Applied Deep Learning - Assignment 1

Peter Helf

October 2025

1 Topic Selection - Animal Pose Estimation

Pose estimation is a well researched topic in the field of computer vision. Many approaches focus on the estimation of human poses[1], [2], [3] on benchmark datasets such as COCO[4], CrowdPose[5], OCHuman[6], and MPII[7].

Models created for human pose estimation have also shown good results on animal pose datasets. ViTPose[1] achieves an AP of 82.4 on the AP-10k[8] dataset. AP-10k contains 10,015 annotated images from 23 different animal families.

The detection of poses can be used for Skeleton-based Action Recognition[9], [10]. By being able to automatically detect certain behaviours of animals, such as pain behaviours, it would be possible to provide better individualized care. This could in turn improve the health and welfare of many livestock species.

2 Project Type - Bring your own method

Several benchmark datasets for the evaluation of animal pose estimation estimation exist. The AP-10k[8] and APT-36k[11] datasets contain labeled images for several animal species. Datasets for individual species, such as mice, marmoset monkeys, and fish[12] also exist.

The newly proposed BUCTD architecture[13] combines the bottom-up and top-down pose estimation approaches. This architecture has shown state of the art performance on the CrowdPose[5] dataset, achieving an AP score 78.5, compared to 76.3 by ViTPose[1]. BUCTD has also been trained on the aforementioned mouse, monkey, and fish datasets achieving AP scores of 99.1, 93.7, and 88.7 respectively. The architecture has however not been tested on the more diverse AP-10k and APT-36k datasets. The aim of the following assignments will be the reimplementation and testing of BUCTD on AP-10k. Further, improvements to the model architecture will be tested in an attempt to improve the results.

3 Dataset - AP-10k

AP-10k[8] consists of 10,015 images containing 23 animal families and 54 species. The images have high-quality keypoint annotations in COCO format. 17 keypoints are labeled on each animal. A visualization of these keypoints can be seen in figure 1 with the definition given in Table 1.

Keypoint	Definition	Keypoint	Definition
1	Left Eye	10	Right Elbow
2	Right Eye	11	Right Front Paw
3	Nose	12	Left Hip
4	Neck	13	Left Knee
5	Root of Tail	14	Left Back Paw
6	Left Shoulder	15	Right Hip
7	Left Elbow	16	Right Knee
8	Left Front Paw	17	Right Back Paw
9	Right Shoulder		

Table 1: Definition of animal keypoints as presented in [8, Table 1]

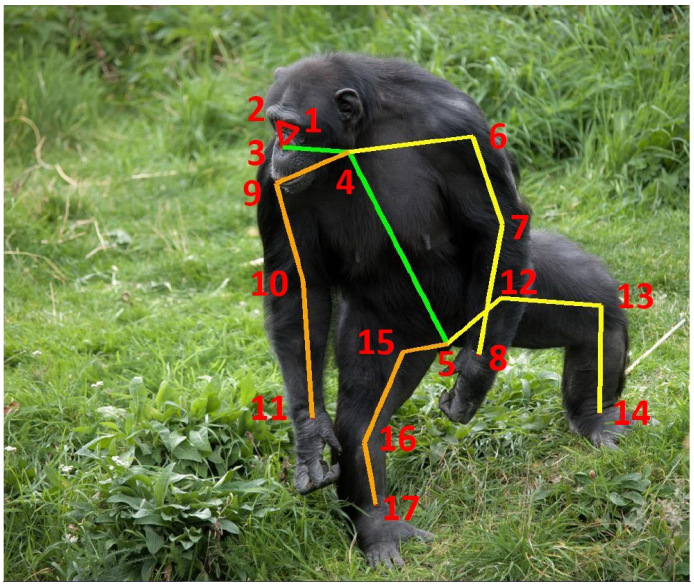

Figure 1: Keypoint definitions as illustrated in [8, Figure 2]

4 Task breakdown

Getting familiar with the dataset - ≈5 hours
Analyzing of the data structure of the dataset. Finding possible needed changes to make it compatible with the BUCTD model.

Recreation and initial training of BUCTD - ≈20 hours
Reimplementation of the model and first training sessions on the dataset. Evaluation of initial results.

Hyperparameter tuning and architecture changes - ≈15 hours
Attempt to improve the performance by hyperparameter tuning and testing changes to individual layers in the model.

Creation of simple Demo Application - ≈5 hours
Development of a simple GUI to run the model on newly uploaded images.

Creating the final report and the presentation - ≈10 hours
Creation of the final project report detailing the approach and the results. Creation of a short presentation.

References

[1] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022.

[2] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[3] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 676–14 686.

[4] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[5] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 863–10 872.

[6] S.-H. Zhang et al., "Pose2seg: Detection free human instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 889–898.

[7] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[8] H. Yu, Y. Xu, J. Zhang, W. Zhao, Z. Guan, and D. Tao, "Ap-10k: A benchmark for animal pose estimation in the wild," *arXiv preprint arXiv:2108.12617*, 2021.

[9] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2969–2978.

[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[11] Y. Yang, J. Yang, Y. Xu, J. Zhang, L. Lan, and D. Tao, "Apt-36k: A large-scale benchmark for animal pose estimation and tracking," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 301–17 313, 2022.

[12] J. Lauer et al., "Multi-animal pose estimation, identification and tracking with deeplabcut," *Nature Methods*, vol. 19, no. 4, pp. 496–504, 2022.

[13] M. Zhou, L. Stoffl, M. W. Mathis, and A. Mathis, "Rethinking pose estimation in crowds: Overcoming the detection information bottleneck and ambiguity," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 14 689–14 699.