LoTE-Animal: A Long Time-span Dataset for Endangered Animal Behavior Understanding

Dan Liu[1]*, Jin Hou[2,3]*, Shaoli Huang[4]*, Jing Liu[1], Yuxin He[5], Bochuan Zheng[2]†, Jifeng Ning[1]†, Jindong Zhang[2]†
[1]Northwest A&F University, [2]China West Normal University, [3]Beijing Normal University,
[4]Tencent AI-Lab, [5]North Sichuan Medical College
{dliu, iliujing}@nwafu.edu.cn, {houj815, heyuxin_nsmc, zhangjd224}@163.com,
shaolihuang@tencent.com, zhengbc@vip.163.com, njf@nwsuaf.edu.cn

Abstract

Understanding and analyzing animal behavior is increasingly essential to protect endangered animal species. However, the application of advanced computer vision techniques in this regard is minimal, which boils down to lacking large and diverse datasets for training deep models. To break the deadlock, we present ***LoTE-Animal****, a large-scale endangered animal dataset collected over 12 years, to foster the application of deep learning in rare species conservation. The collected data contains vast variations such as ecological seasons, weather conditions, periods, viewpoints, and habitat scenes. So far, we retrieved at least 500K videos and 1.2 million images. Specifically, we selected and annotated 11 endangered animals for behavior understanding, including 10K video sequences for the action recognition task, 28K images for object detection, instance segmentation, and pose estimation tasks. In addition, we gathered 7K web images of the same species as source domain data for the domain adaptation task. We provide evaluation results of representative vision understanding approaches and cross-domain experiments. LoTE-Animal dataset would facilitate the community to research more advanced machine learning models and explore new tasks to aid endangered animal conservation. Our dataset will be released with the paper. Our dataset can be found at https://LoTE-Animal.github.io*

1. Introduction

Protecting endangered wildlife is becoming increasingly challenging as global biodiversity declines [16]. Accurate information about wild animals is crucial for implementing effective conservation measures [9, 68]. Various techniques have been developed to monitor wildlife [62, 38, 31, 52, 4, 60, 34], but gathering reliable information remains difficult [3]. Endangered animals are scarce and wary, making it hard to track their movements [68]. And wildlife is often aggressive, making it impossible to implant sensors [71, 61]. Additionally, human interference could disturb animal behavior, causing data collection to deviate from natural conditions [71, 61].

Camera trapping is an effective solution to these problems, allowing for the collection of wildlife image data while ensuring animal welfare [38, 56]. However, camera trap data is vast and contains a significant amount of irrelevant information. Manual sorting and analysis are not only inefficient but also imprecise [38, 26, 62, 66].

Computer vision technology can automatically extract and analyze image information, relying on robust deep learning models for accurate and efficient results [14, 70, 54]. For this reason, it is essential to establish comprehensive large-scale datasets to provide deep learning algorithms with training and testing data [14, 70, 54]. However, existing datasets for animal protection are limited in the following ways:

(1) **Shortage of wildlife data.** Current datasets focus on common domestic or zoo animals [42, 2, 11, 14, 7, 50, 76], whose behaviors are influenced by human interaction, and they may not exhibit natural behavior in the wild. Because wildlife data is scarce, these datasets lack the potential to develop cross-domain models.

(2) **Short time span and discontinuous space.** Some datasets are collected from scattered network images [79], short-term monitoring data [54], or even synthetic images [54]. These data cannot reflect the long-term living status of wild animals. Continuous spatiotemporal data is required to study the specific habits and migration patterns of endangered animals, while discontinuous data is also inadequate for developing complex models related to wildlife growth patterns [80, 48, 37].

(3) **Limited tasks for testing.** Most datasets have only

*These authors contribute equally to this work.
†Corresponding authors: Bochuan Zheng, Jifeng Ning, Jindong Zhang.

one or two data annotations which limit their usefulness for multitask training. Furthermore, they lack environmental information, such as day and night, weather, and location. This makes it difficult to develop accurate ecological prediction models using these datasets [19, 74, 69].

To address these issues, we present LoTE-animal, a long-term and continuous dataset for endangered animal behavior understanding. LoTE-animal has the following three features:

(1) **Abundant wildlife data in natural habitats**. We curated data of 11 endangered wildlife species from the Wolong National Nature Reserve, all of which were captured by trap cameras with minimal human interference. The dataset consists of 10k video sequences and 28k images. We also created a subset of 7k web images that can be used to enhance the generalization performance of deep learning models.

(2) **Long temporal span and spatial continuity**. We collected the raw data of LoTE-animal dataset by monitoring with infrared-triggered trap cameras for 12 years, during which the cameras were fixed at specific locations with clear geographic information. The recorded data includes videos and images of the same population at different stages in the same location, providing valuable information for wildlife research. It is worth noting that the monitoring is still ongoing.

(3) **Rich scene and annotation information**. LoTE-animal dataset provides annotated information on different weather conditions, seasons, and habitat environments, and records images of wildlife in different growth stages under these conditions. We annotated the images with bounding boxes, segmentation masks, skeletal keypoints, and action labels, making them suitable for various computer vision tasks.

In this paper, we trained and tested representative computer vision models based on the wild and web subsets. We also evaluated the generalization performance of the models trained on the web subset. Our results can serve as a reference for the development of deep learning algorithms.

2. Related works

In this section, we introduce easily accessible and open-source animal datasets for animal behavior understanding, and highlight the rarity and potential value of endangered animal datasets.

Previous animal datasets are typically image-based or dedicated to classification tasks. For example, image classification datasets such as Dogs vs Cats [35], Animals with Attributes 2 (AwA2) [75], Bee or Wasp [59], and Fish Recognition Ground-Truth [11, 12] are commonly used. Some special-interest datasets are tailored to specific environments and only for actions performed by specific animals, such as Sheep [55], Cattle [45], Pigs [46], and Salmon [49].

Open-source animal datasets have inspired many works for animal behavior understanding. We introduce several easily accessible and open-source animal datasets in Tab. 1.

These datasets provide pose estimation annotations and additional bounding box annotations, which are provided by Poselets [13] and Animal Pose [14]. Furthermore, TigDog [23, 24], BADJA [57, 8], Synthetic Animal [51], and StanfordExtra [7] include instance segmentation annotations, while TigDog [23, 24] and Animal Kingdom [54] also provide video labels for action recognition.

