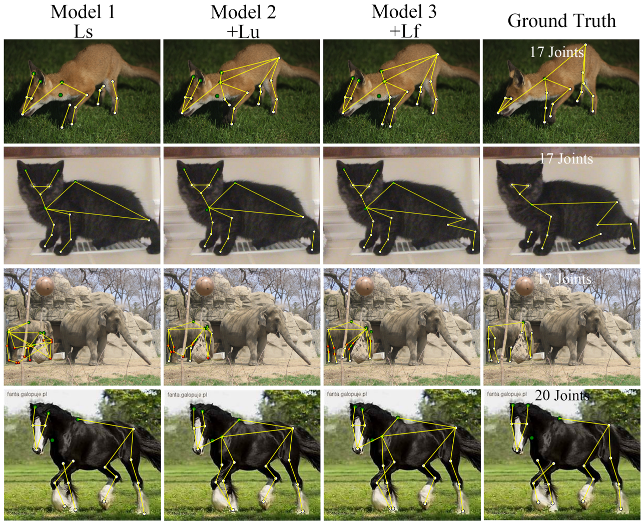

Figure 4: FreeNet can predict accurate joints, including those that are not originally presented in the Ground Truth.

5 MORE VISUAL ANALYSIS

Figure 4 presents examples of predicted landmarks generated by different baselines on the combined datasets (AP-10k and AnimalPose). Consistent with the main paper, FreeNet performs better on harder-to-predict joints, such as those on the rear half of the body and unannotated joints. The first three columns show the prediction results using different loss functions, demonstrating the effectiveness of body part-aware sampling and feedback learning. The fourth column shows the ground truth. FreeNet consistently generates accurate joints despite animal pose variations, including those not initially present in the Ground Truth.

In Figure 5, we present more example images illustrating the scalability of joints, which are 17, 20, 19, 21, and 26, respectively. These correspond to the semantic joint definitions in AP-10k, AnimalPose, and TigDog, two combined datasets (AP-10k and AnimalPose), and three combined datasets (AP-10k, AnimalPose, and TigDog). This illustrates that FreeNet can facilitate fine-grained APE without manual annotations, and its application can be extended to denser joints if more datasets are involved.

REFERENCES

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 702–703.

[2] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems* 31 (2018).

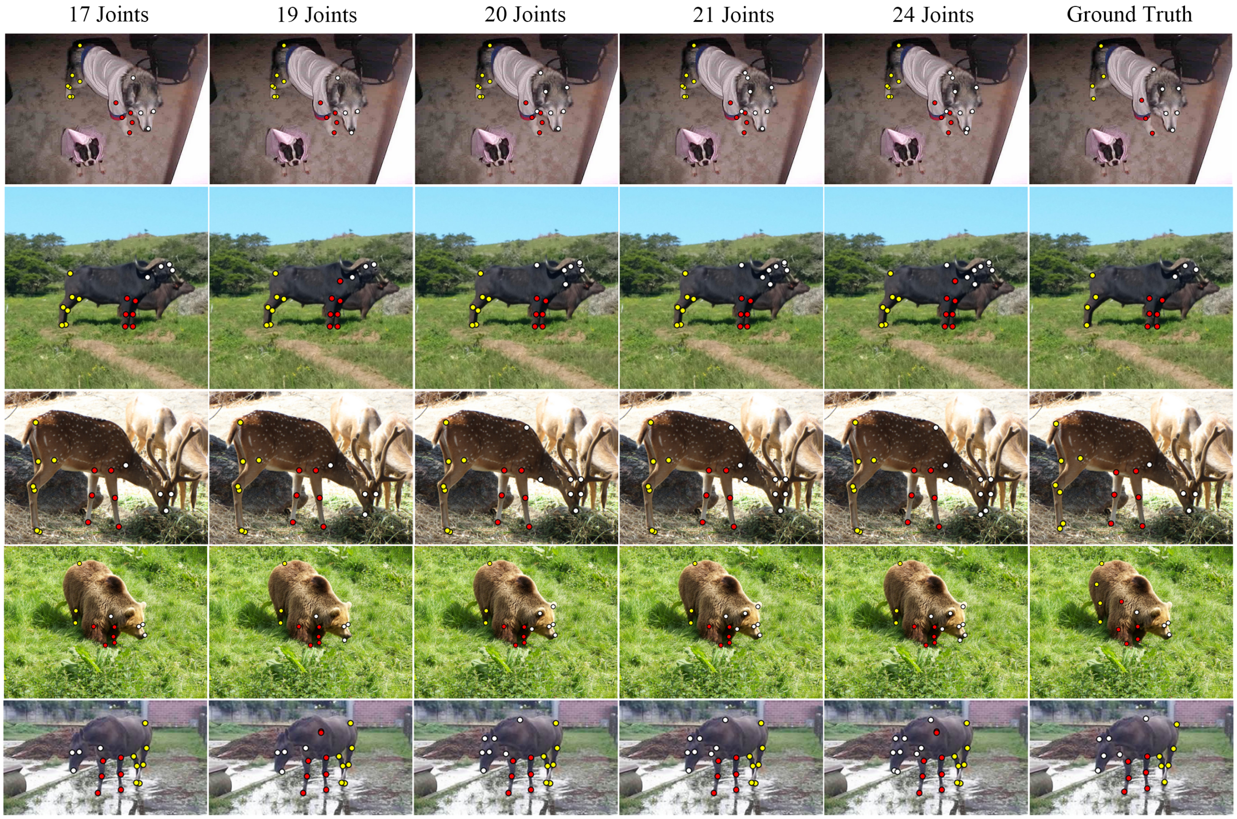

Figure 5: FreeNet can predict fine-grained pose landmarks without additional manual annotations.