MammalNet: A Large-scale Video Benchmark for Mammal Recognition and Behavior Understanding

Jun Chen[1*] Ming Hu[1*] Darren J. Coker[1] Michael L. Berumen[1]
Blair Costelloe[2,3] Sara Beery[4] Anna Rohrbach[5] Mohamed Elhoseiny[1]
[1]King Abdullah University of Science and Technology (KAUST)
[2] Max Planck Institute of Animal Behavior , [3] University of Konstanz
[4]Massachusetts Institute of Technology, [5]University of California, Berkeley

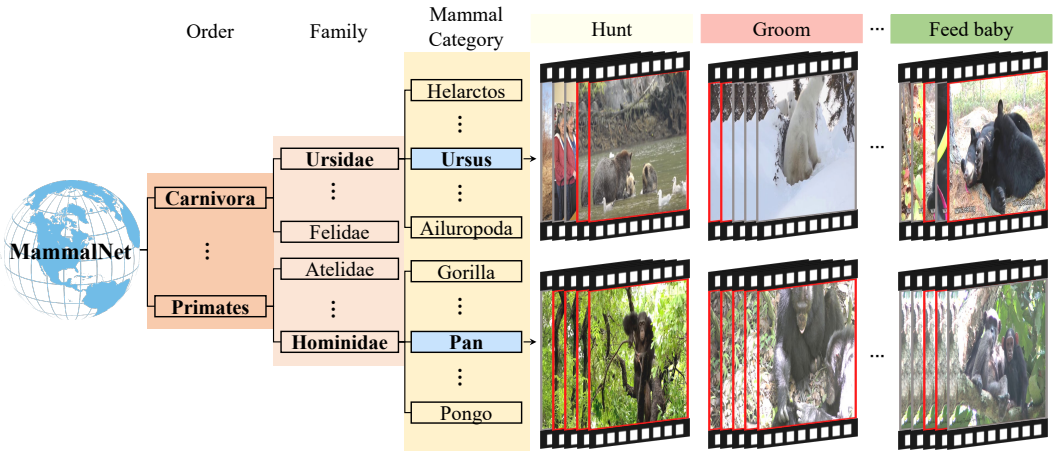

Figure 1. We propose MammalNet, a large-scale video benchmark for recognizing mammals and their behavior. It is built around a biological mammal taxonomy spanning 17 orders, 69 families and 173 mammal categories, and includes 12 common high-level mammal behaviors (e.g., hunt, groom). MammalNet enables the study of animal and behavior recognition, both separately and jointly. It also facilitates investigating challenging compositional scenarios which test models' zero- and low-shot transfer abilities. Moreover, MammalNet includes behavior detection by localizing when a behavior occurs in an untrimmed video. Our dataset is the first to enable animal behavior analysis at scale in an ecologically-grounded manner, and exemplifies multiple challenges for the computer vision community, such as recognition of imbalanced, hierarchical distributions of fine-grained categories and generalization to unseen or seldom seen scenarios.

Abstract

Monitoring animal behavior can facilitate conservation efforts by providing key insights into wildlife health, population status, and ecosystem function. Automatic recognition of animals and their behaviors is critical for capitalizing on the large unlabeled datasets generated by modern video devices and for accelerating monitoring efforts at scale. However, the development of automated recognition systems is currently hindered by a lack of appropriately labeled datasets. Existing video datasets 1) do not classify animals according to established biological taxonomies; 2) are too small to facilitate large-scale behavioral studies and are often limited to a single species; and 3) do not feature temporally localized annotations and therefore do not facilitate localization of targeted behaviors within longer video sequences. Thus, we propose MammalNet*, a new large-scale animal behavior dataset with taxonomy-guided annotations of mammals and their common behaviors. MammalNet contains over 18K videos totaling 539 hours, which is ∼10 times larger than the largest existing animal behavior dataset [36]. It covers 17 orders, 69 families, and 173 mammal categories for animal categorization and captures 12 high-level animal behaviors that received focus in previous animal behavior studies. We establish three benchmarks on MammalNet: standard animal and behavior recognition, compositional low-shot animal and behavior recognition, and behavior detection. Our dataset and code have been made available at:* https://mammal-net.github.io.

*Equal contribution.