Poselets [13] includes 2D pose and bounding box annotations for all animals in the PASCAL 2011 [27] dataset. Animal Pose [14] extended the dataset size by annotating more images from Poselets [13] and Animals 10 [1], and has two subsets: Subset 1 includes five categories from Poselets [13], and Subset 2 includes seven categories with only bounding box annotations from Animals 10 [1]. StanfordExtra [7] released a dataset with 20 keypoints and binary silhouette for each image from StanfordDogs [25, 39]. Although there are no pose labels in StanfordDogs [25, 39], the comprehensiveness of the canine family's images contained in this dataset make it a useful research material for object segmentation and pose estimation [7]. BADJA [57, 8] is a video-based pose annotation animal dataset containing 11 video sequences. By far one of the largest datasets available, AP-10K [79] focuses solely on pose estimation. Animal Kingdom [54] contains 850 species, making it the most species-rich dataset, but the number of each species is small. Table 1 summarizes dataset comparisons. Among them, the datasets demonstrating quantity advantages are Synthetic Animal (50K) [51], iWildCam (280K) [63], Mouse(1000K) [63] and AcinoSet(119K) [41]. Synthetic Animal [51] is a synthetic dataset, while iWildCam [63] comprises real-world wildlife images encompassing 263 distinct categories, primarily focused on wildlife classification recognition. Mouse [63] consists of 6 million frames of unlabeled tracked poses of interacting mice, as well as over 1 million frames with tracked poses and corresponding frame-level behavior annotations. A distinctive addition is the AcinoSet [41], a wildlife cheetah dataset. This dataset incorporates unmarked animal pose estimation using DeepLabCut, furnishing 2D keypoints across 119K frames. Remarkably, only 7.5K image frames underwent manual annotation for this dataset.

These datasets are available from various sources, including the web, domestic environments, and zoos. For example, some datasets are collected from the web, such as Poselets [13], TigDog [23, 24], BADJA [57, 8], Animals 10 [2], Animal Pose [14], AP10K [79], and Animal Kingdom [54]. Other datasets are sourced from domestic environments, such as Stanford Dogs [42] and StanfordExtra [7], while some are sourced from zoos, such as ATRW

Year	Dataset	Species	Source	Tasks				Other task
				Action recognition (video)	Obeject detection (image)	Instance segementation (image)	Pose estimation (image)	
2011	Stanford Dogs [42]	dogs(120 breeds)	domestic	-	-	-	20K	-
2012	Poselets [13]	dog, cat, horse, sheep, cow	web	-	6.2K	-	6.2K	-
2016	TigDog [23, 24]	dog, horse, tiger	web	6.3K	-	6.3K	6.3K	-
2018	BADJA [57, 8]	11 animals	web	-	-	6.7K	-	3D tracking
2018	Animals 10 [2]	dog, cat, horse, spider, butterfly, chicken, sheep, cow, squirrel, elephant	web	-	-	-	26K	-
2019	Animal Pose [14]	dog, cat, cow, horse, sheep and the other 7 categories	web	-	971	-	3.7K	Domain adaption
2019	Synthetic Grevy's Zebra [82]	zebra	synthetic	-	-	-	12K	3D reconstruction
2019	ATRW [44]	amur tiger	zoos	-	-	-	8K	Re-identification
2020	Synthetic Animal [51]	hound, tiger, horse, sheep, elephant	synthetic	-	-	50K	50K	-
2020	StanfordExtra [7]	dogs(120 breeds)	domestic	-	-	12K	12K	3D reconstruction
2021	iWildCam [5]	263 species	wild	-	280K	-	-	-
2021	AcinoSet [41]	cheetah	wild	-	-	-	119K	-
2021	Horse 10 [50]	horse	zoos	-	-	-	8.1K	-
2021	AP10K [79]	23 animal families and 54 species	web	-	-	-	10K	Cross-domain
2021	Mouse [63]	mouse		-	-	-	1000K	Classic Classification, Annotation Style Transfer, New Behaviors
2022	Animal Kingdom [54]	850 species	web	30K	-	-	33K	Video grounding
2023	**LoTE-animal**	**11 endangered species**	**wild**	**10K**	**35K**	**35K**	**35K**	**Cross-domain, semi-supervised, self-supervised**

Table 1: The comparison of other open-source animal datasets. LoTE-animal is shown in **bold**.

[44] and Horse 10 [50]. Zuffi Silvia et al. [82] and Mu Jiteng et al. [51] also attempted to use synthetically generated animals for pose estimation. However, these synthetic datasets are different from real-life animals and may not fully capture the complexities and variations in their behaviors. While some datasets, including TigDog, BADJA, and Animal Kingdom [23, 24, 8, 54, 79], do contain a small number of endangered species, they are typically obtained from documentaries rather than in-situ observations. To truly understand and protect endangered species, it is crucial to collect data directly from their natural habitats through in-situ observations.

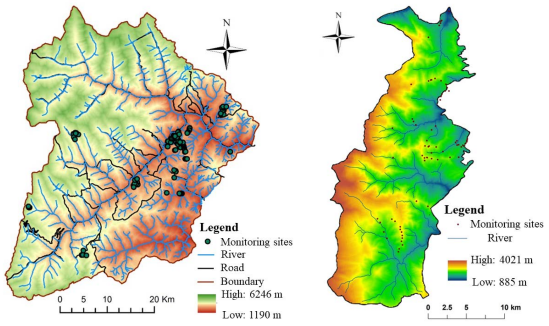
3. Dataset construction

Our dataset consists of two parts: one obtained from the web and the other from trap cameras in undisturbed settings. The web data was obtained through web crawlers that searched for relevant keywords and retrieved 7K images from various online image repositories. The wild data was collected by installing trap cameras in natural habitats of endangered animals. We selected 11 endangered species as our research subjects and annotated their behaviors. All data, including the web-scraped portion, were filtered and classified by the Wildlife Habitat Research Center to ensure the accuracy of the dataset.

3.1. Wild data collection and collation

Data collection. The study area is situated in the Sichuan Wolong National Nature Reserve in southwest China, with geographical coordinates of $102^{\circ}52^{'} - 103^{\circ}24^{'}$E and $30^{\circ}45^{'} - 31^{\circ}25^{'}$N, covering an area of 2,000km^2 and an elevation ranging from 1,150m to 6,250m. The Wolong National Nature Reserve serves as the primary location for data collection, accounting for 95% of the dataset, with monitoring sites depicted in Fig. 1a. In addition, we included supplementary data from the adjacent Mabian Dafengding Nature Reserve, which features the same species, accounting for 5% of our dataset, as shown in Fig. 1b.

