Evaluation of the Effect of Transfer Learning to Multi-Instance Detection of Monkeys

Riza Rae Pineda*,†, Takatomi Kubo*,‡, Masaki Shimada§, and Kazushi Ikeda*
* Division of Information Science, Nara Institute of Science and Technology
§ Department of Animal Sciences, Teikyo University of Science
† E-mail: pineda.riza_rae.pn4@is.naist.jp
‡ E-mail: takatomi-k@is.naist.jp Tel/Fax: +81-743-72-5985

***Abstract*—Multiple object tracking is an open problem in computer vision. Various studies have presented effective and efficient architectures for object tracking and have paved the way towards the construction and development of task-specific assistive tools such as in animal behavior analysis or road traffic monitoring. Despite the growing interest in computer vision systems for animal behavior analysis, no such tools have been developed for multi-animal detection specific to macaques. In this study, we aim to develop a robust Japanese macaque detection model. We also explore how transfer learning affects the detection accuracy of the trained models. With a mean AP_{50} of 83.17%, F1 score of 84%, and a mean $IOU_{AP@50}$ of 70.27%, our Japanese macaque detection model pre-trained on the MacaquePose dataset using YOLOv4 yielded the best performance. Transfer learning from a related task increased the mean average precision at IOU_{50} by 8% and significantly reduced training convergence time.**

I. Introduction

Behavioral studies involve the observation of interactions between organisms in the environment. These observations in animals, supported by evolutionary evidence, have led to a deeper understanding of human evolutionary patterns. Primates, our closest genetic relative, are the subject of many studies aiming to map developmental and mutative patterns to potentially understand human behavior and instincts. Westergaard and Fragaszy [1] have observed in a troop of capuchin monkeys a tradition of using tools such as rocks to open foraged nuts in a group. Non-human primates have also been observed to experience stress when subjected to relocation and social isolation [2], and develop physiological signs of depression such as arched postures and decreased movements as consequences of subordination in troop [2], [3].

Collecting longitudinal data spanning years of the same population is a common practice to establish behavioral findings in monkey behavior research. Primatologists jot down manually observed information of the monkey population and often record videos along with it. With an extensive video archive, performing a thorough review through frame-by-frame analysis is time-consuming.

Current innovations in computer vision enable the development of complex systems that may aid and fast-track aggregation and analysis of visual data. Fast Region-based Convolutional Neural Networks (Fast R-CNN) [4], introduced in 2015, is an object detection architecture that jointly learns classification and refinement of object proposals. The input is passed on to multiple convolutional and max pooling layers. Region proposals are then pooled and fed into fully connected layers for object classification and bounding box estimation. You Only Look Once (YOLO) [5]–[8] is another object detection network that is currently widely used in different studies and applications because of its swift and accurate detection ability in most benchmark datasets. The main design behind this network is the division of input into regions and the prediction of bounding boxes and class probabilities for each region. Detection modules are also performed at different scales, increasing the ability of the network to handle objects in largely varying sizes. Currently leading in object detection benchmarks in terms of detection accuracy and speed, EfficientDet [9] utilizes a weighted bidirectional feature network with a customized compound scaling method. This architecture uses EfficientNet [10] as its backbone network, Bidirectional Feature Pyramid Networks (BiFPN) as its feature network and shared a network for class and box prediction towards the end. These methods have been trained and tested on general object detection datasets such as COCO [11], ImageNet [12], and Pascal VOC [13]. With the emergence of robust architectures for general object detection, developing stronger models for more specific tasks such as human tracking or scene understanding has been made easier and possible.

In recent times, there has been a growing interest in the specific field of computer vision methods for animal behavior analysis [14]–[16]. However, for larger animals in the wild with complex body structures and wide degrees of movement freedom such as chimpanzees, gorillas, and macaques, such systems have yet to be developed. Labugen et. al. [17] proposed a novel visual dataset containing more than 13,000 images of macaques with instance and body key point labels. They also presented a single-monkey pose detection model using DeepLabCut [18].