1. Introduction

Animal species are a core component of the world's ecosystems. Through their behavior, animals drive diverse ecological processes, including seed dispersal, nutrient cycling, population dynamics, speciation, and extinction. Thus, understanding and monitoring the behaviors of animals and their interactions with their physical and social environments is key to understanding the complexities of the world's ecosystems, an objective that is especially critical now given the ongoing biodiversity crisis [12].

Modern sensors, including camera traps, drones, and smartphones, allow wildlife researchers, managers, and citizen scientists to collect video data of animal behavior on an unprecedented scale [43]. However, processing this data to generate actionable, timely insights remains a major challenge. Manual human review and annotation of footage to identify and locate species and behavioral sequences of interest is time-intensive and does not scale to large datasets. Thus, methods for automated animal and behavioral recognition could open the door to large-scale behavioral monitoring and speed up the time to produce usable data, thereby reducing the time to implement management directives.

The first essential step to creating such an AI system for animal and behavior recognition is curating a diverse, representative dataset that allows us to formalize these challenges as computer vision tasks and benchmark potential solutions. Most previous datasets either only cover a limited number of animal and behavior types [4, 38], or do not implement animal labeling [36], or include a small number of videos with insufficient environmental diversity [4,38,48]. Recently, a dataset named "Animal Kingdom" [36] was proposed to study animal actions and is currently the largest existing behavioral dataset, to the best of our knowledge. However, it only contains 4,310 videos totaling 50 hours, which might be insufficient for large-scale animal behavior studies considering its diversity. Furthermore, the authors only focus on the recognition of atomic actions such as yawning, swimming, and flying. These basic actions cannot be easily matched to the higher-order behavioral states that are of primary interest to end users in animal management and conservation [6]. For example, a cheetah that is running may either be hunting, escaping, or playing. Finally, and most importantly, they do not support some important tasks such as animal recognition and behavior detection which are essential for animal behavior understanding.

To overcome the limitations of previous datasets, we propose a new dataset called *MammalNet*. We specifically focus on mammals since they, unlike other animal classes such as birds or insects, usually have more diverse and distinguishable behavior statuses. MammalNet is comprised of 539 hours of annotated videos, which is $\sim$10 times longer than that of the largest available animal behavior dataset. It contains 18,346 videos depicting 12 fundamental high-level behaviors from hundreds of mammal species. Importantly, it focuses on 12 higher-order animal behaviors that are the focus of previous animal behavior literature [3,8,17,33], rather than atomic actions. MammalNet also categorizes animals according to the scientific taxonomy available in Wikipedia, as we show in Fig. 1; hence the dataset can be flexibly expanded in the future by following the same protocols. It includes videos of approximately 800 mammal species in 173 mammal categories. We establish three benchmarks inspired by ecological research needs - standard animal & behavior classification, compositional low-shot animal & behavior recognition, and behavior detection – to promote future study in animal behavior understanding.

Through our experiments, we find that: (1) Correctly recognizing the animals and behaviors is a challenging task even for the state-of-the-art models, especially for less-frequent animals. The top-1 per-class accuracy is 32.5 for animal recognition, 37.8 for behavior recognition, and 17.8 for their joint recognition in our best-performing model. (2) Behavior recognition for unseen animals can be transferred from observations of other seen animals due to their similar features such as appearance and movement style, which can help in studies of animals with less available data. However, to achieve more accurate behavior recognition, having access to videos of the target animals and behaviors is still crucial.

2. Related Work

Automatic animal recognition and behavior detection can help humans monitor and efficiently process animal behavior data [7, 22, 32, 34, 40, 47, 48]. It can massively reduce the labour cost from manually collecting and analyzing animal activities. During the past few years, many datasets [18, 20, 29, 38] have been introduced to develop foundations for animal behavior research. We systematically analyze the previous datasets and summarize several important limitations that prevent them from being used for large-scale animal recognition and behavior understanding:

Lack of behavior understanding. Many previous datasets only focus on animal recognition or pose estimation from images, but lack behavior learning. For example, iNaturalist [45] collects 859,000 images covering more than 5,000 different types of plants and animals. NABird [44] collects 48,562 images of 555 different North American bird species. Also, some works focus on narrow animal recognition such as dogs [27], birds [46, 50] or cats [13]. On the other hand, many works also focus on pose estimation [10,18,21,31,41] and animal face detection [26, 39, 51]. They are generally not applicable to learning animal behavior.