The data collection area spans an elevation range of



(a) Wolong (b) Mabian Dafengding

Figure 1: The map of monitoring sites in wild.

1,806 m to 4,445 m, which encompasses the primary distribution altitude of endangered animals. To capture animal activity within this range, we deployed over 200 infrared-triggered camera traps (Ltl 5210A) at multiple sites, as illustrated in Fig. 2.

Figure 2: The infrared-triggered camera-trapping in wild.

Data collation. We consulted the Mammal Diversity and Geographical Distribution in China catalog [40] to identify, record, and tally the species observed in valid videos. We verified the endangered status of each animal with the International Union for Conservation of Nature (IUCN) Red List of Threatened Species [36]. To the fullest extent, we divided different animals into orders, genera, and families based on the Catalogue of Mammals in China (2021) [81], which fully reflects the biological relationship between species.

The species composition is organized into four orders (①-④), seven families (⑤-⑫), and 11 genera (⑬-㉓), as depicted in Fig. 3. There is only one species in the order *Rodentia*, while the order *Carnivora* is represented by three species and *Primates* by two. The largest representation belongs to the order *Cetartiodactyla*, with five species, making up more than 50% of the total species.

3.2. Manual annotated plan

We followed the COCO standards and annotated multiple tasks including object detection, instance segmentation, pose estimation, and action recognition. The animal labels for the object detection and instance segmentation tasks are species name. For the pose estimation task, we carefully

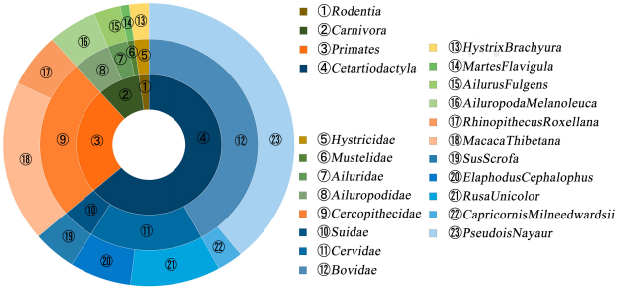


Figure 3: Endangered animal species compasition in LoTE-animal.

compared and considered the definitions of existing animal keypoints, and ultimately selected 17 skeletal keypoints to describe the variations among different species. Lastly, for the action recognition task, we curated typical behaviors of each animal for annotation. Additionally, we annotated three types of information that describe the animal's life pattern, such as season, weather, day and night, to provide more contextual information for animal behavior analysis.

During the annotation process, our annotators underwent rigorous training on the physiognomy, body structure, and distribution of keypoints for each animal species. Highly skilled annotators were then selected for further training on dealing with partial occlusion, and actively participated in the subsequent annotation process. They were instructed to label all visible keypoints, and for occluded keypoints, to estimate their locations based on the animal's body plan, pose, and symmetry property. Any keypoints whose locations could not be accurately estimated were left unannotated. Three annotators were assigned to each image. To address variability among the annotators, we adopted consensus based annotation, in which multiple annotators are asked to annotate the same image, and the final annotation is determined by a consensus among the annotators. This ensured that the resulting dataset contained accurate and high-quality annotations.

4. Dataset statistics

In this section, we present the distribution statistics of our dataset and annotation statistics for each annotation task.

4.1. Data distribution

In Tab. 2, we present the LoTE-animal species list and their corresponding conservation statuses in China. This dataset includes vital umbrella species for safeguarding endangered wildlife, like giant pandas, red pandas, and Sichuan golden monkeys. Additionally, it features primates of high importance to human evolutionary research, including the Tibetan macaque. Notably, these 11 species are fo-

cal conservation targets in China and are listed under the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). This unique status lends our dataset distinct value.

To begin, we present the distribution statistics of our dataset, which include the order, genus, and family distribution, as shown in Fig. 4.

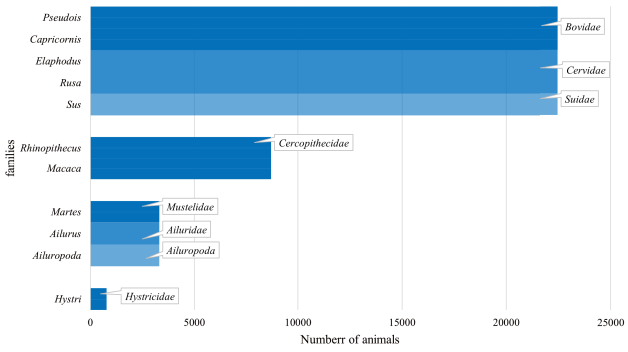


Figure 4: The order, genus and family distribution.

For the domain adaptation task, we design a small source domain dataset by crawling 7K web images of the same species. We then compare this small dataset with the wild subset collected from the Wolong National Nature Reserves, as shown in Fig. 5.

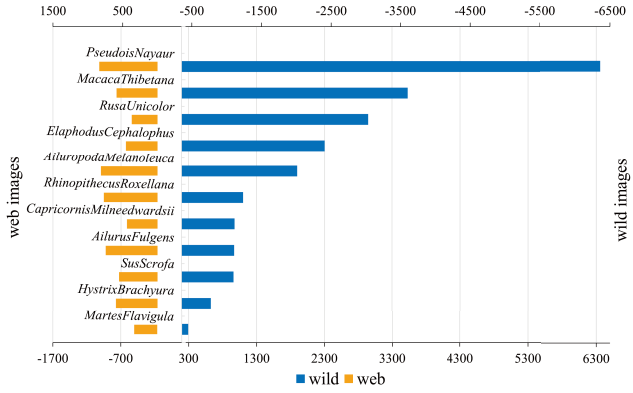


Figure 5: The comparison of web subset and wild subset in LoTE-animal.

Wolong National Nature Reserve, located within the Qinghai-Tibet Plateau's climate region, is characterized by distinct ecological seasons. We recorded the season as additional information and found that the dataset contains the highest number of images during spring and autumn.

Furthermore, each image was annotated with its time of capture, indicating whether it was taken during the day or at night. Interestingly, most animals in the dataset were captured during the daytime, with the exception of the Malay porcupine.

4.2. Annotation statistics

In this section, we present the annotation statistics of our dataset, which include bounding box, segmentation mask, skeletal keypoints, and action labels for endangered animals.

Our team annotated the entire dataset, a time-consuming task that required rich computer vision annotation techniques. In total, we annotated about 35K bounding boxes, 34K segmentation polygons, and 425K skeletal keypoints, as shown in Fig. 6.