In the field of primate research, trooping behavior and relationships between other monkeys are topics of high interest in this field. However, single-animal detection models such as [17] require an additional level of processing to accommodate visual data containing multiple animals, which occurs often. Motivated by current innovations in computer vision to efficiently comb through extensive longitudinal visual data through the automatic detection of target objects [19] and the need to establish more findings in behavioral research of

Japanese macaques, we contribute a robust monkey detection model that is able to detect multiple instances of our target species accurately. We experiment as well on the effect of transfer learning to training convergence and the overall detection accuracy of the trained models.

II. Methodology

Our goal in this work is to construct an accurate Japanese macaque detection model. With this, we experiment on the construction of robust multi-animal detection models as well as inspect the effect of transfer learning from a large general macaque dataset to our target species to the overall detection accuracy. In this section, we will discuss the various datasets and the object detection network used in our study.

A. Dataset

In our study, we used three different datasets with high scene heterogeneity and disparity. First, we utilized the MacaquePose dataset [17] comprised of 13,088 frames with 16,393 unique monkey instances. This dataset contains key point and instance labels of various macaque species such as the rhesus macaque, Japanese macaque, and the like. Each monkey instance is annotated with 17 key points namely, nose and left and right locations of the ears, eyes, shoulders, elbows, wrists, hips, knees, and ankles. However, for this study, we instead computed for the minimum bounding box of each instance and used the resulting set of monkey boxes for experimentation. We focused our subsequent training on Japanese macaques using two datasets: our videos collected from monitoring a troop of this species in the forests of Kinkazan, Miyagi, Japan [20] and Youtube videos containing the target species. We have a total of $3,604$ monkey boxes from recorded clips of a troop of Japanese macaques in the wild. Videos were taken by our collaborator from a lateral angle using a single handheld camera. These recordings were collected by following the individuals of interest, causing scene conditions to highly vary at different times with little to no static reference frame throughout the same clip. To reinforce the generalizability of our model on Japanese macaque detection, we acquired random videos of Japanese macaques from Youtube. Since these were sourced from varying originators in different environments using different cameras, this set has more noise, scene changes, and variations in image quality and resolution. We have a total of 576 frames and 2,020 monkey boxes labeled under this set. Figure 1 contains sample scenes from each dataset and Table I shows the breakdown of each dataset collected.

TABLE I
Dataset characteristics

Dataset	Description	Frames	Boxes
MacaquePose [17]	species: various (rhesus macaques, Japanese macaques, etc.) scenes: various (zoo, hotsprings, etc.)	13,088	16,393
Ours [20]	species: Japanese macaques scenes: forest only	1,231	3,604
Youtube	species: Japanese macaques scenes: various (zoo, snow, hotsprings, etc.) multiple foreign objects	576	2,020

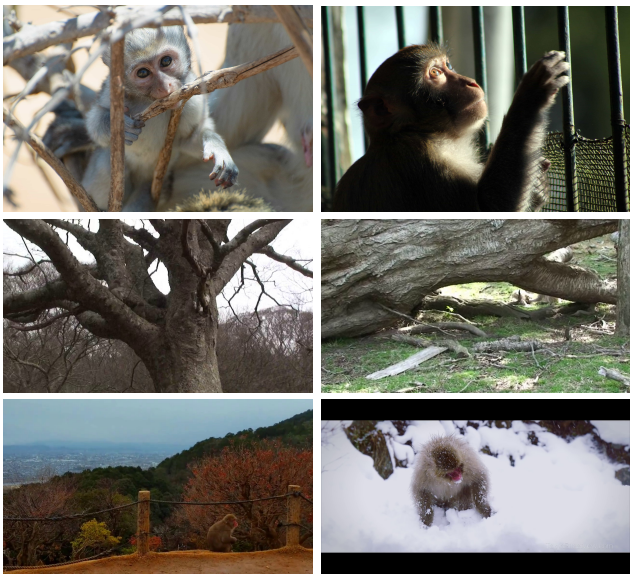

Fig. 1. Different scenes in the datasets: top row contains scenes from the MacaquePose dataset [17], middle row contains scenes from our dataset [20], and the bottom row contains scenes from the Youtube dataset. The high variability of textures and colors in the scenes increases the complexity of detection and spatial-based feature extraction.