Lack of taxonomic diversity for different behaviors. Previous animal behavior datasets have minimal taxonomic coverage - often containing just a single animal species. For example, there are existing behavior recognition datasets for elephants [28], sheep [37], monkeys [5], tigers [16],

Datasets	Dataset Properties								Tasks		
	Publicly Available?	Taxonomy-guided Animal Annotation?	No. of Videos	No. of Actions	No. of Behaviors	No. of Animal Categories	No. of Mammal Categories	Total Duration	Animal Classification	Action/Behavior Recognition	Action/Behavior Detection
Wild Felines [20]	×	×	2,700	3	-	3	3	-	✓	✓	×
Wildlife Actions [29]	×	×	10,600	7	-	32	11	-	✓	✓	×
Animal Kingdom [36]	✓	×	4,301	140	-	850	-	50 (h)	×	✓	×
MammalNet (ours)	✓	✓	18,346	-	12	173	173	539 (h)	✓	✓	✓

Table 1. The comparison among existing animal behavior understanding video datasets. Compared to other datasets, MammalNet annotates the animals by following the scientific mammal taxonomy, focuses on more high-level behavior recognition, has the largest number of mammal categories, collects the largest number of animal behavior videos, totalling 539 hours, and also enables behavior detection tasks.

etc. While useful for the studied species, these datasets do not enable the exploration of behavior across species which is necessary to scale up behavior identification without requiring training examples of every possible combination of species and behavior.

Lack of taxonomy-guided animal annotation. Previous animal behavior datasets either do not include animal recognition as a task [36] or group species according to subjective as opposed to scientific criteria [29, 36]. In contrast, our dataset collects and annotates the videos according to the scientific mammal taxonomy. By following this taxonomy we allow exploration of behavior from an evolutionary perspective, as well as enable standardized and consistent dataset expansion under the same protocol.

Actions vs. Behaviors. Previous works mostly focus on atomic action recognition [20, 29, 36]. For example, some of their action classes are *walk, stand still, fly, etc*. In contrast, MammalNet focuses on higher-level animal behaviors, such as *hunt, feed baby, etc*. Behavior here denotes the main activity instance during a period, and it is usually composed of multiple atomic actions. These complex behaviors are needed to describe and summarize video activity bouts in a manner which is valuable for ecological research.

We further compare our MammalNet dataset with the other existing animal behavior understanding video datasets, and summarize the key difference in Table 1.

3. Constructing MammalNet

The goal of *MammalNet* is to provide a large-scale mammal video dataset that benchmarks both animal and behavior recognition. In this section, we discuss our dataset construction protocol, including the choice of scientific animal taxonomy, crowdsourced annotations, and performing manual quality control during video collection and annotation. Finally, we describe the statistical profile of MammalNet.

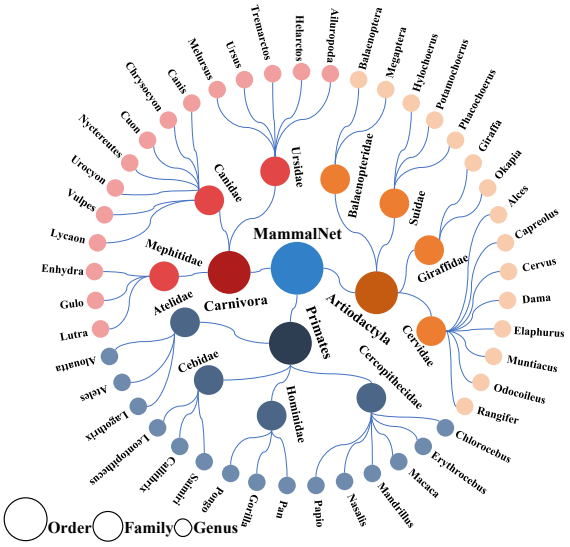

Figure 2. A subset of the mammal taxonomy of MammalNet. It includes 3 orders, 11 families, and 45 genera.

