arXiv:2509.12193v1 [cs.CV] 15 Sep 2025

Domain-Adaptive Pretraining Improves Primate Behavior Recognition

Felix B. Mueller[1] Timo Lueddecke[1] Richard Vogg[1] Alexander S. Ecker[1,2]

[1]Institute of Computer Science and Campus Institute Data Science, University of Göttingen, Germany

[2]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany

felix.mueller@cs.uni-goettingen.de

Abstract

Computer vision for animal behavior offers promising tools to aid research in ecology, cognition, and to support conservation efforts. Video camera traps allow for large-scale data collection, but high labeling costs remain a bottleneck to creating large-scale datasets. We thus need data-efficient learning approaches. In this work, we show that we can utilize self-supervised learning to considerably improve action recognition on primate behavior. On two datasets of great ape behavior (PanAf and ChimpACT), we outperform published state-of-the-art action recognition models by 6.1 %pt. accuracy and 6.3 %pt. mAP, respectively. We achieve this by utilizing a pretrained V-JEPA model and applying domain-adaptive pretraining (DAP), i. e. continuing the pretraining with in-domain data. We show that most of the performance gain stems from the DAP. Our method promises great potential for improving the recognition of animal behavior, as DAP does not require labeled samples. Code is available at https://github.com/ecker-lab/dap-behavior

1. Introduction

There has been a growing interest in computer vision methods for animal behavior in recent years. Understanding animal behavior is crucial for many different fields, like cognition, ecology, and animal conservation. In this domain, non-human primate cognition is of high interest, due to its complexity and relation to human cognition.

Animal behavior monitoring using computer vision is widely adopted in lab settings [24, 27], but these methods not easily transferable to more complex visual environments [33]. They exploit the characteristics of one very constrained environment, which is often not possible in the wild. For in-the-wild settings, we are typically able to collect large amounts of data, but only parts of the data are relevant for the research problem at hand and a crucial limiting factor is labeling cost. This makes data-efficient machine learning methods highly relevant.

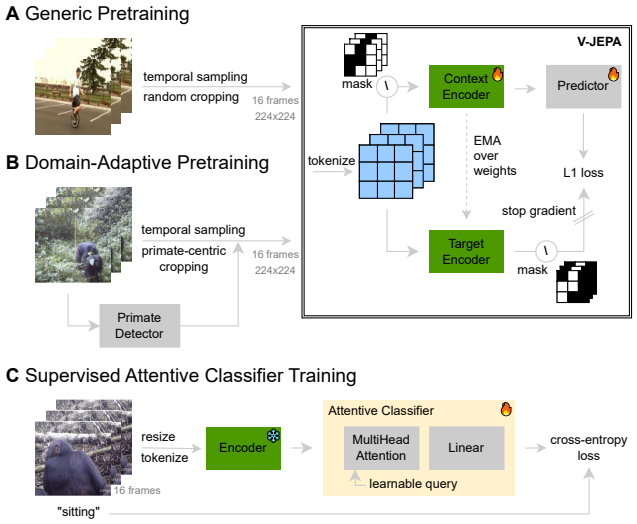


Figure 1. Overview over our methodology and the V-JEPA training paradigm. EMA: exponential moving average. 🔥: trainable parameters. ❄: frozen parameters. **A.** Self-supervised training utilizing human action datasets, done by [2]. **B.** Our domain-adaptive pretraining utilizes the same learning paradigm as in A. **C.** We freeze the encoder from A and B and train a classifier on top.

In this work, we explore self-supervised learning as a data-efficient learning method. In self-supervised learning, models are trained to learn useful representations of input data *without requiring labeled data*. Recently, there has been considerable progress in video representation learning using masked autoencoding objectives [2, 32].

We apply self-supervised learning to the task of *behavior recognition*, i.e. performing action recognition with primates as subjects. We perform domain-adaptive pretraining on a V-JEPA [2] backbone already pretrained on human video data and then train an attentive classifier on top of the frozen backbone (Fig. 1). We thereby beat the previous best published results on behavior recognition on PanAf500 [6] and ChimpACT [22]. On both datasets, we see that DAP gives large performance improvements compared to using the pretrained V-JEPA backbone directly.