B. Monkey Detection using You Only Look Once (YOLOv4)

We trained a monkey detection model using YOLOv4 [7], [8]. This object detection framework is composed of three modules: backbone, neck, and head. YOLOv4 uses CSPDarknet-53 as its backbone network, as shown in Figure 2. This version of Darknet-53 reintroduces its residual blocks as Cross-Stage Partial blocks (CSP) to mitigate the heavy computational load that it usually requires. The receptive fields are then enhanced using Spatial Pyramid Pooling (SPP) [21] and parameters are aggregated from different backbone levels using a Path Aggregation Network (PANet) [22]. The resulting feature maps of the neck module are then passed on to the YOLOv3 [6] layers to predict classes and the bounding boxes of objects. YOLOv4 [7], [8] proposes new data augmentation techniques such as mosaicking and cut mix, aside from photometric and geometric distortions, to increase the variability of the training images and the robustness of the detection model.

1) Spatial Pyramid Pooling (SPP): Spatial Pyramid Pooling (SPP) [21] is a pooling method where the input is partitioned into divisions in different levels and aggregates local features in them. This enables the processing of variable-sized or variable-scaled images and consequently, reduces over-fitting. Moreover, due to the features being pooled at different scales, this method improves the flexibility or scale-invariance of the overall network and generally increases its

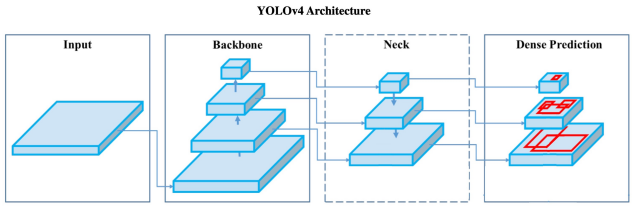

Fig. 2. The YOLOv4 Architecture [7], [8]. The input images are passed to the backbone network, CSPDarknet-53, for feature extraction. Spatial Pyramid Pooling (SPP) and a Path Aggregation Network (PANet) in the neck stage then improves the robustness of detection by enhancing the receptive fields and allowing for parameter-sharing across different backbone levels. Predictions using the YOLOv3 [6] layers are then performed at different stages.

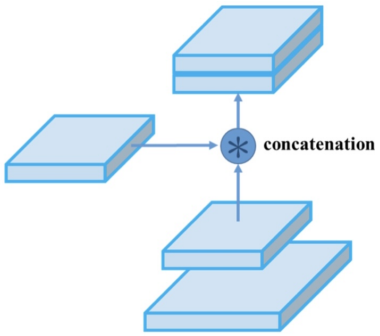

Fig. 3. Modified Path Aggregation Network (PANet) [22] for YOLOv4. Information is propagated from the lower layers to the upper layers with parameters from different backbone layer through a concatenation operator [7], [8]

detection ability.

2) Path Aggregation Network (PANet): A Path Aggregation Network (PANet) [22] propagates accurate localization information from the lower to the topmost layers which improve the quality of the region proposals of the network. We used a modified version of PANet where the shortcut connection is replaced with concatenation, as shown in Figure 3.

3) Mish activation: We used *Mish* [23] as the activation function in replacement of ReLU [24]. This function is given by

$$f_{mish}(x) = xtanh(softplus(x)) \quad (1)$$

Similar to *Swish* [25], *Mish* is smooth, self-regularized, and non-monotonic. Unlike ReLU [24], this function is continuously differentiable which avoids singularities and allows for smoother gradient-based learning with no side-effects. As shown in Table II, *Mish* performs better with CSPDarknet-53 than *ReLU* on the MS COCO dataset.