3.1. Animal Taxonomy and Behavior Collection

Animal taxonomy construction. Accurate animal recognition is one of the main goals defined by MammalNet. To ensure this task is scientifically relevant, we aimed to collect and structure our list of target animals based on the mammal taxonomy. First, we collected a diverse list of mammals, approximately 800 mammal species and sub-species, from the **National Geographic** [2] and **Animal A-Z** [1] websites. Next, we manually mapped each mammal onto the taxonomic structure (including class, order, family, genus, tribe, sub-family and species) in mammal taxonomy from Wikipedia. In total, the classes included in MammalNet cover 17 orders, 69 families, and 173 mammal categories. A taxonomic subset is visualized in Fig. 2.

Behavior collection. We aim to enable the study of complex, high-level animal behavior as opposed to the simpler atomic actions emphasized in previous work [29, 36]. Be-

Figure 3. The examples for the annotated target behavior boundaries. The frames marked in red boxes denote the annotated temporal boundaries for the target behavior.

havior here represents the major activity being displayed during a period of a video, and can be viewed as a series of atomic actions collectively serving a higher-level purpose. For example, hunting behavior, as defined in MammalNet, can often be decomposed into running, chasing and killing actions, etc. Identification of behaviors at this level of definition is more useful for ecologists and zoologists [3, 8, 17, 33] compared to atomic actions.

Inspired by previous biological and ecological studies [3, 8, 17], we consider 12 fundamental mammal behaviors under 5 different groups in our study. They are respectively:

Foraging behaviors: eat food, drink water, hunt.

Reproductive behaviors: mate, feed baby, give birth.

Hygiene behaviors: groom.

Agnostic behaviours: fight.

Maintenance behaviors: urinate, defecate, sleep, vomit.

3.2. Video Collection and Quality Assurance

Dataset curation is usually a costly process requiring a lot of manual annotation by humans. Some datasets are annotated by domain experts to encourage more reliable labeling, particularly for challenging tasks, but this expertise comes at a much higher cost and is thus hard to scale. We adapted a semi-automatic crowdsourcing approach to collect and annotate our datasets, inspired by many previous works [9, 15, 23, 49].

Online video retrieval. Our goal was to collect videos depicting each animal in our database performing each of the 12 focal behaviors. To achieve this, we queried YouTube with the text combinations of each animal common name and behavior, e.g., *tiger hunting*, and downloaded videos where the title included our queried animal name and behavior. However, some animal names are too rare to generate enough relevant videos, e.g., *Brown hyena* or *Spotted hyena*. In such cases, we used a more common name, *hyena*, to represent the union of those animals. In order to retrieve more relevant videos from the search engine, we also expanded each animal with their synonyms as given in Wikipedia. For example, we expanded the original *artic fox* with other equivalent names such as *white fox*, *polar fox* and *snow fox*. Each video was downloaded at its highest available resolution.

Data filtering. During video retrieval, we downloaded the videos that are accessible for people $<$16 years old to avoid the videos with violent content. We also prioritized videos that are shorter than 10 minutes in duration to limit the total storage.

However, some videos might have irrelevant content due to the inaccuracy of text-based retrieval. For example, the downloaded videos might 1) display cartoon or toy animals or unrealistic environments such as games or movies, 2) display a static image instead of a continuous video, 3) involve a lot of human-animal interaction, e.g., human feeding an animal, as opposed to focusing on animal behavior, 4) not contain the specified animal and/or behavior.

To alleviate these issues, we employed Amazon Mechanical Turk workers to verify the presence of the animal and behavior, and identify other quality issues. We assigned each video to three different workers and provided them with the pictures of an animal with its common name, and an expected behavior. We asked the workers to verify if this animal and behavior indeed appear in the video. Only the videos that "pass" by all the three workers were kept for the following behavior localization annotation. Before the workers started the verification, we first provided them the verification instructions and asked them to complete a corresponding qualification test with 20 multiple choice questions to ensure only qualified workers could participate in our task.

