Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding

Xun Long Ng‡   Kian Eng Ong‡   Qichen Zheng‡   Yun Ni‡   Si Yong Yeo   Jun Liu*

Information Systems Technology and Design, Singapore University of Technology and Design, Singapore

{xunlong_ng, kianeng_ong}@mymail.sutd.edu.sg   {qichen_zheng, ni_yun, siyong_yeo, jun_liu}@sutd.edu.sg

Abstract

Understanding animals' behaviors is significant for a wide range of applications. However, existing animal behavior datasets have limitations in multiple aspects, including limited numbers of animal classes, data samples and provided tasks, and also limited variations in environmental conditions and viewpoints. To address these limitations, we create a large and diverse dataset, Animal Kingdom, that provides multiple annotated tasks to enable a more thorough understanding of natural animal behaviors. The wild animal footages used in our dataset record different times of the day in extensive range of environments containing variations in backgrounds, viewpoints, illumination and weather conditions. More specifically, our dataset contains 50 hours of annotated videos to localize relevant animal behavior segments in long videos for the video grounding task, 30K video sequences for the fine-grained multi-label action recognition task, and 33K frames for the pose estimation task, which correspond to a diverse range of animals with 850 species across 6 major animal classes. Such a challenging and comprehensive dataset shall be able to facilitate the community to develop, adapt, and evaluate various types of advanced methods for animal behavior analysis. Moreover, we propose a Collaborative Action Recognition (CARe) model that learns general and specific features for action recognition with unseen new animals. This method achieves promising performance in our experiments. Our dataset can be found at https://sutdcv.github.io/Animal-Kingdom.

1. Introduction

A better understanding of how animals behave and move in the wild is not only a cornerstone of behavioral sciences [24] but also crucial for a wide range of applications. Animal behavioral analysis plays a pivotal role in conservation efforts [17, 62] and wildlife management [49]. Each year, the world spends at least $75 billion [43, 68] in conservation efforts. Before researchers can better protect the wildlife, they often need to first monitor the animals. Researchers often monitor animals using video cameras, without having to attach physical markers that may alter their natural behaviors [24]. By monitoring and analyzing the behaviors of the animals, they can gain deeper insights into their health and needs [24], and even detect movement-related injuries [24]. In addition, detecting and analyzing changes in animal behaviors allow researchers to learn about new behaviors [1, 72]. Besides, animal behavioral analysis has also helped pharmaceutical scientists understand the effects of experimental interventions on behaviors, and neuroscientists learn more about brain activity across different temporal scales [1, 24, 29, 42, 52]. These findings also help researchers better understand human diseases and develop drugs that are suitable for use in humans. Moreover, insights gained from analyzing animals' behaviors have also played an important role in aiding engineers in designing bio-inspired robots [46, 83] to efficiently perform specialised functions (*e.g.*, lifesaving). Animators and game developers can also make use of the insights to realistically render animals in animations and games [8, 83]. In summary, the analysis of animal behaviors is not only significant for ecological protection, but also significant in a wide range of applications, and thus has received more and more research attention recently [8,17,34,35,38,39,46,53,77,79].

Animal action analysis is especially important in the understanding of the health and needs of animals. Given the diversity of animals' actions, and how actions can occur and switch within split seconds, the use of video footages enables us to monitor multiple animals round the clock [1]. In addition, by localizing keypoints and employing pose estimation, we can analyze and identify changes in the animal's pose to better interpret their actions and behaviors. Before we analyze animal behaviors, we may need to first identify the frames of interest, since oftentimes a large part of the animal footages may not even contain any animal [12, 44]. However, the vast volume of wildlife animal footages inherently make it both labour and time-intensive to identify animals and actions. In view of this, employing video ground-

---

*Corresponding author.

‡These authors contribute equally to this work.

ing [47, 81] will be effective to localize the animals and behaviors of interest in long video streaming.

Previous studies [8, 72] show that a large and comprehensive dataset is important to develop robust deep learning models. However, many existing animal behavioral analysis datasets have limitations in multiple aspects: **(1)** Relatively small dataset and scarcity of extensive and well-annotated labels [72], which reduce the generalization ability and transferability of models [72]. In contrast, having a large and comprehensive dataset would be helpful to mitigate such issues. **(2)** Limited number and limited diversity of animals [1, 72], as many of the existing datasets are designed for and to study specific groups of animals (*e.g.*, mammals only). However, there is a far greater diversity of animals in nature, and the understanding of wildlife should not be restricted to only a few specific classes. Furthermore, there exists a huge variety in shapes, sizes (*e.g.*, different stages in life cycle), body patterns (*e.g.*, camouflage), and number of limbs or appendages (*e.g.*, wings) within and across animal classes. Therefore, having an extensive range and diverse representation of animals in a dataset would improve the generalization ability of the model for analysing behaviors of various animal classes [72]. **(3)** Uniform environmental conditions (*e.g.*, laboratory conditions or specific habitat) prevent the transferability of models to a different context (*e.g.*, another habitat), because the models are often not able to generalize well when the new data differs from those previously used for training [72]. Moreover, animals in the wild are naturally found at various locations, in different environmental conditions, and at different time periods, which in turn will all affect the appearances and behaviors of the animals [24]. However, many of the existing datasets are confined to specific or uniform environmental set-ups. Therefore, it is essentially important to possess a sizable range of environmental conditions and backgrounds. **(4)** Limited number of tasks and annotations for a more comprehensive analysis of animal behaviors. Most of the current datasets provide annotations of one task only, while different tasks can facilitate the understanding of animal behaviors in different aspects.

Therefore, we create a diverse dataset, Animal Kingdom, for animal behavioral analysis from video grounding task that identifies and extracts the segments of relevant videos, to animal action recognition and pose estimation tasks to better understand animal behaviors. The natural behaviors of animals in the wild are generally dynamic, complex and noisy [42,52,69], and our diverse dataset is a good reflection of the realities in the wild. Besides constructing the Animal Kingdom dataset, in this paper, we also design a simple yet effective Collaborative Action Recognition (CARe) model for action recognition with unseen new types of animals.