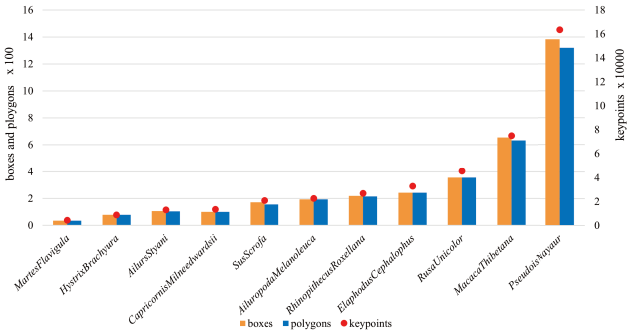


Figure 6: The annotation statistics. The primary axis (left) is the number of bounding boxes and polygons (mask) and the secondary axis (right) is the number of skeletal keypoints.

We also gathered statistics on the number of boxes in each image, revealing that approximately 15K images contain only one animal, while around 6.5K images contain between two and six animals, with a few images featuring more than seven animals. Social animals such as *PseudoisNayaur*, *MacacaThibetana*, and *RhinopithecusRoxellana*, tend to appear in large groups, with some images featuring more than 20 animals. On average, there are 1.6 animals per image, as demonstrated in Fig. 7.

Furthermore, we gathered data on the area ratios of each animal in comparison to the entire image, as demonstrated in Fig. 8. Typically, the majority of the animals have an area ratio peak of 0.1, indicating that smaller animals are more prevalent. Additionally, we observed that a smaller number of animals have an area ratio of more than 0.3 in LoTE-animal, indicating a scarcity of larger animals in the dataset.

Moreover, we also examined the skeletal keypoints of the animals, as depicted in Fig. 9. From these statistics, we can infer that predicting the skeletal keypoints of animal limbs (left front palm, right front palm, left back palm, and right back palm) is challenging.

To further investigate animal behavior, we also collected statistics on the number of behavior video clips, encompass-

Protection level	Specie	Specie(*Latin*)	Category(*Latin*)	Family(*Latin*)	Order(*Latin*)
I	Giant Panda	*Ailuropoda Melanoleuca*	*Ailuropoda*	*Ailuropodidae*	*Carnivora*
I	Golden Snub-nosed Monkey	*Rhinopithecus Roxellana*	*Rhinopithecus*	*Cercopithecidae*	*Primates*
II	Red Panda	*Ailurus Fulgens*	*Ailurus*	*Ailuridae*	*Carnivora*
II	Yellow-throated Marte	*Martes Flavigula*	*Martes*	*Mustelidae*	*Carnivora*
II	Tibetan Macaque	*Macaca Thibetana*	*Macaca*	*Cercopithecidae*	*Primates*
II	Wild Boar	*Sus Scrofa*	*Sus*	*Suidae*	*Cetartiodactyla*
II	Sambar	*Rusa Unicolor*	*Rusa*	*Cervidae*	*Cetartiodactyla*
II	Tufted Deer	*Elaphodus Cephalophus*	*Elaphodus*	*Cervidae*	*Cetartiodactyla*
II	Chinese Serow	*Capricornis Milneedwardsii*	*Capricornis*	*Bovidae*	*Cetartiodactyla*
II	Blue Sheep	*Pseudois Nayaur*	*Pseudois*	*Bovidae*	*Cetartiodactyla*
III	Porcupine	*Hystrix Brachyura*	*Hystri*	*Hystricidae*	*Primates*

Table 2: The endangered animal species list of LoTE-animal. I, II, III is wildlife protection levels and endangered levels in China. *Italics* are Latin names.

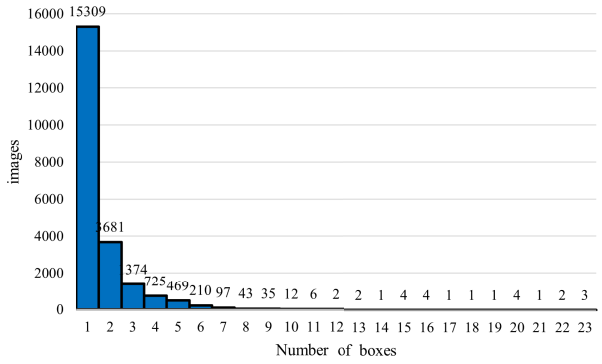


Figure 7: Number of boxes in each image.

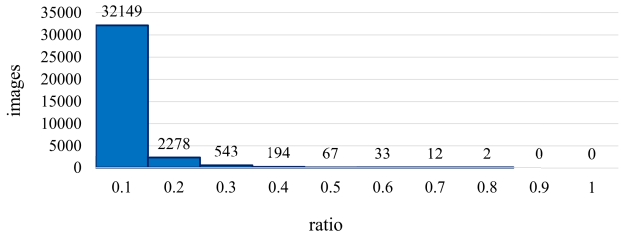


Figure 8: The area ratios of each animal in one image.

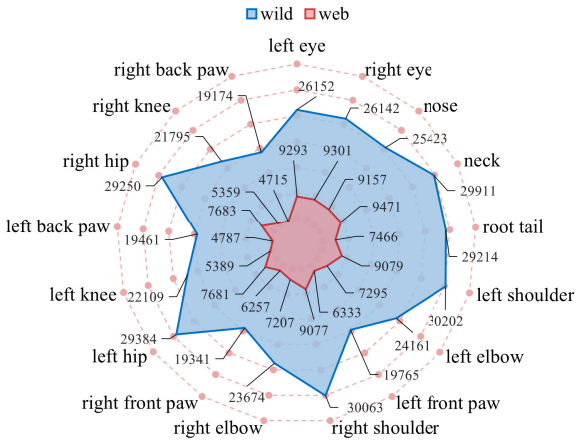


Figure 9: The distribution of skeletal keypoints.