2. Related Work

Over the last years, a growing number of datasets for animal behavior recognition have been published, both general [25, 29] and primate specific. The latter range from zoo recordings [14, 22] over in-the-wild recordings using camera traps [4, 6] to drone footage [11].

There is a growing number of works on animal behavior recognition: Bain *et al*. [1] showed that they can discriminate two audiovisually distinctive actions in wild chimpanzees using deep learning. Various other works try to recognize primate behavior using deep learning [15, 19, 34], improve behavior recognition with text descriptions [5], or adapt the CLIP model [28] for retrieval of camera trap events using natural language [16].

The goal in self-supervised learning is to learn useful representations from unlabeled data. This allows to perform a downstream task (e.g. ImageNet classification) without much additional training, either by training a classification head on top of the frozen backbone model or by finetuning for few iterations. While there exists a plethora of self-supervised learning methods for images [7, 9, 26, 36], approaches for video input are more rare and are typically based on masked autoencoding. In masked autoencoding, the training objective is to predict masked regions of a video given the unmasked regions. The reconstruction loss can either be calculated directly in pixel space [32] or in latent space as is done in V-JEPA [2] and SIGMA [30]. Most recent self-supervised learning approaches use attention-based models, either Vision Transformer (ViT) [10] or a variation of it.

In this work, we focus on V-JEPA [2] as video backbone, as it is based on a large, diverse pretraining dataset, utilizing HowTo100M, Kinetics, and Something-Something-v2.

3. Methodology

To perform primate behavior recognition, we take a pretrained V-JEPA backbone and perform domain-adaptive pretraining without labels to adjust it to the primate domain. We then train an attentive classifier on top of the frozen backbone model to perform action recognition.

In this section, we describe the self-supervised pretraining setup (Section 3.1), a primate-centric sampling strategy for the domain-adaptivce pretraining (Section 3.2) and the attentive classifier used for action recognition (Section 3.3).

3.1. Self-Supervised Pretraining

We use the V-JEPA model, which consists of a context encoder E, a target encoder $\overline{E}$, and a predictor P. V-JEPA works on a video tensor $x \in \mathbb{R}^{16\times224\times224}$ consisting of 16 frames sampled with a temporal stride of 4 and with a resolution of 224×224. The input is patchified into patches of size $2 \times 16 \times 16$, tokenized, and combined with sine-cosine positional embeddings to produce a set of $N = 1568$ tokens $X \in \mathbb{R}^{N\times d}$, with $d = 1024$ being the token dimension.

During training, part of the input sequence is masked and the training objective is to predict the latent representation of masked tokens given the unmasked tokens (see Fig. 1). Let Mask and Mask^C be functions that split a set of tokens into the set of masked and unmasked tokens, respectively. Then the V-JEPA pretraining loss for an unlabeled data sample X is formulated as

$$L_{\text{MAE}} = \left\| P\left(E\left(\text{Mask}(X)\right)\right) - \text{Mask}^C\left(\overline{E}(X)\right)\right\|_1 . \quad (1)$$

The parameters of context encoder E and predictor P (see Fig. 1) are trained. To avoid collapse to trivial solutions, the parameters of target encoder $\overline{E}$ are not trained but computed as an exponential moving average over the past weights of the context encoder E. For generic pretraining, both teacher and student weights are initialized to the same random weights. For DAP, the final weights of teacher and student from generic pretraining are used.

The training objective L_{MAE} was used during V-JEPA pretraining and is also used for the domain-adaptive pretraining we perform.

3.2. Primate-Centric Sampling

We find that primate datasets are less object-centric compared to common datasets for human action recognition, like Something-Something-v2. Videos might be almost empty with one primate at a distance or they might contain many primates at the same time. To capture fine-grained details of individual primates without increasing the input size (and therefore memory and runtime complexity) of Vision Transformer models, we preprocess our primate-specific datasets for domain-adaptive pretraining by using Grounding DINO [21] as off-the-shelf open-vocabulary object detector to crop out parts of the video that contain primates (see Section 4.2 for details).