2. Related Work

2.1. Animal Behavioral Understanding Datasets

Analyzing poses and actions of animals in the wild allows researchers to objectively quantify natural behaviors [1, 24, 42, 52, 57, 72] and uncover new behaviors [1, 72]. In the computer vision community, it has inspired lots of works for animal behavioral analysis [5, 8, 17, 26, 34, 35, 38, 39, 46, 53, 54, 73, 77, 79]. Despite the great demand for large unifying animal behavior datasets [1, 4, 72], most of the current datasets are still relatively small, disparate, animal and environment-specific. Moreover, these datasets are typically either mainly image-based or dedicated to animal classification tasks only (*e.g.*, iNaturalist [71], Animals with Attributes [76], Caltech-UCSD Birds [75], Florida Wildlife Animal Trap [20]), or focused on one or few species only [14, 30, 40, 48, 75], or taken in specific environments (*e.g.*, Snapshot Serengeti [65], Fish4Knowledge [28]). We summarize some of the notable animal behavioral analysis datasets in Table 1.

Many existing animal behavior datasets are tailored to specific environments and target only a few actions performed by specific animals. For example, the behaviors of mice [23, 70], worms [27], monkeys [2] and flies [24, 55] are often examined in the laboratory settings, sometimes to understand the effects of experimental interventions [11]. Behaviors of animals, such as sheep [50], cows [36], and pigs [38] in livestock industry, and salmons [39] in fishing and aquaculture industries, are most frequently studied.

Meanwhile, the understanding of wildlife behaviors for ecological understanding and environmental protection [17] has garnered lots of interest of the community [8, 17, 35, 59, 79]. However, most of them are still largely limited to a few or specific classes of animals. Animal Pose [8] is a smaller subset created from the publicly available VOC2011, which is targeted at only 5 mammal species and used solely for pose estimation. In contrast, by far one of the largest datasets available, AP-10K [79] focuses solely on pose estimation, and it is confined to only mammals. Likewise, the action recognition dataset in [35] focuses solely on 7 action categories. Besides these, there are also some attempts to adopt synthetically generated animal data [46, 60] for pose estimation on real-life images of animals.

In summary, most of these datasets focus specifically on a small number of action classes and types of animals, yet there exists far more animals out there in the animal kingdom. All of these also illustrate the evident challenges of obtaining considerably diverse and comprehensive set of animal data. In comparison to all the existing animal behavior datasets, our Animal Kingdom dataset demonstrates noteworthy advantages in terms of having a bigger dataset size for a number of tasks (*i.e.*, video grounding, action recognition and pose estimation), with rich an-

notations of multi-label actions and a much vaster diversity of animal classes found in a wide range of diverse background scenes with different weather conditions, low light and night scenes, and taken at different viewpoints. Hence, our dataset provides a challenging and comprehensive benchmark for the community to develop and test different types of models for animal behavior analysis.

2.2. Animal Behavioral Analysis Methods

Some animal pose estimation models have been proposed [24, 29, 41, 52]. Various body parts of the animals are segmented and used for pose normalization in [67]. Some works use synthetic images to learn poses of real animals [31, 34, 60]. 3D models [46, 84] are even used in this endeavour. Another approach is to leverage knowledge from synthetic to real animals [31], or from both human and animal poses [8] to extend pose knowledge to other animals.

Current animal action recognition models are mainly adapted from existing human action recognition methods. Some of the popular ones are CNN [6] with LSTM [38, 50], Mask R-CNN [17, 59] with VGG [17], I3D [35], (2+1)D ResNet [59], and SlowFast [59]. Different from existing animal action recognition methods, in this work, considering the diversity of animals and the fact that the same class of action (e.g., *eating*) can look similar yet different when performed by different types of animals, we develop a simple yet effective Collaborative Action Recognition (CARe) model that extracts and utilizes both general and specific features to effectively identify the actions of unseen new types of animals.

3. The Proposed Animal Kingdom Dataset

3.1. Dataset Description

Our Animal Kingdom dataset contains 50 hours of long videos for video grounding, 50 hours of video clips for action recognition, and 33K annotated frames for pose estimation. In addition, our dataset possesses significant properties as listed below.

Diverse types of animals and actions. Having a large diversity of animals and actions is useful to understand animal behaviors in the wild, and important for training various types of models. Our dataset identifies 140 classes of actions exhibited by a wide range of terrestrial and aquatic animals, which make up to 850 unique species that span over 6 key animal classes (e.g., reptiles, birds, and fishes as seen in Fig. 1). Even for the same animal, the behaviors differ in different scenarios (e.g., amphibious animals like frogs which live and move differently in water and on land). Such is the intriguing lives of animals which add dimensions of complexities to action recognition in our dataset, and also influence the postures and movements of the animals, thus giving rise to a variety of animal poses. On the other hand, some animals having the same outward appearance live in
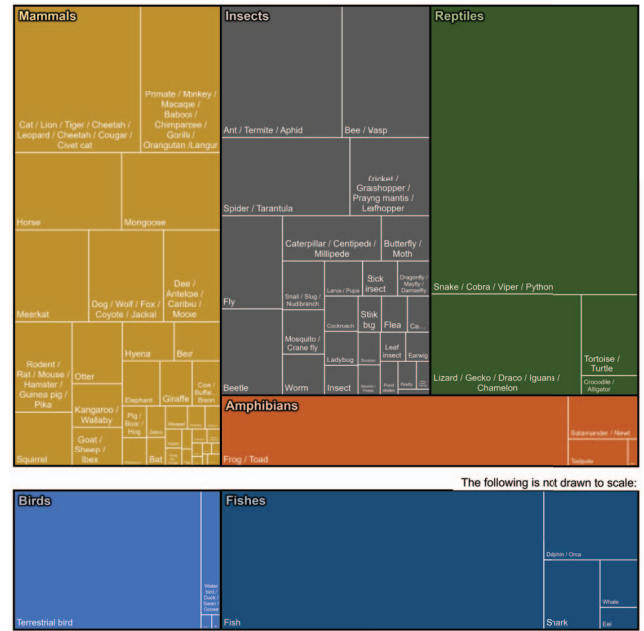

Figure 1. Distribution of clips of over 850 species of animals in 6 major animal classes, classified based on their appearance, number of limbs and how they move, and further divided into sub-classes.
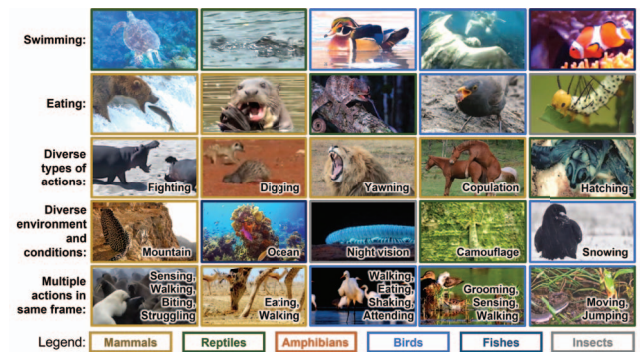

Figure 2. Examples of actions across various animal classes in our dataset. Rows 1 and 2 show how the same set of actions differs across various animal classes. Row 3 shows examples of various actions in our dataset. Row 4 shows sample features of our dataset, ranging from diverse environments to varying illumination and weather conditions. Row 5 shows multiple actions performed by multiple animals in the same frame.

totally different habitats and thus have different forms of movements (e.g., a terrestrial spider that crawls on land but a water spider that swims in water). All of these examples in our dataset illustrate the complex yet beautiful diversity of wildlife, and highlight the benefit of containing a diverse and good range of animal classes to encompass various actions and behaviors that occur naturally in the wild.