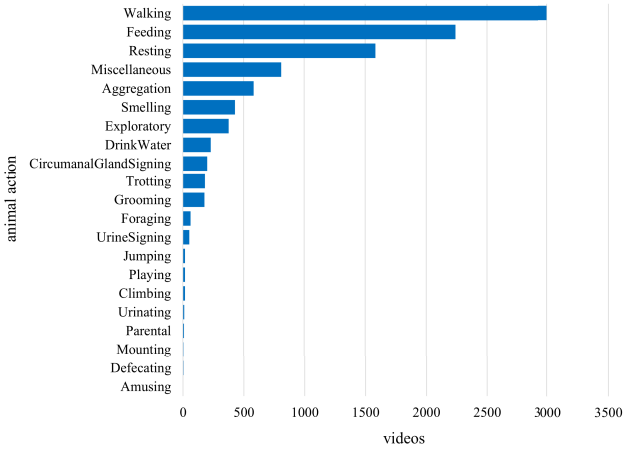


Figure 10: Number of videos of each animal behavior in the dataset.

ing 21 common and special behaviors of endangered animals. The distribution of animal behavior is long-tailed, as shown in Fig. 10.

Finally, we provide some annotation samples in Fig. 11, including images of single animals, images with two animals, and images with multiple occluded animals. We also showcase images of small object animals in the fourth row.

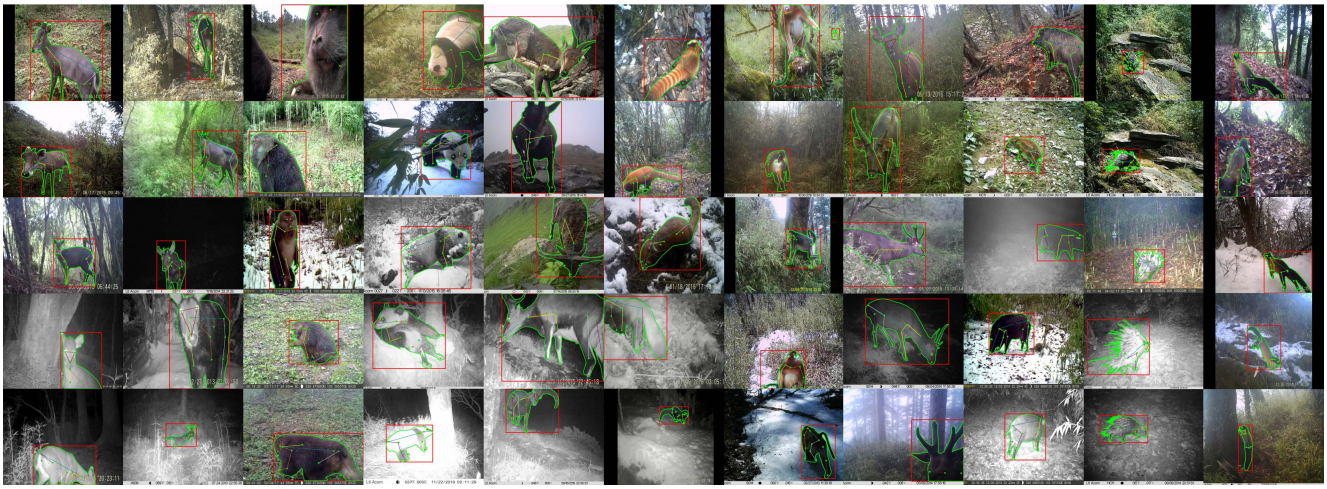


Figure 11: A glance at animal species in LoTE-Animal.

5. Experiment

In this section, we present our experimental setup and evaluation results for the proposed endangered animal dataset. Specifically, we perform four types of supervised learning tasks and three types of cross-domain tasks to compare model performance. Our objective in conducting these experiments is to gain a deeper understanding of the dataset by assessing the performance of common computer vision techniques on it.

5.1. Implementation details

We implement experiments in several representative computer vision models on 4 GeForce RTX 2080Ti GPU with 48GB or 4 GeForce RTX 3090 GPU with 96GB. The PyTorch 1.8.0 deep learning architecture is used, and the programming language is Python 3.8. Unless otherwise specified, all models are implemented using the publicly available MMLAB framework [17, 20, 21, 22] or officially released code by default. Hyperparameters are set uniformly across all experiments, and the models are optimized with default parameters in the codebase. We terminate training when the loss converges stably, or when the training schedule (measured in epochs) is completed. Specifically, we use a training schedule denoted by 1×for 12 epochs, 2×for 24 epochs, and 3×for 36 epochs. In addition, we report evaluation results under the average precision (AP) metric, with cross denoting cross-domain adaptation tasks.

5.2. Results of supervised learning and cross-domain tasks

We randomly divided the web and wild images into two separate train set, validation set, and test set in a ratio of 7:1:2 for each species. We transplant some popular computer vision frameworks for comparison. We also implemented cross domain tasks for object detection, instance segmentation, pose estimation to analyze and compare web images without behavior information and wild images with behavior information.

There are two kinds of experiments, one is the performance comparison of models trained separately in web column and wild column, the other is the comparison of generalization performance in web subset web column and cross column, as shown in Tab. 3. We visualize the AP performance of the three cross-domain tasks, as shown in Fig. 12

Our experiments reveal significant improvements in the performance of object detection and instance segmentation algorithms on both the web and wild subsets. Specifically, the r50 accuracy of object detection increased from 55.45% to 70.65% and from 69.75% to 76.13% on the web and wild subsets, respectively, with similar improvements observed for the r101 model. For instance segmentation, the r50 accuracy increased from 57.16% to 64.98% and from 68.47% to 67.00% on the web and wild subsets, respectively, with similar improvements observed for the r101 model. The evaluated the pose estimation method showed lower accuracy, ranging from 36.55% to 51.17% on the web subset and from 36.55% to 43.74% on the wild subset. The accuracy of the action recognition method is only tested on the wild subset with the lowest is 24.76% and the highest is 60.02%.

In the second type of experiment, we evaluated cross-domain adaptation performance on two subsets: training on the web or wild subset and testing on the other opposite subset. The results in the cross column of Tab. 3 indicate that all models exhibit a significant decrease in performance when tested on the cross subset. Specifically, for object detection models, the performance decrease ranges from 52.49% to 60.18%. For instance segmentation models, the decrease is between 51.95% and 57.96%. For pose estimation models,