Using these modules, we performed training in two stages. The first stage involves training a general macaque detection model using two versions of YOLOv4 on the MacaquePose [17] dataset. We then utilized these as the base models for Japanese macaque detection using our own dataset and Youtube datasets on the used YOLOv4 architectures.

TABLE II
TEST PERFORMANCE OF CSPDARKNET-53 [7], [8] WITH ACTIVATIONS RELU [24] AND MISH [23] ON THE MS COCO DATASET

Model	Size	Data Augmentation	ReLU	Mish
CSP-Darknet53	(512 × 512)	No	64.5%	**64.9%**
CSP-Dakrnet53	(608 × 608)	No	-	**65.7%**
CSP-Darknet53+ PANet+SPP	(512 × 512)	Yes	64.5%	**64.9%**

III. EXPERIMENTS

In this section, we discuss the experiments performed for monkey detection. Transfer learning has been observed to provide better results when training models even on unrelated tasks versus starting from scratch. We design our experiments to analyze the changes in the detection accuracy when employing methods such as transfer learning and introducing dataset variability. We staged our training into two phases: general macaque detection and Japanese macaque detection and split each dataset into 70% training, 20% testing, and 10% validation. This gives us the breakdown shown in Table III. We performed all our experiments using an NVIDIA RTX 3090 GPU with 24GB memory.

TABLE III
BREAKDOWN OF EACH DATASET USED IN THE EXPERIMENTS

Dataset	Training	Validation	Testing
MacaquePose [17]	9160	1312	2616
Ours [20]	862	123	246
Youtube	403	59	114

*1) **First Stage: Training a General Macaque Detection Model on the MacaquePose Dataset**:* For this initial stage, we computed for the minimum bounding boxes of each monkey instance label in the MacaquePose dataset. We then trained a general macaque detection model using the YOLOv4 architecture with a network size of (416x416) with batch and subdivision sizes as 64 and 16, respectively. We set our learning rate to 0.001, momentum at 0.949, and decay at 0.0005. We also enabled data augmentations with the following parameter values: saturation at 1.5, exposure at 1.5, hue at 0.1, and mosaicking enabled. For model validation during training, we used the MacaquePose dataset. To measure the accuracy of our trained detector, we used the testing set comprised of 2,616 frames with 3,517 corresponding boxes from the same dataset.

*2) **Second Stage: Training a Japanese Macaque Detection Model on Our and Youtube Datasets**:* For the second stage, we trained separate models on each training dataset using the same validation set and the pre-trained general macaque detection model as the starting point. We tested the trained models at this stage on our testing set and the Youtube dataset separately to inspect the generalizability of the model as well as its accuracy in the primary dataset that we will be using for our monkey tracking and behavioral research.

We performed another set of experiments to verify the significance of transfer learning in the development of an accurate monkey detection system. We trained these models using three different initial weights: random, pre-trained on the

COCO dataset, and pre-trained on the MacaquePose dataset. Based on the results of this experimentation, we will observe how transfer learning affects the overall detection accuracy versus using randomly initialized weights.

YOLOv4 Parameters. For this stage, we used the general monkey detection model trained using the original YOLOv4 architecture as the initial weights of the subsequent training experiments and kept the same YOLOv4 parameters from stage 1. The other hyperparameters such as learning rate, momentum, and decay were kept at 0.001, 0.949, and 0.0005, respectively.

IV. Results and Discussion

In this section, we will present and discuss the results of our experiments.

A. First Stage: General Macaque Detection

Initially, we trained YOLOv4 on the MacaquePose [17] dataset. Shown in Table IV are the mean average precision value at IOU^{50} (mAP_{50}) of 90.92%, F1 score of 90%, and mean IOU of 71.46% of the trained model.