Diverse environmental conditions. The various environmental and weather conditions enable better understanding of animal behaviors in natural ecosystems. In the wild, animals are found in a variety of environments (e.g., grassland, forests, rivers, oceans) and environmental conditions

19003

Table 1. Comparison of our dataset with some of the previous animal behavioral analysis datasets. Our dataset contains more annotated data for various tasks, more diverse animal classes, multiple types of environmental and weather conditions, which shall provide a challenging and comprehensive benchmark for animal behavior understanding in images and videos.

Dataset	Publicly available?	Diverse types of animals						No. of species	Task 1: Video Grounding		Task 2: Action Recognition		Task 3: Pose Estimation	Types of scene									Weather			
		Mammals	Reptiles	Birds	Amphibians	Fishes	Insects		No. of annotated long videos	No. of statements	No. of annotated video clips	No. of annotated action classes	No. of labelled images	Night scene	Low light	Complex background	Mountain	Forest	Grassland	Desert	Ocean	Underwater	Windy	Foggy	Rain	Snow
Broiler Chicken [14]	×	×	×	✓	×	×	×	NA	×	×	NA	6	556	NA	NA	NA	×	×	×	×	×	×	NA	NA	NA	NA
Fish Action [53]	×	×	×	×	×	✓	×	NA	×	×	95	5	×	×	✓	✓	×	×	×	×	✓	✓	×	×	×	×
Salmon Feeding [39]	×	×	×	×	×	✓	×	1	×	×	76	2	×	×	✓	✓	×	×	×	×	✓	✓	×	×	✓	×
Wild Felines [17]	×	✓	×	×	×	×	×	3	×	×	2,700	3	×	✓	✓	✓	×	✓	✓	×	×	×	NA	NA	NA	NA
Pig Tail-biting [38]	×	✓	×	×	×	×	×	1	×	×	4,396	2	×	×	✓	×	×	×	×	×	×	×	×	×	×	×
Wildlife Action [35]	×	✓	✓	✓	✓	✓	✓	106	×	×	10,600	7	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	NA	NA	NA	NA
Animal Pose [8]	✓	✓	×	✓	×	×	×	5	×	×	×	×	4,666	✓	✓	✓	✓	✓	✓	×	×	×	×	✓	×	✓
Horse-30 [40]	✓	✓	×	×	×	×	×	3	×	×	×	×	8,144	×	×	✓	×	×	✓	×	×	×	×	×	×	×
AP-10K [79]	✓	✓	×	×	×	×	×	54	×	×	×	×	10,015	✓	✓	✓	✓	✓	✓	✓	✓	×	×	✓	×	✓
Macaque Pose [30]	✓	✓	×	×	×	×	×	NA	×	×	×	×	13,083	✓	✓	✓	×	×	×	×	×	×	×	✓	×	×
Dogs [3]	✓	✓	×	×	×	×	×	1	×	×	13	4	2,200	×	✓	×	×	×	×	×	×	×	×	×	×	×
Animal Kingdom (Ours)	✓	✓	✓	✓	✓	✓	✓	850	4,301 (50h)	18,744	30,100 (50h)	140	33,099	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

(*e.g.*, raining, snowing). The animal footages in our dataset are captured at different times of day, under different illumination conditions (*e.g.* low light, sunset, backlight, different underwater depths), sometimes leading to shadow cast and low foreground-background contrast. This challenge is compounded when some animals (*e.g.*, crows or ravens in Fig. 2) lack distinctive colours for various body parts, while others (*e.g.*, tree lizards in Fig. 2) camouflage very well, blending seamlessly into their environment [56]. Some animals (*e.g.*, chameleons in Fig. 2) may even change their appearance rapidly in response to the environment. Animals move around in habitats with complex compositions and dynamic background with other animals in close proximity, or plants swaying in the wind [56]. All of these are found in our dataset and present the significant yet practical challenges of animal behavioral analysis in the wild.

Varied viewpoints and various types of footages. Obtaining real footages documenting how animals live and behave in the wild is of great significance to ecological studies and conservation efforts [62]. The animal footages, which were captured by nature enthusiasts and professionals who used different equipment (*e.g.*, night vision), come in various forms (*e.g.*, documentaries, animal trap videos, and personal camera recordings), and were taken from various vantage points (*e.g.*, birds' eye view, bottom up, underwater). This brings diversities of captured viewpoints in our dataset.

Various compositions of scenes with fine-grained multi-label actions. Our dataset encapsulates the lives and actions of wildlife, with clips depicting the fine-grained action(s) of one animal to multiple types of animals in a single frame. This leads to the creation of a multi-label dataset comprising fine-grained actions to aptly describe each unique action of animal(s) in the clip. However, this also inevitably leads to a long-tailed distribution of actions, which poses a significant challenge for accurate models, and deserves attention to deepen research into the building

of more robust models for the practical world.

3.2. Dataset Tasks and Annotations

Similar to the experience detailed in [72], our diverse dataset for video grounding, action recognition and pose estimation was meticulously assembled together through the collective effort of 23 individuals (including biology experts). We manually identified and provided framewise annotations of both animal and action descriptions for 50 hours of videos that were collected from YouTube videos. A total of 3 rounds of quality checks were performed to uphold the annotation quality. Please refer to the Supplementary and our website for more details of our dataset.

