Supplementary Material for ScarceNet: Animal Pose Estimation with Scarce Annotations

Chen Li Gim Hee Lee
Department of Computer Science, National University of Singapore
lichen@u.nus.edu gimhee.lee@comp.nus.edu.sg

Details on how to decide the threshold loss value. We select the reliable pseudo labels based on the pseudo label based loss (Eqn. (2) in the main paper), *i.e.* only samples with loss smaller than the threshold l_r are be used. Note that it is difficult to directly set l_{r} since $\mathcal{L}_{\mathrm{p}}^{i}$ does not fall in a specific range, and therefore we decide l_{r} based on a percentile score. Specifically, given a batch of images as input, we first compute the pseudo label based loss for all joints $\mathcal{L}_{\mathrm{p}}^{B} = \{\mathcal{L}_{\mathrm{p}}^{i}|_{i=1}^{B\times J}\}$, where B denotes the batch size. The threshold loss value is then computed as:

$$l_{\mathrm{r}} = \mathrm{percentile}(\mathcal{L}_{\mathrm{p}}^{B}, c), \tag{1}$$

where $\mathrm{percentile}(\cdot,\cdot)$ returns the value of the c^{th} percentile.

Strong-weak augmentation. We apply two sets of augmentations in our framework as described in our main paper. Specifically, we leverage RandAugment [1] as the strong augmentation $\mathcal{P}$, where we remove the translation and shearing augmentation since the object needs to be at the center of the image for pose estimation. For the weak augmentation $\mathcal{T}$, we apply the widely used augmentation operations in human pose estimation, including rotation $(-20^{\circ} - 20^{\circ})$ and scaling $(0.9 - 1.1)$.

Network details. We adopt the HRNet-w32 as our backbone with an input size of $256 \times 256 \times 3$. The HRNet maintains high-resolution representation through the whole network in order to learn more accurate features for pose estimation. This high-resolution feature is fed into a linear prediction layer to regress joint locations after fusing with features at lower resolutions. In our framework, we replace the single-branch prediction layer with a multi-branch prediction layer, where each branch consists of a Bottleneck residual block [2] followed by a linear layer. The output channel size of the last linear layer is one because we only regress the heatmap for one joint in each branch.

Training details. We implement our network in Pytorch and the parameters are optimized using the Adam [3] optimizer with the default parameters. We first train the MBHR-Net on the labeled data with the supervised loss. Following [6], we train the network with the Adam optimizer for 210 epochs. The initial learning rate is set to 10^{-3}, and is dropped to 10^{-4} and 10^{-5} at the 170^{th} and 200^{th} epochs, respectively. We then generate pseudo labels with the pre-trained model following [5]. We remove samples with low confidence score, and apply the small-loss trick to select reliable pseudo labels from the remaining ones. The distance threshold in the agreement check is empirically set to 0.6 in our experiments. The weights for different loss terms are also set empirically as $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 1$ and $\lambda_4 = 2$. The training of the first stage takes 2-5 hours depends on the number of the labeled data. The second stage takes around 21 hours on two RTX 2080Ti GPUs.

Training details for the domain adaptation setting. We test the effectiveness of our approach in the domain adaptation setting in the main paper. We replace the MBHR-Net in our framework with the ResNet backbone [5] for fair comparison. A domain discriminator [4] is also applied in the feature space to facilitate the knowledge transfer from synthetic to real images. Note that our approach and CC-SSL [5] has simpler network compared to MDAM-MT [4], which introduces an extra refinement module for self-knowledge distillation.

More qualitative results. We show qualitative results for more animal species under the 25 labels per animal species setting in the first three rows of Fig. 1. We can see we that our network is able to estimate the pose accurately for diverse animal species when only 25 images for each species are labeled.

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF*

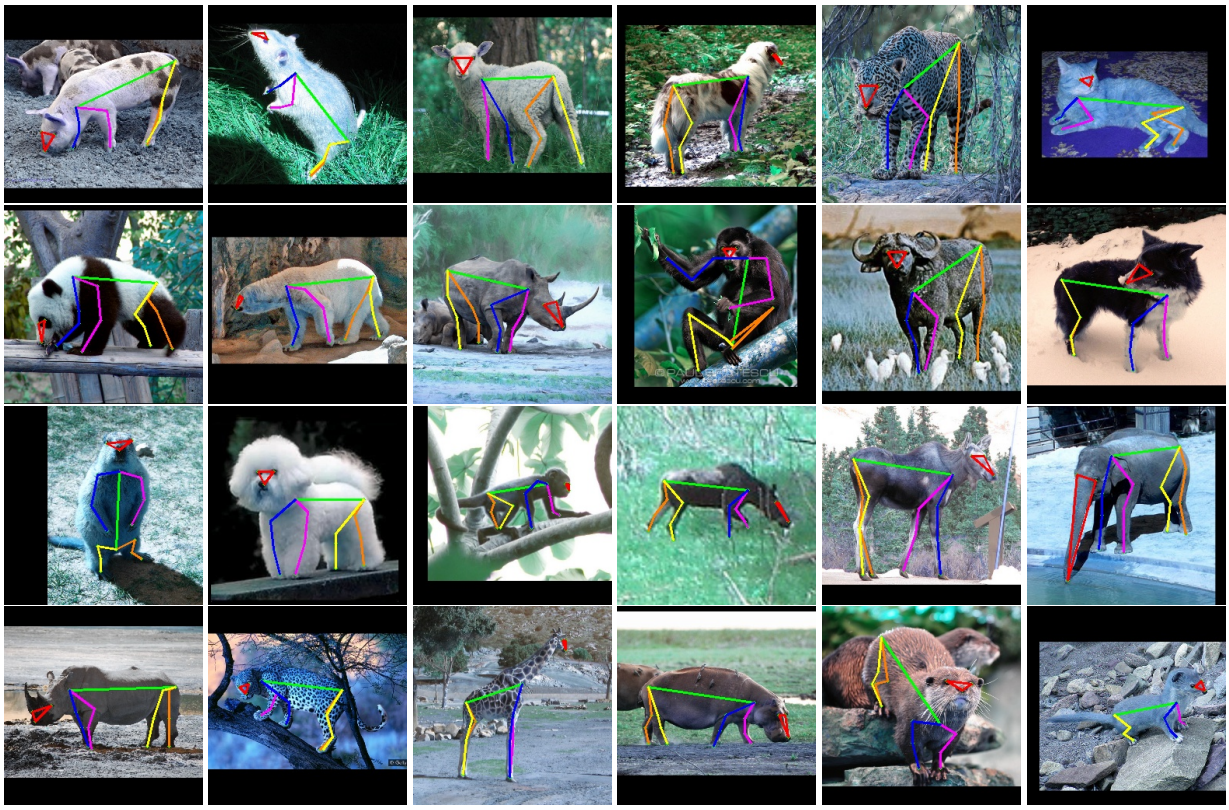

Figure 1. Qualitative results when 25 images per species are labeled.

conference on computer vision and pattern recognition workshops, pages 702–703, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1491, 2021. 1

[5] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12386–12395, 2020. 1

[6] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 1