Model	Year	Schedule	Backbone	AP				$AP_{.5}$				$AP_{.75}$			
				web	wild	web ↓ wild	wild ↓ web	web	wild	web ↓ wild		web	wild	web ↓ wild	
Object Detection															
Faster R-CNN [58]	NeurIPS2015	1×	r50	55.45	69.75	26.35	26.54	85.11	93.07	45.78		63.15	81.35	27.29	
			r101	58.83	71.04	31.22	29.55	86.27	93.30	49.32		67.08	81.93	35.49	
FCOS [67]	ICCV2019	1×	r50	40.59	70.13	16.16	20.16	60.87	91.90	25.26		44.75	80.15	17.75	
			r101	42.16	72.31	17.02	22.49	62.27	93.27	26.44		46.75	82.15	18.88	
DETR [15]	ECCV2020	1×	r50	56.10	39.70	10.70	6.25	71.60	60.97	20.30		51.90	44.37	10.10	
Sparse R-CNN [65]	CVPR2021	3×	r50	65.01	73.41	29.06	33.51	83.34	94.48	42.36		70.42	82.18	31.75	
			r101	68.25	74.42	28.63	35.62	83.86	94.49	41.00		74.23	82.73	30.94	
TOOD [30]	ICCV2021	2×	r50	68.45	75.09	29.84	21.87	85.14	93.83	41.41		75.25	84.75	33.10	
			r101	69.58	75.79	32.38	25.01	85.60	94.09	44.34		76.16	84.58	35.32	
DiffusionDet [18]	arXiv2022	2×	r50	70.65	76.13	30.49	32.59	86.45	95.34	42.80		74.76	85.01	32.71	
			r101	69.75	76.23	27.48	35.20	85.00	95.40	37.57		74.00	85.60	29.72	
Instance Segementation															
Mask R-CNN [32]	ICCV2017	1×	r50	57.16	68.47	27.04	24.65	82.91	93.89	42.58		65.64	82.12	30.87	
			r101	59.59	69.47	28.64	28.37	83.86	93.91	43.22		68.72	84.50	33.74	
YOLOACT [10]	ICCV2019	55e	r50	57.48	61.41	24.16	30.14	80.78	89.33	37.67		63.17	72.08	26.65	
			r101	58.39	62.47	26.03	34.37	81.03	90.33	40.00		64.61	73.41	29.60	
SOLOv2 [73]	ECCV2020	55e	r50	62.46	63.86	27.56	12.90	85.48	89.28	38.84		75.22	73.92	31.60	
			r101	63.83	64.64	28.45	10.56	83.73	91.23	40.70		69.56	75.00	32.20	
PointRend [43]	CVPR2020	3×	r50	66.59	73.42	29.98	25.71	86.06	95.05	41.38		75.91	86.90	34.90	
QueryInst [28]	ICCV2021	55e	r50	64.98	67.00	30.85	29.70	86.32	93.12	46.14		75.53	79.69	35.83	
			r101	67.20	71.84	33.80	29.45	87.02	96.31	48.29		76.83	87.25	39.99	
Pose Estimation															
HourglassNet [53]	ECCV2016	210e	hourglass	46.93	41.86	21.99	36.60	85.04	83.58	52.48		45.75	37.03	14.51	
ResNet [33]	CVPR2016		r50	46.05	41.02	21.33	34.93	85.93	82.31	53.80		42.91	35.91	12.39	
			r101	48.89	41.15	22.08	34.19	86.73	82.92	54.85		48.95	35.85	13.95	
HRNet [64]	CVPR2019		w32	51.00	43.74	25.74	41.13	88.54	84.56	59.93		52.87	39.94	17.77	
			w48	51.17	43.25	26.30	41.58	88.55	84.82	58.47		52.70	39.46	17.85	
SCNet [47]	CVPR2020		r50	46.93	40.65	21.55	34.40	86.84	82.97	54.80		44.64	35.36	12.87	
			r101	47.07	40.56	20.85	34.79	86.62	83.00	53.45		46.48	34.85	12.37	
LiteHRNet [78]	CVPR2021		r18	39.21	36.55	16.45	28.27	81.20	80.90	45.76		32.00	28.42	7.71	
			r30	38.94	37.62	16.05	27.62	80.86	81.01	44.86		31.35	29.50	8.44	
PVTV2 [72]	CVMJ2022		b2	48.85	40.77	23.48	36.25	88.15	82.35	58.65		49.02	34.75	14.75	
Action Recognition															
				top 1				top5				mean accurary			
SlowOnly [29]	ICCV2019	256e	r50	79.39				98.39				60.02			
		196e	r101	68.98				97.03				42.44			
SlowFast [29]	ICCV2019	256e	r50	71.79				97.54				43.88			
			r101	N/A				N/A				N/A			
TPN [77]	CVPR2020	150e	r50	55.81				94.32				24.76			
TimeSformer [6]	ICML2021	15e	spaceOnly	61.70				96.68				39.04			
			jointST	68.63				97.89				40.72			
			divST	70.24				97.99				43.29			

Table 3: The results of four supervised learning vision tasks and three cross-domain tasks on two subsets, respectively. The underlined models are trained on 3090 GPU and the rest on 2080Ti GPU. The best, second and third results are shown in red, orange, blue.

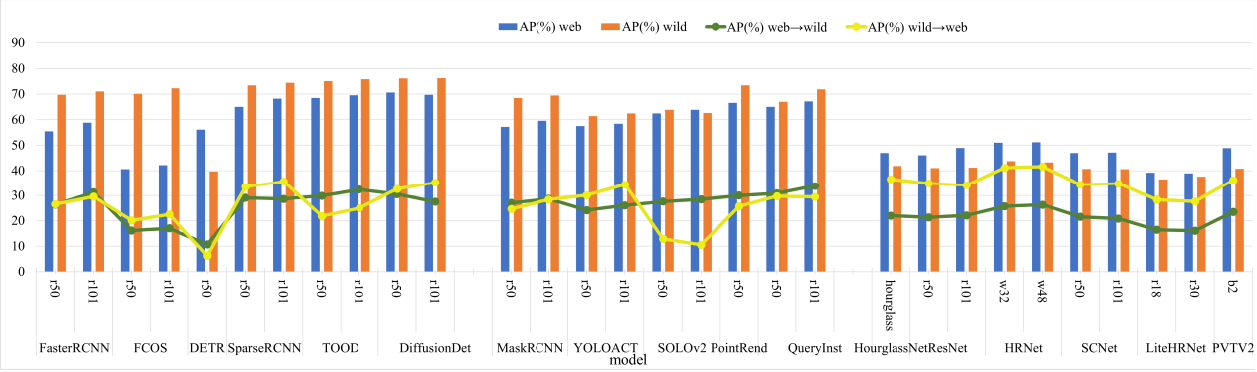


Figure 12: The AP performance of meodels on web dataset, wild dataset, web→wild and wild→web.

the decrease ranges from 49.54% to 54.83%.