Action recognition with multi-label fine-grained actions. In action recognition, the model takes in an input video clip, and outputs the action labels for it. Actions of animals can occur within split seconds (*e.g.*, jumping), and up to minutes (*e.g.*, sexual display) for more complex behavioral patterns. Our diverse dataset contains 50 hours of video clips, comprising over 850 animal species and 140 fine-grained action classes that were taken from a list used by ethologists [19, 22, 58, 63]. The collection of actions and behaviors (see Supplementary) encompasses life events (*e.g.*, molting), to daily activities (*e.g.*, feeding), to social interactions (*e.g.*, playing). The video clips last an average of 6 seconds, and range from 1 second to 117 seconds.

The complexities brought by the natural world, present a set of practical challenges in analysing animal actions in the wild. **(1)** There exists a large intra and inter-class variation of actions in our dataset. As seen in the first two rows in Fig. 2, the same type of action (*e.g.*, eating) can look differently when performed by different types of animals. Hence, we need to design robust methods to recognize the actions of various animals. **(2)** In nature, an animal can perform more than one action at the same time. Besides, there can be multiple animals in the same frame, with different animals

performing different actions. All of these lead to the multi-label actions, and can also result in a long-tailed distribution of actions. **(3)** As the videos of animals were captured in the wild, some actions (*e.g.*, molting) occur less frequently than others (*e.g.*, eating) in nature. This inevitably leads to an uneven long-tailed distribution of actions, thus encouraging the development of advanced strategies for handling long-tailed distributions.

When constructing our animal action recognition dataset, we follow the work of [82] and divide the distribution into three different segments based on the number of samples in each action class. Specifically, we group all the 140 action classes in our dataset into the head segment (17 action classes that have more than 500 samples each), the middle segment (29 action classes that have 100 to 500 samples each), and the tail segment (94 action classes that have fewer than 100 samples each). The head segment contains the more frequently occurring actions in nature (*e.g.*, sensing and eating). The middle segment consists of actions that are commonly observed in nature (*e.g.*, climbing, grooming, and digging), while the tail segment is made up of actions that are less frequently observed in nature (*e.g.*, molting and doing somersault).

Video grounding based on descriptions. In video grounding, an input query sentence, describing the scene with the animals and behaviors of interest, is provided to the model. Then the model needs to output the relevant temporal segment with the start and end timing (Fig. 3), much like a video clip search engine. Video grounding plays a pivotal role in improving productivity for ecological research, because a large part of the ecological surveillance videos may not contain the footage of the animals and behaviors of interest [12,44], while video grounding provides a convenient way for users to search for the relevant time segment in long videos by describing the scenarios of interest.

The key challenge of video grounding comes in the form of identifying the behaviors and delineating the animals from the complex background to return relevant results based on the language descriptions. The video grounding task in our dataset contains 50 hours of videos (a total of 4301 long video sequences) with 18,744 annotated sentences. Each video sequence contains 3-5 sentences.

Figure 3. Samples of the video grounding task. Given the language description, we need to detect the corresponding time segment.

Pose estimation for inferring animal keypoints. In pose estimation, the model takes in an input image of the animal, and predicts the locations of its joints. As animals can have vastly different anatomical structures across classes, this poses a challenge for pose estimation. As such, we organize the images according to five main animal classes: mammals, reptiles, amphibians, birds, and fishes, to construct differentiated understanding of the poses for each animal class. We institute a common set of keypoints across the five animal classes (Fig. 4). These keypoints correspond to the equivalent of human poses, and are also in line with the recommendations in [72]. We define a total of 23 keypoints: 1 head, 2 eyes, 4 parts of mouth, 2 shoulders, 2 elbows, 2 wrists, 1 mid torso, 2 hips, 2 knees, 2 ankles, and 3 parts of the tail. For the upper limbs of birds (*i.e.*, wings) and fishes (*i.e.*, fins), their shoulders, elbows, and wrists are defined in accordance to how their upper limbs move. Similarly, for the 'lower' limbs of fishes which do not have legs, their hips, knees, and ankles are annotated along the edge of their tail section, because their tails control their swimming movements [7].

In summary, our Animal Kingdom dataset possesses several significant properties for various types of animal behavior analysis tasks. In addition, the large intra-class variations, multi-label actions, and long-tailed distributions, present practical challenges to understanding animal behaviors in the wild. Hence, our dataset provides a challenging and diverse benchmark to facilitate the research community to develop, adapt, and evaluate various types of advanced methods for animal behavioral analysis.

4. The Proposed CARe Model

When constructing our dataset, we observe that the same class of action can look similar yet different when performed by different types of animals. For example, the action "eating" shares similarities yet still differs between chameleons and otters, as shown in Fig. 2. Meanwhile, training network models to analyze and recognize actions of various types of animals can be an interesting problem. However, even if we can collect a large number of video action samples for a set of animals, it can still be difficult to exhaustively collect and annotate lots of samples for various types of species for network training, considering that there are more than 7 million species in the world [45].
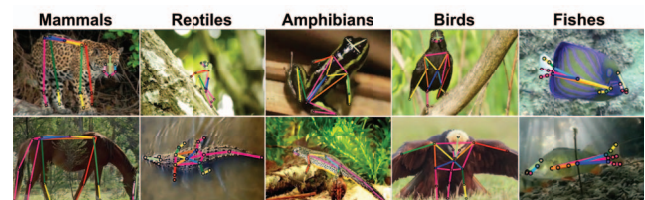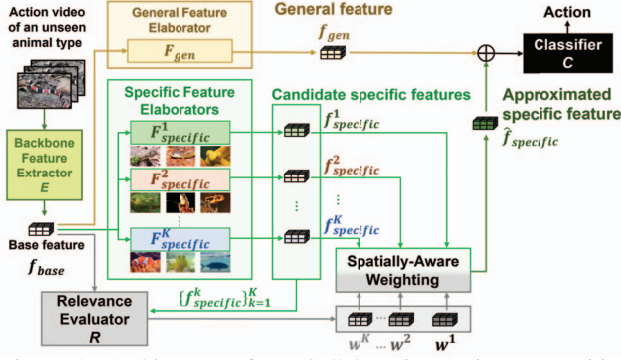

Figure 4. Examples of animal poses in our dataset

Figure 5. Architecture of our Collaborative Action Recognition (CARe) model. During testing, given an video action sample of an unseen type of animal, the approximated specific feature will be obtained by combining the outputs of the K existing specific feature elaborators in a spatially-aware weighting manner. Then the approximated specific feature will be fused with the general feature, and finally fed to the classifier C to predict the action class.