We perform the domain-adaptive pretraining on three-second video snippets v (see Section 4.1 on how we chunk the in-domain pretraining data into video snippets). We crop, resize, and downsample v both spatially and temporally to produce a video tensor x for the self-supervised pretraining,

$$x = \text{Sample}_{16}(\text{CropResize}(v, b)). \quad (2)$$

CropResize spatially crops all frames of the video snippet v to a bounding box $b \in \mathbb{R}^4$ and resizes them to 224×224, while Sample_{16} extracts a subsample of 16 frames. To crop spatially, we apply the primate detector on the center frame of v and randomly select one of the predicted bounding boxes. We increase the bounding box edge lengths by a factor of 1.25 on each side and then crop with a random

scale between 0.3 and 1.0 and a random aspect ratio between 0.75 and 1.35, yielding the bounding box b. We discard video snippets in which no primate was detected. To crop temporally, Sample_{16} takes a random segment of 64 frames (2s) out of the three-second snippet, which it reduces to 16 frames by applying a temporal stride of 4.

3.3. Supervised Attentive Classifier Training

Our goal is to predict the behavior of an animal specified by a bounding box b in an input video snippet v. We obtain a latent representation of the behavior by applying the pretrained encoder on a cropped video snippet

$$h = \overline{E}(\text{Sample}_{16}(\text{CropResize}(v, b))) \tag{3}$$

with $h \in \mathbb{R}^{N \times d}$.

We choose our attentive classifier to be a single multihead cross-attention layer which performs attention between a learnable query token $q \in \mathbb{R}^d$ and the latent representations h (keys and values), followed by a linear layer mapping to predicted class probabilities $\hat{y} \in [0, 1]^C$ with $C > 1$. Given a latent representation h, we predict class probabilities $\hat{y}$ as

$$\hat{y} = \sigma(\text{Linear}(\text{CrossAttention}(q, h))) \tag{4}$$

with σ being the *softmax* function for single-label classification and the *sigmoid* function for multi-label classification. We use a cross entropy loss function for single-label ground truth $L_{1L} = -\sum_{c=1}^{C} y_c \log \hat{y}$. In case of multi-label samples, we use binary cross entropy, $L_{ML} = -\sum_{c=1}^{C} [y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c)]$, with $y \in \{0, 1\}^C$ being the ground truth label for (x, b). We optimize only the parameters of Linear and CrossAttention. $\overline{E}$ is frozen.

Application to datasets. We use two datasets for action recognition: PanAf500 and ChimpACT. Both datasets provide ground truth bounding box information as part of the evaluation protocol, but pose the behavior recognition task in two different ways: snippet-wise or frame-wise.

For *snippet-wise* classification (like PanAf500), the input is a video snippet v, a sequence of bounding $(b_1, \ldots, b_{|v|})$ specifying how the animal moves through v, and one ground truth label y. To perform classification with our approach, we set b to be the union over the bounding boxes $(b_1, \ldots, b_{|v|})$. The PanAf evaluation video snippets are only 16 frames long, so we use no temporal stride for this dataset and directly input the spatially cropped snippets into the model.

For *frame-wise* classification (like ChimpACT), the input is a video v^*, one point in time t, one bounding box b at time t, and one ground truth label y for this bounding box. To perform classification with our approach, we select a 2s video snippet v from $v*$ centered at time t.

4. Experiments

4.1. Datasets and Evaluation Setup

We evaluate our approach on two video datasets of primates, PanAf [6] and ChimpACT [22].

The **PanAf** dataset contains camera trap videos from 18 field sites in tropical Africa capturing chimpanzees and gorillas. The dataset consists of 20,000 coarsely annotated videos (PanAf20k) and 500 videos with fine-grained bounding box, track and frame-wise behavior annotation (PanAf500). Each video is 15 s long. PanAf500 has frame-wise and per-individual single-label annotations of nine classes: sitting, walking, standing, hanging, climbing up, sitting on back, running, camera interaction, and climbing down (in descending order of frequency).

Following the PanAf action recognition evaluation protocol, we train and evaluate the attentive classifier on 16-frame snippets of sustained single behavior of a single ape and predict a single behavior label for each snippet. Performance is evaluated using Top-1 Accuracy and average per-class accuracy.