6. Discussion

Object detection and instance segmentation algorithms applied to the LoTE-animal dataset have shown significant improvements in accuracy with the evolution of algorithms. In cross-domain instance segmentation tasks, the generalization ability has also been gradually enhanced. However, the overall accuracy drop in cross-domain tasks is still relatively significant. Therefore, when constructing algorithms, in addition to considering accuracy improvement, enhancing its generalization performance should also be taken into account, as direct transfer of web models has significant limitations. Analysis of the results on the wild test set reveals that the proportion of wildlife images in the missed detection and missed segmentation images is very low.

Pose estimation requires first locating the animal's skeletal keypoints and then making pose judgments, which is more challenging than object detection and instance segmentation. Therefore, the accuracy of the results is significantly lower than that of the former two. HourglassNet, ResNet, and HRNet are improved networks for animal pose estimation [79], although their performance is significantly lower than that of newly proposed networks such as SCNet [47], LiteHRNet [78], and PVTV2 [72] in human pose estimation. However, after training on the LoTE-animal dataset, the former models outperform the latter, indicating that there are differences between human and animal pose estimation, and simple transfer cannot fully leverage their strengths. The results show that although the web subset has fewer samples, its accuracy is higher than that of the wild subset after training, which is because the animal images in the web subset are generally tracked and the animal's main target is complete after selection, while in wild data, animal limbs are more frequently occluded and located at the edges, making the task more difficult. However, some web models can achieve more than 60% of the performance of wild models in cross-domain testing, indicating that models established based on web data for pose estimation tasks have certain generalization capabilities, and new algorithm construction can explore their potential.

For action recognition task, the statistical information of the dataset shows a long-tailed distribution of animal videos. The lack of training samples results in decreased accuracy. Therefore, in algorithm construction, it is a worthwhile direction to enhance the training of tail-end images to improve model accuracy.

At present, there are significant limitations to datasets constructed based on network data, and their performance is relatively poor in practical applications. In the short term, if developing deep learning models for specific endangered animal analyses based on existing algorithms, continuous time and space-uninterrupted raw data need to be collected to construct datasets, but this method is difficult. Another approach is to develop deep learning models with generalization capabilities, such as developing cross-domain deep learning models using transfer learning. The dataset established in this study provides a foundation for the development and testing of such models.

7. Conclusion

In this paper, we present a animal dataset with a long time span, and simultaneously incorporate endangered animals computer vision tasks. Our evaluation of the dataset using mainstream deep learning algorithms yielded valuable insights for optimizing algorithms designed specifically for wildlife conservation. Our work is expected to inspire further research in animal behavioral analysis and understanding, which is vital for the preservation of endangered species and ecological balance.

Acknowledgements. This project is supported by National Natural Science Foundation of China (Grant No.61876153, 62176217, U21A20193 and 42071279) and Interdisciplinary Research Fundation for Doctoral Candidates of Beijing Normal University (Grant BNUXKJC2221).

References

[1] Corrado Alessio. Animals-10 dataset. (2022, Aug 23). 2

[2] C Alessio. animals-10 dataset. *Accessed: Dec*, 2019. 1, 2, 3

[3] Jeanne Altmann. Observational study of behavior: sampling methods. *Behaviour*, 49(3-4):227–266, 1974. 1

[4] Jake Astill, Rozita A Dara, Evan DG Fraser, Bruce Roberts, and Shayan Sharif. Smart poultry management: Smart sensors, big data, and the internet of things. *Computers and Electronics in Agriculture*, 170:105291, 2020. 1

[5] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset, 2021. 3

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. 8

[7] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision*, pages 195–211. Springer, 2020. 1, 2, 3

[8] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *Asian Conference on Computer Vision*, pages 3–19. Springer, 2018. 2, 3

[9] Simon A Black. Assessing presence, decline, and extinction for the conservation of difficult-to-observe species. In *Problematic Wildlife II*, pages 359–392. Springer, 2020. 1

[10] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019. 8

[11] Bastiaan J Boom, Phoenix X Huang, Cigdem Beyan, Concetto Spampinato, Simone Palazzo, Jiyin He, Emmanuelle Beauxis-Aussalet, Sun-In Lin, Hsiu-Mei Chou, Gayathri Nadarajan, et al. Long-term underwater camera surveillance for monitoring and analysis of fish populations. *VAIB12*, 2012. 1, 2

[12] Bastiaan J Boom, Phoenix X Huang, Jiyin He, and Robert B Fisher. Supporting ground-truth annotation of image datasets using clustering. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1542–1545. IEEE, 2012. 2

[13] Lubomir Bourdev and Jitendra Malik. Dataset of keypoints and foreground annotations for all categories of pascal 2011, 2012. 2, 3

[14] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9498–9507, 2015. 1, 2, 3

[15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 8

[16] Gerardo Ceballos, Paul R Ehrlich, and Peter H Raven. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proceedings of the National Academy of Sciences*, 117(24):13596–13602, 2020. 1

[17] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[18] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 8

[19] Sylvain Christin, Éric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644, 2019. 2

[20] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 7

[21] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 7

[22] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 7

[23] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Articulated motion discovery using pairs of trajectories. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2151–2160, 2015. 2, 3

[24] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Behavior discovery and alignment of articulated object classes from unstructured video. *International Journal of Computer Vision*, 121(2):303–325, 2017. 2, 3

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[26] Anna Rita Di Cerbo and Carlo M Biancardi. Monitoring small and arboreal mammals by camera traps: effectiveness and applications. *Acta Theriologica*, 58(3):279–283, 2013. 1

[27] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html. 2

[28] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6910–6919, October 2021. 8

[29] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 8

[30] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, 2021. 8

[31] Arthur Francisco Araújo Fernandes, João Ricardo Rebouças Dórea, and Guilherme Jordão de Magalhães Rosa. Image analysis and computer vision applications in animal sciences: an overview. *Frontiers in Veterinary Science*, 7:551269, 2020. 1

[32] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 8

[33] 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 (CVPR)*, June 2016. 8

[34] J. Hou, Y. He, H. Yang, T. Connor, and S. Zhou. Identification of animal individuals using deep learning: A case study of giant panda. *Biological Conservation*, 242:108414, 2020. 1

[35] Microsoft Research Institute. Dogs vs. Cats. (2022, Aug 14). 2

[36] IUCN. The iucn red list of threatened species. https://www.iucnredlist.org, 2022. 4