Nevertheless, since the same action class of different animals can share some common (generalizable) characteristics, it becomes possible to train a general network based on video samples of the annotated animal set, and then apply this network for recognizing the same set of actions of even unseen new animals (*i.e.*, animal types that are not included during model training, but are of interest later). In this way, well generalizable features shared across different animals can be extracted by the network for action recognition. However, extracting general features only while ignoring the specific features of actions of different animal types can limit the action recognition performance, because the specific features, which though may be unique to a certain animal type only, can still convey discriminative information for recognizing the actions of this animal type.

To address this issue, we propose a simple yet effective Collaborative Action Recognition model (CARe), which is able to recognize the same set of actions even for unseen types of animals, by leveraging both the general and specific features, as shown in Fig. 5. Assuming that we have the data of K types of observed animals for training, then the training data can be denoted as $\mathbf{D} = \{D_k\}_{k=1}^{K}$, where $D_k = \{(x_n^k, y_n^k)\}_{n=1}^{N_k}$. Here x_n^k represents a video action sample, y_n^k represents its ground-truth action label, and N_k is the number of samples for the k-th observed animal type.

Our basic model is composed of the backbone feature extractor E, as well as the relatively lightweight general feature elaborator F_{gen} (shared by all animal types) and specific feature elaborators $\{F_{specific}^k\}_{k=1}^{K}$ (each specific for a certain type of animal), and the classifier C. The backbone feature extractor E (an I3D model) is used to extract base feature $f_{base} = E(x)$ from each input video sample x, which is then fed to the general feature elaborator F_{gen} to get the general feature $f_{gen} = F_{gen}(f_{base})$, and a specific feature elaborator $F_{specific}^k$ to get the specific feature

$f_{specific}^k = F_{specific}^k(f_{base})$. Such a base model can be trained on the training set $\mathbf{D}$ as follows:

$$\hat{y}_n^k = C\Big(F_{gen}\big(E(x_n^k)\big), F_{specific}^k\big(E(x_n^k)\big)\Big)$$
$$\ell = \ell_{\text{CE}}(\hat{y}_n^k, y_n^k) \tag{1}$$

This indicates, for each training sample x_n^k, its general and specific features are concatenated and finally fed to the classifier C for predicting the action label $\hat{y}_n^k$, and the training of this base model is performed under the supervision of the cross-entropy loss ℓ_{CE}. Via this scheme, during training, action samples from all the K types of training animals will pass through the general feature elaborator (F_{gen}), while only the samples from the k-th animal type will pass through the specific elaborator $F_{specific}^k$. Thus, the general feature elaborator (that is well trained on a large amount of data of various (*i.e.*, K) types of animals) is able to learn to extract well generalizable features. Meanwhile, each specific elaborator $F_{specific}^k$ (that is trained on the data of a specific type of animal) will learn to extract features that tend to be more specific for the corresponding type of animal. Thus, these two sets of features, which are complementary, can be fused for action recognition.

Though the aforementioned basic model can generate general and specific features for a seen animal type, it still does not address the scenario for handling unseen animal types, since we do not have specific feature elaborators for animal types beyond these K trained types.

Considering that different animals can share different levels of similarities in their appearances and behaviors (*e.g.*, cheetahs share many more similarities with tigers, compared to other animals like snakes), when handling an input video of an unseen animal type, to obtain its specific feature, a possible solution is to leverage existing elaborators based on different levels of similarities, to mine and approximate the specific feature for the unseen animal.

To this end, we design a relevance evaluator R that is a sub-network for assessing the similarities of the action sample (x) of the unseen animal type w.r.t all the K observed animal types. To achieve this, we first input x to the backbone feature extractor E to get the base feature (f_{base}), which is then fed to all the K specific elaborators, yielding a set of intermediate features $\{f_{specific}^k\}_{k=1}^{K}$. Then we estimate the similarities as: $\mathbf{w} = R\big(f_{base}, \{f_{specific}^k\}_{k=1}^{K}\big)$, where $\mathbf{w} = \{w^k\}_{k=1}^{K}$ represents the set of similarity (relevance) scores of this sample w.r.t the K observed animal types. Thus, we can then approximate the specific feature of this unseen animal type's action sample (x), via the collaboration of the existing K specific elaborators as:

$$\hat{f}_{specific} = \frac{1}{K} \sum_{k=1}^{K} w^k \cdot f_{specific}^k \tag{2}$$

Moreover, considering an animal can have different similarity degrees w.r.t another type of animal at different spa-

tial positions (different body parts), w^k can also be implemented as a tensor (*i.e.*, $w^k \in \mathbb{R}^{h \times w}$) instead of a scalar, such that the approximated specific feature of the unseen animal type is produced based on its different similarities to different observed animal types and also different spatial positions on the feature maps (with size $h \times w$). In this way, the K feature elaborators collaborate with one another in a spatially-aware manner to yield optimal approximation. After specific features have been approximated, the general and approximated specific features are fused and fed to the classifier C to predict the action class, as shown in Fig. 5.

As mentioned before, our basic model can be trained on the training set (**D**) of the observed animals. However, the relevance evaluator R needs to be generalizable to unseen new animals, which makes the training of R challenging. Inspired by meta learning [13, 18, 25, 32], which is a "learning to learn" technique that can learn to generalize by splitting the original training set to meta-train and meta-test sets to mimic target testing scenarios during model training, we here apply meta-learning to train our R for improving its generalization ability to unseen animal types. Specifically, following the mimicking scheme of meta learning, we split the training set containing the K types of animals into a virtual meta-train set (with $K-1$ seen types of animals), and a virtual meta-test set (with one "unseen" type of animal). Based on these sets, we then take advantage of the domain generalization meta-learning (MLDG) algorithm [32] that can effectively optimize our module R.

By alternately training our base model using Eq. (1) and training R via MLDG, our CARe model obtains generalization ability, and can produce reliable general and approximated specific features that are useful for recognition of the same set of actions of unseen animals. More implementation details of CARe can be found in the Supplementary.

5. Experiments

We evaluate state-of-the-art methods on our dataset: I3D [10], X3D [15], and SlowFast [16] for action recognition; VSLNet [81] and LGI [47] for video grounding; and HRNet [64] and HRNet-DARK [80] for pose estimation. We use their original codes and adapt them for our Animal Kingdom dataset.