The **ChimpACT** dataset contains 2 h of video footage of zoo-housed chimpanzees recorded at the Leipzig Zoo. The recordings focus on one specific individual (Azibo) over the span of four years, but the dataset captures the whole chimpanzee group consisting of over 20 individuals. The dataset contains both indoor and outdoor recordings. ChimpACT has 23 behaviors annotated, which are be grouped into locomotion, social interaction, object interaction, and other. ChimpACT has ground truth multi-label behavior annotations for each ape in each frame. We evaluate mean average precision (mAP) following the AVA protocol [17].

Pretraining Data. For PanAf500, we utilize both PanAf500 and PanAf20k for pretraining, as PanAf20k is in-domain. For ChimpACT, we pretrain on all ChimpACT videos. In both cases we do not utilize any label information during pretraining even if available (all primate bounding boxes are obtained using the primate detector). We create samples by splitting all videos in the respective datasets into three-second snippets. We use a stride of 1.5s for PanAf500+20k, yielding 155k samples, and a stride of 1s for ChimpACT, yielding 6k samples.

4.2. Architectural Choices

Following [2], we choose a ViT-L model with video input for E and a narrow 12-layer transformer for the V-JEPA predictor P. E has 304M parameters, P has 22M parameters. All parameters are fine-tuned during domain-adaptive pretraining (DAP), which we perform for 14.4k steps with an effective batch size of 80, resulting in 1.2M samples seen during DAP. We choose a learning rate of 6×10^{-6} with warmup and cosine annealing. Following [2], we increase

Method	Top-1 Acc (%)	C-Avg Acc (%)
X3D [12]	80.04	56.10
I3D [8]	79.29	42.15
3D ResNet-50 [18]	77.45	55.17
MViTV2 [20]	81.09	54.91
TimeSformer [3]	79.45	<u>56.38</u>
V-JEPA (no DAP)	<u>83.68</u>	**57.75**
V-JEPA (DAP)	**87.24**	<u>56.37</u>

Table 1. Results on the behavioral action recognition task on the **PanAf500** test set. Baseline results are taken from [6]. For each baseline method we chose the best-performing setup based on Top-1 accuracy. DAP: Domain-adaptive pretraining. C-Avg Acc: Average per-class accuracy. We implemented the evaluation setup based on the description in [6], see Section 4.1.

Method	mAP	mAP_L	mAP_O	mAP_S
ACRN [31]	24.4	<u>58.7</u>	**33.8**	14.7
LFB [35]	23.2	45.0	31.2	17.7
SlowOnly [13]	22.3	52.3	31.2	13.8
SlowFast [13]	24.3	56.8	31.5	15.6
AlphaChimp* [23]	**34.3**	50.3	31.3	**29.3**
V-JEPA (no DAP)	26.0	**60.7**	29.2	17.8
V-JEPA (DAP)	<u>30.7</u>	55.7	<u>33.4</u>	<u>26.8</u>

Table 2. Results on the spatiotemporal action detection track on the **ChimpACT** test set. Baseline results are taken from [22], for each baseline method we chose the best-performing setup reported. DAP: Domain-adaptive pretraining. mAP: mean of average precision over classes. We report both the overall mAP and the mAP on Locomotion, Object interaction and Social interaction. *AlphaChimp is concurrent unpublished work.

the weight decay during training from 0.01 to 0.1, instead of decreasing it. DAP takes around 3 hours on 4 A100 GPUs.

We use zero-shot Grounding Dino [21] for primate detection (82.7 mAP_{All} on the PanAf500 primate detection task) as it can be applied easily to different datasets. We find the prompt "monkey.primate.ape." and a box detection threshold of 0.2 to work well.

The attentive classifier has 4M parameters (Eq. 4). When training it, we freeze the encoder $\overline{E}$. We train for up to 30 epochs on the labeled train sets and perform early stopping based on validation accuracy. We report results on the test set, calculating class scores on a single view per test sample.

5. Results

V-JEPA with domain-adaptive pretraining (DAP) outperforms plain V-JEPA by 3.5 percentage points in top-1 accuracy on PanAf500 (Table 1). Both V-JEPA with and without DAP outperform the previous best model. Even though we did not take specific measures to handle class imbalance, we still achieve competitive class-average accuracy. On the ChimpACT dataset we see the same pattern (Table 2): domain-adaptive pretraining clearly outperforms standard V-JEPA on mAP. Both V-JEPA with and without DAP outperform the previous best models.