TABLE IV
Performance evaluation of the general macaque detector (%)

Model	MacaquePose
mAP_{50}	90.92
Precision	91
Recall	89
F1 Score	90
mIOU	71.46

B. Second Stage: Japanese Macaque Detection

We used the general macaque detection model trained using the original YOLOv4 in stage 1 as the initial weights of our training experiments in stage 2. As shown in Table V, our model achieved a mAP_{50} of 83.17%, F1 score of 84% and meanIOU of 70.27% on our test set. With a high precision score of 90%, the model seldom boxes incorrect objects. Despite having large disparities in image resolution, scenes, and monkey box sizes from the frames in our own video dataset, our Youtube-trained model yielded an acceptably high accuracy of 70.78% mAP_{50} on our test images.

TABLE V
Performance evaluation (mAP_{50}) of the trained Japanese macaque detection models on the frames in our test dataset

Training	MacaquePose	Ours	Youtube
Validation	Ours	Ours	Ours
mAP_{50}	21.56	83.17	70.78
Precision	60	90	72
Recall	26	79	71
F1 Score	36	84	71
mIOU	39.29	70.27	55.77

We also tested on the Youtube test set to further check the generalizability of our detection models trained on different datasets. Similar with the previously-discussed results in Table V, we achieved the following test results on our Youtube test set on Table VI. The model trained on the Youtube dataset achieved the highest mAP_{50} of 71.5% and mean IOU of 62.69%. The MacaquePose model trained with our validation set yielded the lowest with mAP_{50} scores of below 30% on both test sets.

TABLE VI
Performance evaluation (mAP_{50}) of the trained Japanese macaque detection models on the Youtube test dataset

Training	MacaquePose	Ours	Youtube
Validation	Ours	Ours	Ours
mAP_{50}	26.64	49.58	71.5
Precision	71	61	80
Recall	28	48	66
F1 Score	40	54	73
mIOU	48.83	45.9	62.69

Shown in Table VII is the test performance of the trained models, using different initial weights, on our test set. The model with pre-trained MacaquePose weights yielded the highest detection performance at 83.17%. Training converged faster using pre-trained weights than with randomized weights. Shown in Figure 4 are the training loss and the validation mAP_{50} plots of the training run using the random, pre-trained on COCO [11], and pre-trained on MacaquePose [17], respectively. Observing the behavior of the validation accuracy plot across the different graphs, both the pre-trained models yielded the best mAP a few hundred steps after the first thousandth iteration. While both pre-trained models are already showing a downward validation accuracy trend and a training loss of less than 2 at the second thousandth iteration, we still see an increasing trend on the model initialized with random weights. This suggests a boost in the speed of training convergence for models with pre-trained weights on a general or a related task versus randomized weights.

Comparing the results of all the ablation studies performed in this study, transfer learning from a more related task improves the accuracy of the detection model and allows for faster convergence.

V. Conclusion

In this paper, we proposed a new multi-instance detection model for monkeys that can be used in different behavioral researches that involve visual data. With a mean AP_{50} of 83.17% our system was able to accurately detect Japanese macaques in varying environments and conditions. We also show that transfer learning improves detection accuracy and reduces training convergence time.

TABLE VII
Performance evaluation of the trained Japanese macaque detection models on our test set

Model	Initial Weights	mAP_{50} (%)	F1 Score(%)
YOLOv4	random	75.4	77
YOLOv4	pre-trained on MS COCO [11]	79.14	79
YOLOv4	pre-trained on MacaquePose [17]	**83.17**	**84**