5.1. Action Recognition Results

For our first type of experiments, we perform multi-label action recognition. We first use Multi-Label Binarizer [51] on our multi-label animal action recognition dataset, which contains all the 30.1K video clips, to stratify the action classes and split both the video clips and each action class into 80% train and 20% test. Following [61], we adopt the mean Average Precision (mAP) as our evaluation metric. As mentioned in Sec. 3.2, our dataset contains a long-tailed distribution of multi-label actions and we follow [82]

Table 2. Results of action recognition

Method	mAP			
	overall	head	middle	tail
Baseline (Cross Entropy Loss)				
I3D [10]	16.48	46.39	20.68	12.28
SlowFast [16]	20.46	54.52	27.68	15.07
X3D [15]	25.25	60.33	36.19	18.83
Focal Loss [37]				
I3D [10]	26.49	64.72	40.18	19.07
SlowFast [16]	24.74	60.72	34.59	18.51
X3D [15]	28.85	64.44	39.72	22.41
LDAM-DRW [9]				
I3D [10]	22.40	53.26	27.73	17.82
SlowFast [16]	22.65	50.02	29.23	17.61
X3D [15]	30.54	62.46	39.48	24.96
EQL [66]				
I3D [10]	24.85	60.63	35.36	18.47
SlowFast [16]	24.41	59.70	34.99	18.07
X3D [15]	30.55	63.33	38.62	25.09

to identify the head, middle and tail segments. Thus, we evaluate the overall mAP and also segment-specific mAP of action recognition. Note that as our dataset has long-tailed distribution, we also evaluate methods for handling long-tailed distribution [9, 37, 66] on our dataset, and apply these methods on action recognition methods. Specifically, we test the cost-sensitive learning (Focal Loss [37]) and re-weighting methods (LDAM-DRW [9] and EQL [66]), in conjunction with different action recognition methods (I3D [10], X3D [15], and SlowFast [16]). As shown in Table 2, action recognition methods with approaches for handling long-tailed distribution (*e.g.*, Focal Loss) achieve higher mAP for the head, middle and tail segments, compared to the baselines. This shows the possibility and brings opportunities for the community to explore various approaches to handle challenges including long-tailed and multi-label problems.

For our second type of experiment, we perform action recognition for unseen types of animals. Specifically, we pick video clips of six action classes (*i.e.*, moving, eating, attending, swimming, sensing, and keeping still) that are widely shared by various types of animals, to train and test our CARe model for action recognition of unseen new types of animals. We select four animal types (*i.e.*, birds, fishes, frogs, snakes) for training, and another five animal types (*i.e.*, lizards, primates, spiders, orthopteran insects, water fowls) for testing. In this experiment, we use the video clips, each containing only one action instance. As such, we use accuracy as the evaluation metric for this experiment. We conduct ablation studies, as well as compare our method with two recent domain generalization methods [33, 74], and all experiments are conducted by using I3D as the backbone. For the ablation study, we first construct 'CARe without specific feature', and 'CARe without general feature' (*i.e.*, only using approximated specific feature). We also construct 'CARe without spatially-aware weighting' where similarity weights **w** are implemented as

19007

Table 3. Results of action recognition of unseen animals

Method	Accuracy (%)
Episodic-DG [33]	34.0
Mixup [74]	36.2
CARe without specific feature	27.3
CARe without general feature	38.2
CARe without spatially-aware weighting	37.1
CARe (Our full model)	39.7

a set of scalars, while **w** are a set of tensors in our full CARe model. Results in Table 3 show that our full model (CARe) achieves the best results, demonstrating the effectiveness of taking advantage of the general feature and also spatially-aware approximated specific features for recognizing actions of unseen types of animals.

5.2. Video Grounding Results

Following [21, 47, 81], we use both mean Intersection over Union (IoU), and "Recall@n, IoU=μ" as our evaluation metrics. Mean IoU means the average IoU of all test samples. As for "Recall@n, IoU=μ", we first compute the IoU of the results with ground truth (*i.e.*, length between the start and end time). The returned result is considered correct if IoU is larger than the threshold μ. Hence, "Recall@n, IoU=μ" refers to the percentage of language queries in the test set with at least one correct result among the top-n returned results. We use n=1 and $\mu \in \{0.1, 0.3, 0.5, 0.7\}$, and report the results in Table 4.

Table 4. Results of video grounding

Method	Recall@1				mean IoU
	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.7	
LGI [47]	50.84	33.51	19.74	8.94	22.90
VSLNet [81]	53.59	33.74	20.83	12.22	25.02

5.3. Pose Estimation Results

We define three protocols to identify potential challenges of pose estimation that arise due to intra and inter-animal class variations.

Protocol 1: The whole dataset with all animal species is split to the training set (80% of samples) and test set (20% of samples). This means all species are contained in both the train and test sets, for evaluating the efficacy of models when estimating pose of all animals.

Protocol 2: We adopt the leave-k-out setting by selecting $k = 12$ animal sub-classes (*e.g.*, felines, turtles, etc.) that are reserved and used only for testing, while the other animal sub-classes are used for training. The selected sub-classes come from the five major classes. Therefore, the sub-classes in the test set do not appear in the training set. This evaluates the generalization ability of the model for pose estimation of unseen animal classes.

Protocol 3: We group all the samples in our dataset to five major classes (*i.e.*, mammals, amphibians, reptiles, birds, and fishes). And for each major class, we define its

Table 5. Results of pose estimation

Protocol	Description	PCK@0.05	
		HRNet [64]	HRNet-DARK [80]
Protocol 1	All	66.06	66.57
Protocol 2	Leave-k-out	39.30	40.28
Protocol 3	Mammals	61.59	62.50
	Amphibians	56.74	57.85
	Reptiles	56.06	57.06
	Birds	77.35	77.41
	Fishes	68.25	69.96

own training set (80% samples) and test set (20% samples). This is to evaluate the pose estimation performance of each major class of animals.

Following [78], we adopt Percentage of Correct Keypoints (PCK) as our evaluation metric. PCK@α measures the accuracy of localizing body joints within the distance threshold α. The predicted keypoint is considered to be correct if it falls within the distance threshold calculated by $\alpha \times max(height, width)$ of the animal bounding box. Here we set α to 0.05. As shown in Table 5, we evaluate two state-of-the-art pose estimation methods [64, 80] on our dataset. In Protocol 3, the results for reptiles and amphibians are lower than other animals, possibly due to the challenges of estimating their keypoints, as they come in various forms of shapes and textures with varying positions of their joints within its class, and they camouflage well into their environments. We observe that the result of Protocol 2 is comparatively lower than the rest, illustrating the challenges in generalizing the pose to unseen new animal classes. This shall bring opportunities for the community to develop novel methods to better learn and generalize pose information across various animal classes.