5.1. Ablations

The V-JEPA attentive classifier is quite heavy, using multihead cross-attention followed by a two-layer MLP, and a final projection layer, resulting in 12M parameters [2]. We simplified it by dropping the MLP, which results in only a third of the parameters required while producing similar or better performance (Table 3). We assume that the attentive classifier is overparametrized given the small evaluation dataset size and thus reducing the number of parameters reduces overfitting.

	Param Count	PanAf500 Top-1 Acc %		ChimpACT mAP %	
		no DAP	DAP	no DAP	DAP
Ours	4M	**83.7**	**87.2**	25.97	**30.74**
V-JEPA	12M	82.5	87.1	**27.15**	29.74

Table 3. We compare our simplified attentive classifier using only multihead attention to the V-JEPA classifier using a full transformer block.

6. Conclusion

Our results show that domain-adaptive pretraining is a helpful tool when utilizing large pretrained backbone models for animal behavior tasks. A few GPU hours are enough to considerably improve downstream performance using DAP.

A limitation of our work is that we crop input videos to bounding boxes. While this allows to utilize the limited input resolution of ViTs better, it disregards global context and necessitates one forward pass per region of interest.

Self-supervised training allows us to utilize vast amounts of unlabeled data. Future work should explore to scale up DAP. Combining large and diverse sources of animal data might learn better representations and aid behavior recognition. This might especially improve performance on tail classes for which it might be hard to learn good representations from small datasets.

Acknowledgments

The project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 454648639 – SFB 1528. The authors gratefully acknowledge the computing time granted by the Resource Allocation Board and provided on the supercomputer

Emmy/Grete at NHR-Nord@Göttingen as part of the NHR infrastructure. The calculations for this research were conducted with computing resources under the project nib00021.

References

[1] Max Bain, Arsha Nagrani, Daniel Schofield, Sophie Berdugo, Joana Bessa, Jake Owen, Kimberley J. Hockings, Tetsuro Matsuzawa, Misato Hayashi, Dora Biro, Susana Carvalho, and Andrew Zisserman. Automated audiovisual behavior recognition in wild primates. 7(46):eabi4883. Publisher: American Association for the Advancement of Science. 2

[2] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. V-JEPA. 1, 2, 3, 4

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 813–824. PMLR, 2021. 4

[4] Otto Brookes, Maksim Kukushkin, Majid Mirmehdi, Colleen Stephens, Paula Dieguez, Thurston C. Hicks, Sorrel Jones, Kevin Lee, Maureen S. McCarthy, Amelia Meier, Emmanuelle Normand, Erin G. Wessling, Roman M. Wittig, Kevin Langergraber, Klaus Zuberbühler, Lukas Boesch, Thomas Schmid, Mimi Arandjelovic, Hjalmar Kühl, and Tilo Burghardt. The PanAf-FGBG dataset: Understanding the impact of backgrounds in wildlife behaviour recognition, . 2

[5] Otto Brookes, Majid Mirmehdi, Hjalmar Kuhl, and Tilo Burghardt. ChimpVLM: Ethogram-enhanced chimpanzee behaviour recognition, . 2

[6] Otto Brookes, Majid Mirmehdi, Colleen Stephens, Samuel Angedakin, Katherine Corogenes, Dervla Dowd, Paula Dieguez, Thurston C. Hicks, Sorrel Jones, Kevin Lee, Vera Leinert, Juan Lapuente, Maureen S. McCarthy, Amelia Meier, Mizuki Murai, Emmanuelle Normand, Virginie Vergnes, Erin G. Wessling, Roman M. Wittig, Kevin Langergraber, Nuria Maldonado, Xinyu Yang, Klaus Zuberbühler, Christophe Boesch, Mimi Arandjelovic, Hjalmar Kühl, and Tilo Burghardt. PanAf20k: A large video dataset for wild ape detection and behaviour recognition. . 1, 2, 3, 4