[37] Avelino Javer, Michael Currie, Chee Wai Lee, Jim Hokanson, Kezhi Li, Céline N Martineau, Eviatar Yemini, Laura J Grundy, Chris Li, QueeLim Ch'ng, et al. An open-source platform for analyzing and sharing worm-behavior data. *Nature methods*, 15(9):645–646, 2018. 1

[38] ZOE Jewell. Effect of monitoring technique on quality of conservation science. *Conservation Biology*, 27(3):501–508, 2013. 1

[39] Le Jiang, Caleb Lee, Divyang Teotia, and Sarah Ostadabbas. Animal pose estimation: A closer look at the state-of-the-art, existing gaps and opportunities. *Computer Vision and Image Understanding*, page 103483, 2022. 2

[40] Zhigang Jiang. *China's mammal diversity and geographic distribution*. Science Press, 1 edition, 8 2015. An optional note. 4

[41] Daniel Joska, Liam Clark, Naoya Muramatsu, Ricardo Jericevich, Fred Nicolls, Alexander Mathis, Mackenzie W Mathis, and Amir Patel. Acinoset: a 3d pose estimation dataset and baseline models for cheetahs in the wild. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13901–13908. IEEE, 2021. 2, 3

[42] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011. 1, 2, 3

[43] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 8

[44] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: a benchmark for amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019. 3

[45] Yun Liang, Fuyou Xue, Xiaoming Chen, Zexin Wu, and Xiangji Chen. A benchmark for action recognition of large animals. In *2018 7th International Conference on Digital Home (ICDH)*, pages 64–71. IEEE, 2018. 2

[46] Dong Liu, Maciej Oczak, Kristina Maschat, Johannes Baumgartner, Bernadette Pletzer, Dongjian He, and Tomas Norton. A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs. *Biosystems Engineering*, 195:27–41, 2020. 2

[47] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 8, 9

[48] Tim CD Lucas. A translucent box: interpretable machine learning in ecology. *Ecological Monographs*, 90(4):e01422, 2020. 1

[49] Håkon Måløy, Agnar Aamodt, and Ekrem Misimi. A spatio-temporal recurrent network for salmon feeding action recognition from underwater videos in aquaculture. *Computers and Electronics in Agriculture*, 167:105087, 2019. 2

[50] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1859–1868, 2021. 1, 3

[51] 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. 2, 3

[52] Suresh Neethirajan. The role of sensors, big data and machine learning in modern animal farming. *Sensing and Bio-Sensing Research*, 29:100367, 2020. 1

[53] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 8

[54] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034, 2022. 1, 2, 3

[55] Kehinde Owoeye and Stephen Hailes. Online collective animal movement activity recognition. *arXiv preprint arXiv:1811.09067*, 2018. 2

[56] Stuart L Pimm, Sky Alibhai, Richard Bergl, Alex Dehgan, Chandra Giri, Zoë Jewell, Lucas Joppa, Roland Kays, and Scott Loarie. Emerging technologies to conserve biodiversity. *Trends in ecology & evolution*, 30(11):685–696, 2015. 1

[57] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 3

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jun 2017. 8

[59] Tanya Sachdev Gauri S. Kate T. carlobueza12@gmail.com Rtnewnan Pooja Sehgal Nikhilesh Jasuja

Musky Capichi1235. Rocky C. Heater, Carlo Bueza. Bee vs wasp. (2022, Nov 23). 2

[60] Abhinav Sharma, Arpit Jain, Prateek Gupta, and Vinay Chowdary. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873, 2020. 1

[61] Justine A Smith, Kaitlyn M Gaynor, and Justin P Suraci. Mismatch between risk and response may amplify lethal and non-lethal effects of humans on wild animal populations. *Frontiers in Ecology and Evolution*, 9:604973, 2021. 1

[62] Robin Steenweg, Mark Hebblewhite, Roland Kays, Jorge Ahumada, Jason T Fisher, Cole Burton, Susan E Townsend, Chris Carbone, J Marcus Rowcliffe, Jesse Whittington, et al. Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment*, 15(1):26–34, 2017. 1

[63] Jennifer J. Sun, Tomomi Karigo, Dipam Chakraborty, Sharada P. Mohanty, Benjamin Wild, Quan Sun, Chen Chen, David J. Anderson, Pietro Perona, Yisong Yue, and Ann Kennedy. The multi-agent behavior dataset: Mouse dyadic social interactions, 2021. 2, 3

[64] 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 (CVPR)*, June 2019. 8

[65] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. SparseR-CNN: End-to-end object detection with learnable proposals. *arXiv preprint arXiv:2011.12450*, 2020. 8

[66] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019. 1

[67] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019. 8

[68] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):1–15, 2022. 1

[69] John Joseph Valletta, Colin Torney, Michael Kings, Alex Thornton, and Joah Madden. Applications of machine learning in animal behaviour studies. *Animal Behaviour*, 124:203–220, 2017. 2

[70] Lukas von Ziegler, Oliver Sturman, and Johannes Bohacek. Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology*, 46(1):33–44, 2021. 1

[71] Kristen A Walker, Andrew W Trites, Martin Haulena, and Daniel M Weary. A review of the effects of different marking and tagging techniques on marine mammals. *Wildlife Research*, 39(1):15–30, 2011. 1

[72] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 8, 9

[73] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting objects by locations. In *Proc. Eur. Conf. Computer Vision (ECCV)*, 2020. 8

[74] Ben G Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545, 2018. 2

[75] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 2

[76] Yongqin Xian, H. Christoph Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018. 1

[77] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[78] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10440–10450, 2021. 8, 9

[79] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021. 1, 2, 3, 9

[80] Qiuyan Yu, Wenjie Ji, Lara Prihodko, C Wade Ross, Julius Y Anchang, and Niall P Hanan. Study becomes insight: ecological learning from machine learning. *Methods in Ecology and Evolution*, 12(11):2117–2128, 2021. 1

[81] WEI Fuwen YANG Qisen WU Yi JIANG Xuelong LIU Shaoying LI Baoguo YANG Guang LI Ming ZHOU Jiang LI Son Yibo GE Deyan LI Sheng YU Wenhua CHEN Bingyao ZHANG Zejun ZHOU Caiquan WU Shibao ZHANG Li CHEN Zhongzheng CHEN Shunde DENG Huaiqing JIANG Tinglei ZHANG Libiao SHI Hongyan LU Xueli LI Quan LIU Zhu Yaqian LI Yuchun. Catalogue of mammals in China(2021). *Acta Theriologica sinca*, 41(05):487–5016, 2021. 4

[82] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5359–5368, 2019. 3