6. Conclusion

We introduce a challenging animal behavioral analysis dataset, comprising over 850 species of animals in the animal kingdom for video grounding, action recognition, and pose estimation. We also introduce a Collaborative Action Recognition (CARe) model with improved ability to recognize actions of unseen new animal types. The variations within and between classes of animals for both action recognition and pose estimation illustrate the advantage of our dataset being diverse in nature, and yet at the same time highlight the challenges and reinforce the compelling need for further research in the areas of video grounding, animal action recognition and pose estimation. We expect that our work will inspire further works in animal behavioral analysis, and understanding animal behaviors can make our world a better place for both wildlife and humans.

References

[1] David J. Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, 2014. 1, 2

[2] Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. *Nature Communications*, 11(1):1–12, 2020. 2

[3] Shanis Barnard, Simone Calderara, Simone Pistocchi, Rita Cucchiara, Michele Podaliri-Vulpiani, Stefano Messori, and Nicola Ferri. Quick, accurate, smart: 3D computer vision technology helps assessing confined animals' behaviour. *PLOS One*, 11(7):e0158748, 2016. 4

[4] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iWildCam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 2

[5] Cigdem Beyan and Robert B Fisher. Detection of abnormal fish trajectories using a clustering based hierarchical classifier. In *British Machine Vision Conference (BMVC)*, 2013. 2

[6] James P Bohnslav, Nivanthika K Wimalasena, Kelsey J Clausing, Yu Y Dai, David A Yarmolinsky, Tomás Cruz, Adam D Kashlan, M Eugenia Chiappe, Lauren L Orefice, Clifford J Woolf, et al. DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. *ELife*, 10:e63377, 2021. 3

[7] Iman Borazjani and Mohsen Daghooghi. The fish tail motion forms an attached leading edge vortex. *Proceedings of the Royal Society B: Biological Sciences*, 280(1756):20122071, 2013. 5

[8] 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 (ICCV)*, pages 9498–9507, 2019. 1, 2, 3, 4

[9] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7

[10] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 7

[11] Carlos Fernando Crispim-Junior, Fernando Mendes de Azevedo, and José Marino-Neto. What is my rat doing? Behavior understanding of laboratory animals. *Pattern Recognition Letters*, 94:134–143, 2017. 2

[12] Fagner Cunha, Eulanda M. dos Santos, Raimundo Barreto, and Juan G. Colonna. Filtering empty camera trap images in embedded systems. In *CVPR Workshop*, pages 2438–2446, 2021. 1, 5

[13] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, pages 16145–16154, 2021. 7

[14] Cheng Fang, Tiemin Zhang, Haikun Zheng, Junduan Huang, and Kaixuan Cuan. Pose estimation and behavior classification of broiler chickens based on deep neural networks. *Computers and Electronics in Agriculture*, 180:105863, 2021. 2, 4

[15] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213, 2020. 7

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *CVPR*. 7

[17] Liqi Feng, Yaqin Zhao, Yichao Sun, Wenxuan Zhao, and Jiaxi Tang. Action recognition using a spatial-temporal network for wild felines. *Animals*, 11(2), 2021. 1, 2, 3, 4

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017. 7

[19] National Centre for the Replacement Refinement and Reduction of Animals in Research. General ethograms. Accessed November 11, 2021 [Online]. Available: https://nc3rs.org.uk/sites/default/files/documents/EvaluatingEnvironmentalEnrichment/General%20ethograms.pdf. 4

[20] Crystal Gagne, Jyoti Kini, Daniel Smith, and Mubarak Shah. Florida wildlife camera trap dataset. *arXiv preprint arXiv:2106.12628*, 2021. 2

[21] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *CVPR*, pages 5267–5275, 2017. 8

[22] Joseph Garner. Ethogram comparison. Accessed November 11, 2021 [Online]. Available: https://mousebehavior.org/ethogram-comparison/. 4

[23] Brian Q Geuther, Asaf Peer, Hao He, Gautam Sabnis, Vivek M Philip, and Vivek Kumar. Action detection using a neural network elucidates the genetics of mouse grooming behavior. *ELife*, 10:e63207, 2021. 2

[24] Jacob M Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R Costelloe, and Iain D Couzin. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *ELife*, 8:e47994, 2019. 1, 2, 3

[25] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *CVPR*, pages 6163–6172, 2020. 7

[26] Shilpi Gupta, Prerana Mukherjee, Santanu Chaudhury, Brejesh Lall, and Hemant Sanisetty. DFTNet: Deep fish tracker with attention mechanism in unconstrained marine environments. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021. 2

[27] 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. 2

[28] Daniela Giordano Lynda Hardman Fang-Pang Lin Jessica Yun-Heh Chen-Burger, Robert Fisher. Fish4Knowledge, 2020. 2

[29] Pierre Karashchuk, Katie L Rupp, Evyn S Dickinson, Sarah Walling-Bell, Elischa Sanders, Eiman Azim, Bingni W Brunton, and John C Tuthill. Anipose: a toolkit for robust markerless 3d pose estimation. *Cell Reports*, 36(13):109730, 2021. 1, 3

[30] Rollyn Labuguen, Jumpei Matsumoto, Salvador Blanco Negrete, Hiroshi Nishimaru, Hisao Nishijo, Masahiko Takada, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. MacaquePose: A novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in Behavioral Neuroscience*, page 268, 2021. 2, 4

[31] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *CVPR*, pages 1482–1491, 2021. 3

[32] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 7

[33] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019. 7, 8

[34] Siyuan Li, Semih Gunel, Mirela Ostrek, Pavan Ramdya, Pascal Fua, and Helge Rhodin. Deformation-aware unpaired image translation for pose estimation on laboratory animals. In *CVPR*, pages 13158–13168, 2020. 1, 2, 3

[35] Weining Li, Sirnam Swetha, and Mubarak Shah. Wildlife action recognition using deep learning, 2020. Accessed November 11, 2021 [Online]. Available: https://www.crcv.ucf.edu/wp-content/uploads/2018/11/Weining_L_Report.pdf. 1, 2, 3, 4

[36] 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

[37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 7

[38] 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. 1, 2, 3, 4

[39] 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. 1, 2, 4

[40] 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 (WACV)*, pages 1859–1868, 2021. 2, 4