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. DINO. 2

[8] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017. 4

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. SimCLR, partially read. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Isla Duporge, Maksim Kholiavchenko, Roi Harel, Scott Wolf, Dan Rubenstein, Meg Crofoot, Tanya Berger-Wolf, Stephen Lee, Julie Barreau, Jenna Kline, Michelle Ramirez, and Charles Stewart. BaboonLand dataset: Tracking primates in the wild and automating behaviour recognition from drone videos. 2

[12] Christoph Feichtenhofer. X3D: expanding architectures for efficient video recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 200–210. Computer Vision Foundation / IEEE, 2020. 4

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019. 4

[14] Michael Fuchs, Emilie Genty, Adrian Bangerter, Klaus Zuberbühler, and Paul Cotofrei. From forest to zoo: Great ape behavior recognition with ChimpBehave, . 2

[15] Michael Fuchs, Emilie Genty, Klaus Zuberbühler, and Paul Cotofrei. ASBAR: an animal skeleton-based action recognition framework. recognizing great ape behaviors in the wild using pose estimation with domain adaptation, . Pages: 2023.09.24.559236 Section: New Results. 2

[16] Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. WildCLIP: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. 2

[17] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. 3

[18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 3154–3160. IEEE Computer Society, 2017. 4

[19] Yujie Lei, Pengmei Dong, Yan Guan, Ying Xiang, Meng Xie, Jiong Mu, Yongzhao Wang, and Qingyong Ni. Postural behavior recognition of captive nocturnal animals based on deep learning: a case study of bengal slow loris. 12(1):7738. Publisher: Nature Publishing Group. 2

[20] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4794–4804. IEEE, 2022. 4

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. 2, 4

[22] Xiaoxuan Ma, Stephan P. Kaufhold, Jiajun Su, Wentao Zhu, Jack Terwilliger, Andres Meza, Yixin Zhu, Federico Rossano, and Yizhou Wang. ChimpACT: A longitudinal dataset for understanding chimpanzee behaviors. 1, 2, 3, 4

[23] Xiaoxuan Ma, Yutang Lin, Yuan Xu, Stephan P. Kaufhold, Jack Terwilliger, Andres Meza, Yixin Zhu, Federico Rossano, and Yizhou Wang. Alphachimp: Tracking and behavior recognition of chimpanzees, 2024. 4

[24] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. 21 (9):1281–1289. Publisher: Nature Publishing Group. 1

[25] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. pages 19023–19034. 2

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. 2

[27] Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022. 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[29] Javier Rodriguez-Juan, David Ortiz-Perez, Manuel Benavent-Lledo, David Mulero-Pérez, Pablo Ruiz-Ponce, Adrian Orihuela-Torres, Jose Garcia-Rodriguez, and Esther Sebastián-González. Visual WetlandBirds dataset: Bird species identification and behavior recognition in videos. 2

[30] Mohammadreza Salehi, Michael Dorkenwald, Fida Mohammad Thoker, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Sigma: Sinkhorn-guided masked video modeling. In *European Conference on Computer Vision*, pages 293–312. Springer, 2024. 2

[31] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, pages 335–351. Springer, 2018. 4

[32] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. 35:10078–10093. VideoMAE. 1, 2

[33] Richard Vogg, Timo Lüddecke, Jonathan Henrich, Sharmita Dey, Matthias Nuske, Valentin Hassler, Derek Murphy, Julia Fischer, Julia Ostner, Oliver Schülke, Peter M. Kappeler, Claudia Fichtel, Alexander Gail, Stefan Treue, Hansjörg Scherberger, Florentin Wörgötter, and Alexander S. Ecker. Computer vision for primate behavior analysis in the wild. version: 1. 1

[34] Charlotte Wiltshire, James Lewis-Cheetham, Viola Komedová, Tetsuro Matsuzawa, Kirsty E. Graham, and Catherine Hobaiter. DeepWild: Application of the pose estimation tool DeepLabCut for behaviour tracking in wild chimpanzees and bonobos. 92(8):1560–1574. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.13932. 2

[35] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-term feature banks for detailed video understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 284–293. Computer Vision Foundation / IEEE, 2019. 4

[36] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. iBot. 2