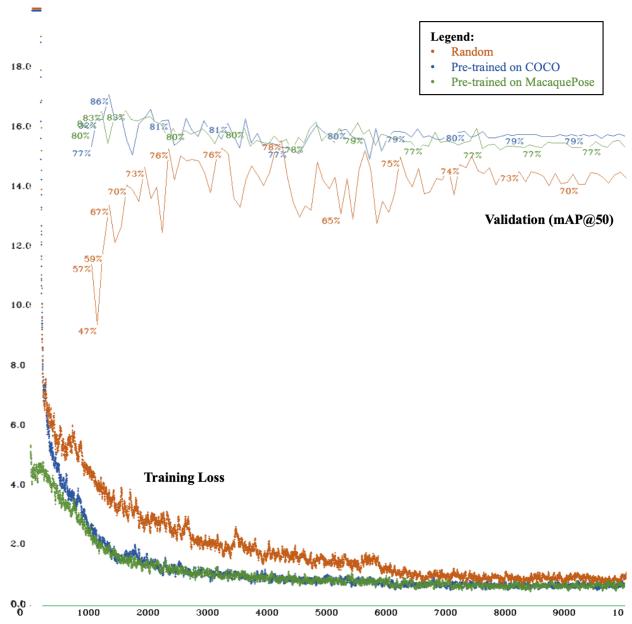

Fig. 4. Training Loss versus Validation Accuracy (mAP_{50}). The orange, blue, and green graphs show the training loss and validation accuracy plots of the run using randomized, pre-trained weights on the COCO dataset, and pre-trained weights on the MacaquePose dataset, respectively. The orange-colored graph starts at a loss of 1759.67. The blue-colored graph begins at a loss score of 1646. The green-colored graph shows an initial loss of 5.1, significantly lower than the first two above. Both pre-trained models yielded the highest validation accuracy a few hundred steps after the first thousandth iteration, with the latter leading by a few hundred iterations. While both pre-trained models are already showing a downward validation accuracy trend and a training loss of less than 2 at the second thousandth iteration, we still see an increasing trend on the model initialized with random weights. This suggests a longer training convergence time required for models with randomized weights versus those pre-trained with a general or a related task.

VI. Acknowledgments

This work was supported in part by KAKENHI grant number 18K19821 and by a postgraduate scholarship from the Engineering Research and Development for Technology (ERDT), Philippines.

References

[1] G. WESTERGAARD and D. Fragaszy, "The manufacture and use of tools by capuchin monkeys (cebus apella)," *Journal of Comparative Psychology*, vol. 101, pp. 159–168, 06 1987.

[2] J. S. Meyer and A. F. Hamel, "Models of stress in nonhuman primates and their relevance for human psychopathology and endocrine dysfunction," *ILAR journal*, vol. 55, no. 2, pp. 347–360, 2014, 25225311[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25225311

[3] S. L. Willard and C. A. Shively, "Modeling depression in adult female cynomolgus monkeys (macaca fascicularis)," *American Journal of Primatology*, vol. 74, no. 6, pp. 528–542, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajp.21013

[4] R. B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015. [Online]. Available: http://arxiv.org/abs/1504.08083

[5] J. Redmon, "Darknet: Open source neural networks in c," http://pjreddie.com/darknet/, 2013–2016.

[6] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020.

[8] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 13 029–13 038.

[9] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," *CoRR*, vol. abs/1911.09070, 2019. [Online]. Available: http://arxiv.org/abs/1911.09070

[10] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[11] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[13] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, p. 303–338, Jun. 2010. [Online]. Available: https://doi.org/10.1007/s11263-009-0275-4

[14] M. Clapham, E. Miller, M. Nguyen, and C. T. Darimont, "Automated facial recognition for wildlife that lack unique markings: A deep learning approach for brown bears," *Ecology and Evolution*, vol. 10, no. 23, pp. 12 883–12 892, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.6840

[15] R. Sarfati, J. Hayes, E. Sarfati, and O. Peleg, "Spatio-temporal reconstruction of emergent flash synchronization in firefly swarms via stereoscopic 360-degree cameras," *Journal of the Royal Society, Interface*, vol. 17, p. 20200179, 09 2020.