[41] 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. *Nature Neuroscience*, 21(9):1281–1289, 2018. 3

[42] Mackenzie Weygandt Mathis and Alexander Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*, 60:1–11, 2020. Neurobiology of Behavior. 1, 2

[43] Donal P McCarthy, Paul F Donald, Jörn PW Scharlemann, Graeme M Buchanan, Andrew Balmford, Jonathan MH Green, Leon A Bennun, Neil D Burgess, Lincoln DC Fishpool, Stephen T Garnett, et al. Financial costs of meeting global biodiversity conservation targets: current spending and unmet needs. *Science*, 338(6109):946–949, 2012. 1

[44] Zhongqi Miao, Ziwei Liu, Kaitlyn M. Gaynor, Meredith S. Palmer, Stella X. Yu, and Wayne M. Getz. Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3(10):885–895, Oct 2021. 1, 5

[45] Camilo Mora, Derek P Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. How many species are there on earth and in the ocean? *PLOS Biology*, 9(8):e1001127, 2011. 5

[46] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *CVPR*, pages 12386–12395, 2020. 1, 2, 3

[47] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*, 2020. 2, 7, 8

[48] Natalia Neverova, Artsiom Sanakoyeu, Patrick Labatut, David Novotny, and Andrea Vedaldi. Discovering relationships between object categories via universal canonical maps. In *CVPR*, pages 404–413, 2021. 2

[49] Hung Nguyen, Sarah J Maclagan, Tu Dinh Nguyen, Thin Nguyen, Paul Flemons, Kylie Andrews, Euan G Ritchie, and Dinh Phung. Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 40–49. IEEE, 2017. 1

[50] Kehinde Owoeye and Stephen Hailes. Online collective animal movement activity recognition. *arXiv preprint arXiv:1811.09067*, 2018. 2, 3

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 7

[52] Talmo D Pereira, Diego E Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S-H Wang, Mala Murthy, and Joshua W Shaevitz. Fast animal pose estimation using deep neural networks. *Nature Methods*, 16(1):117–125, 2019. 1, 2, 3

[53] Shah Atiqur Rahman, Insu Song, Maylor KH Leung, Ickjai Lee, and Kyungmi Lee. Fast action recognition using negative space features. *Expert Systems with Applications*, 41(2):574–587, 2014. 1, 2, 4

[54] Maheen Rashid, Xiuye Gu, and Yong Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *CVPR*, pages 6894–6903, 2017. 2

[55] Primoz Ravbar, Kristin Branson, and Julie H Simpson. An automatic behavior recognition system classifies animal behaviors using movements and their temporal context. *Journal of Neuroscience Methods*, 326:108352, 2019. 2

[56] Benjamin Risse, Michael Mangan, Luca Del Pero, and Barbara Webb. Visual tracking of small animals in cluttered natural environments using a freely moving camera. In *ICCV*, pages 2840–2849, 2017. 4

[57] Lianne Robinson and Gernot Riedel. Comparison of automated home-cage monitoring systems: Emphasis on feeding behaviour, activity and spatial learning following pharmacological interventions. *Journal of Neuroscience Methods*, 234:13–25, 2014. 2

[58] Paul E Rose and Lisa M Riley. Conducting behavioural research in the zoo: A guide to ten important methods, concepts and theories. *Journal of Zoological and Botanical Gardens*, 2(3):421–444, 2021. 4

[59] Frank Schindler and Volker Steinhage. Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecological Informatics*, 61:101215, 2021. 2, 3

[60] Moira Shooter, Charles Malleson, and Adrian Hilton. SyDog: A synthetic dog dataset for improved 2D pose estimation. In *CVPR Workshop*, 2021. 2, 3

[61] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, pages 510–526. Springer, 2016. 7

[62] Abhineet Singh, Marcin Pietrasik, Gabriell Natha, Nehla Ghouaiel, Ken Brizel, and Nilanjan Ray. Animal detection in man-made environments. In *WACV*, pages 1438–1449, 2020. 1, 4

[63] Richard Stafford, Anne E Goodenough, Kathy Slater, William S Carpenter, Laura Collins, Heather Cruickshank, Sarah Downing, Sally Hall, Katie McDonald, Heather McDonnell, et al. Inferential and visual analysis of ethogram data using multivariate techniques. *Animal Behaviour*, 83(2):563–569, 2011. 4

[64] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 7, 8

[65] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, 2(1):1–14, 2015. 2

[66] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, pages 11662–11671, 2020. 7

[67] Luming Tang, Davis Wertheimer, and Bharath Hariharan. Revisiting pose-normalization for fine-grained few-shot recognition. In *CVPR*, pages 14352–14361, 2020. 3

[68] Biosafety Unit. Global biodiversity outlook 5, 2020. Accessed November 11, 2021 [Online]. Available: https://www.cbd.int/gbo/gbo5/publication/gbo-5-en.pdf. 1

[69] Elsbeth A van Dam, Lucas PJJ Noldus, and Marcel AJ van Gerven. Deep learning improves automated rodent behavior recognition within a specific experimental setup. *Journal of Neuroscience Methods*, 332:108536, 2020. 2

[70] Elsbeth A. van Dam, Lucas P.J.J. Noldus, and Marcel A.J. van Gerven. Deep learning improves automated rodent behavior recognition within a specific experimental setup. *Journal of Neuroscience Methods*, 332:108536, 2020. 2

[71] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 2

[72] 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, 2, 4, 5

[73] Jung-Hua Wang, Shih-Kai Lee, Yi-Chung Lai, Cheng-Chun Lin, Ting-Yuan Wang, Ying-Ren Lin, Te-Hua Hsu, Chang-Wen Huang, and Chung-Ping Chiang. Anomalous behaviors detection for underwater fish using ai techniques. *IEEE Access*, 8:224372–224382, 2020. 2

[74] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020. 7, 8

[75] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 2

[76] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning – A comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*, 2020. 2

[77] Qiumei Yang, Deqin Xiao, and Sicong Lin. Feeding behavior recognition for group-housed pigs with the faster R-CNN. *Computers and Electronics in Agriculture*, 155:453–460, 2018. 1, 2

[78] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2878–2890, 2012. 8

[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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1, 2, 4

[80] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*. 7, 8

[81] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, 2020. 2, 7, 8

[82] Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry Davis. VideoLT: Large-scale long-tailed video recognition. 2021. 5, 7

[83] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *CVPR*, pages 3955–3963, 2018. 1

[84] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, pages 6365–6373, 2017. 3