[16] D. McIntosh, T. P. Marques, A. B. Albu, R. Rountree, and F. D. Leo, "Movement tracks for the automatic detection of fish behavior in videos," *CoRR*, vol. abs/2011.14070, 2020. [Online]. Available: https://arxiv.org/abs/2011.14070

[17] R. Labuguen, J. Matsumoto, S. B. Negrete, H. Nishimaru, H. Nishijo, M. Takada, Y. Go, K.-i. Inoue, and T. Shibata, "Macaquepose: A novel "in the wild" macaque monkey pose dataset for markerless motion capture," *Frontiers in Behavioral Neuroscience*, vol. 14, p. 268, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbeh.2020.581154

[18] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, 2018. [Online]. Available: https://www.nature.com/articles/s41593-018-0209-y

[19] D. Schofield, A. Nagrani, A. Zisserman, M. Hayashi, T. Matsuzawa, D. Biro, and S. Carvalho, "Chimpanzee face recognition from videos in the wild using deep learning," *Science Advances*, vol. 5, 09 2019.

[20] M. Shimada and C. Sueur, "Social play among juvenile wild japanese macaques (macaca fuscata) strengthens their social bonds," *American Journal of Primatology*, vol. 80, no. 1, p. e22728, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajp.22728

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *CoRR*, vol. abs/1406.4729, 2014. [Online]. Available: http://arxiv.org/abs/1406.4729

[22] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *CoRR*, vol. abs/1803.01534, 2018. [Online]. Available: http://arxiv.org/abs/1803.01534

[23] D. Misra, "Mish: A self regularized non-monotonic neural activation function," *CoRR*, vol. abs/1908.08681, 2019. [Online]. Available: http://arxiv.org/abs/1908.08681

[24] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *CoRR*, vol. abs/1505.00853, 2015. [Online]. Available: http://arxiv.org/abs/1505.00853

[25] P. Ramachandran, B. Zoph, and Q. Le, "Swish: a self-gated activation function," 10 2017.

[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," Apr 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[27] T. Matsuzawa, "Hot-spring bathing of wild monkeys in shiga-heights: origin and propagation of a cultural behavior," *Primates*, vol. 59, no. 3, pp. 209–213, May 2018. [Online]. Available: https://doi.org/10.1007/s10329-018-0661-z

[28] S. Kawamura, "The process of sub-culture propagation among japanese

macaques," *Primates*, vol. 2, no. 1, pp. 43–60, Mar 1959. [Online]. Available: https://doi.org/10.1007/BF01666110

[29] M. Kawai, "Newly-acquired pre-cultural behavior of the natural troop of japanese monkeys on koshima islet," *Primates*, vol. 6, no. 1, pp. 1–30, Aug 1965. [Online]. Available: https://doi.org/10.1007/BF01794457

[30] K. R. Wright, J. A. Mayhew, L. K. Sheeran, J. A. Funkhouser, R. S. Wagner, L.-X. Sun, and J.-H. Li, "Playing it cool: Characterizing social play, bout termination, and candidate play signals of juvenile and infant tibetan macaques (macaca thibetana)," *Zoological research*, vol. 39, no. 4, pp. 272–283, Jul 2018, 29766979[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29766979

[31] T. Matsuzawa, "Sweet-potato washing revisited: 50th anniversary of the primates article," *Primates*, vol. 56, no. 4, pp. 285–287, Oct 2015. [Online]. Available: https://doi.org/10.1007/s10329-015-0492-0

[32] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," *CoRR*, vol. abs/1807.06521, 2018. [Online]. Available: http://arxiv.org/abs/1807.06521

[33] J. Lauer, M. Zhou, S. Ye, W. Menegas, T. Nath, M. M. Rahman, V. Di Santo, D. Soberanes, G. Feng, V. N. Murthy, G. Lauder, C. Dulac, M. W. Mathis, and A. Mathis, "Multi-animal pose estimation and tracking with deeplabcut," *bioRxiv*, 2021. [Online]. Available: https://www.biorxiv.org/content/early/2021/04/30/2021.04.30.442096