3.3. Animal Behavior Localization Annotation

The target behavior typically does not span the whole video. To localize the part where the target behavior is actually occurring in the video, we asked the AMT workers to manually annotate the respective temporal boundaries. Namely, we asked five different qualified workers to annotate the start and end frames for the target behavior in each video. In the end, each video received at least 5 annotated behavior boundaries. To achieve robust annotation agreement, we used the complete linkage algorithm [14] to cluster different temporal boundaries and merge them into one or

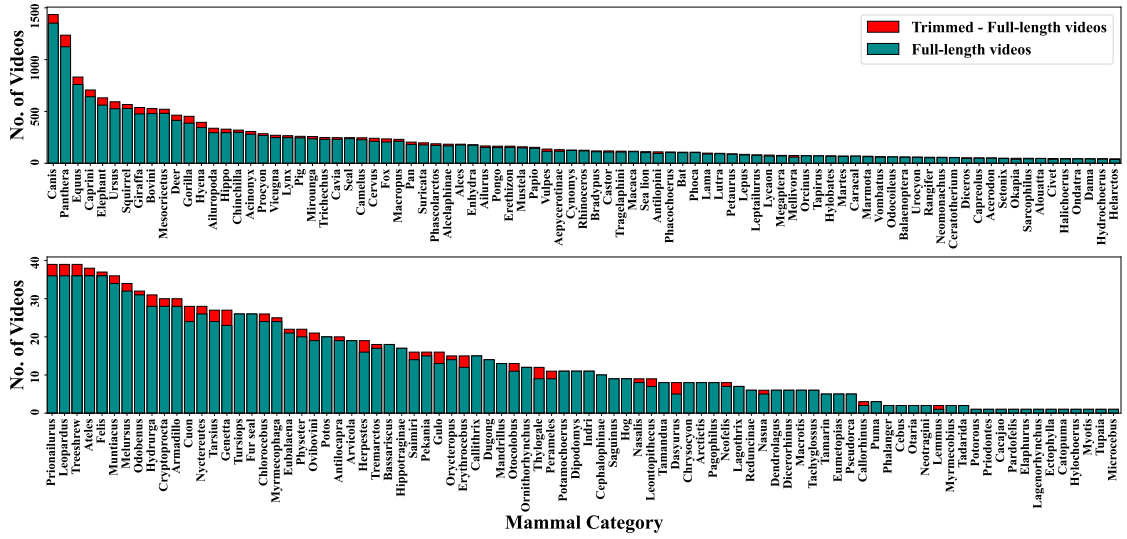

Figure 4. Number of videos per each mammal category. We rank the categories according to their trimmed videos frequency.

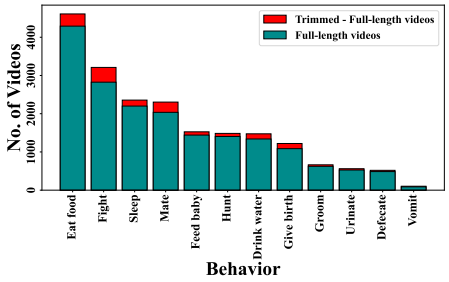

Figure 5. Number of videos per each behavior. We rank the behavior according to their trimmed videos frequency.

several more stable ones that received multiple agreements. Note, that a single video might have multiple discrete occurrences of a given behavior, and thus have multiple boundary definitions. We show several examples for annotated target behavior boundaries in Fig. 3.

3.4. Recognition at lowest feasible taxonomic level

We categorize animals according to the lowest and feasible taxonomic classification, rather than the species level, for the following reasons: 1) YouTube videos contain the ambiguous and inexpert species labels. 2) We are not using expert annotators, and even experts often cannot reliably identify animals at the species or even genus level based on crowdsourced or field images [25]. Thus, it is more practical to classify animals in our dataset to the lowest feasible taxonomic level rather than species. As a result, our taxonomy contains 173 distinguishable taxonomic classification levels including: sub-family, tribe and genus.

3.5. MammalNet Statistics

The final MammalNet dataset contains 18,346 videos (539 hours); after we trim the videos according to the annotated behavior boundaries, it increases to 20,033 trimmed video instances (394 hours). In total, it covers 173 mammal categories, 69 families and 17 orders in our dataset, with the total of 173 mammal categories defined for our recognition tasks. MammalNet also contains 12 different common behaviors. The average duration for the untrimmed and trimmed videos is 106 and 77 seconds, respectively. Over 54% of videos reach HD resolutions (1280 $\times$ 720). We demonstrate the data distribution in terms of each animal category and behavior in Fig. 4 and 5. We observe that the number of collected videos per type follow a long-tail distribution.

4. Experimental Results

We construct three main tasks on MammalNet: 1) Standard animal and behavior classification on trimmed videos, 2) Compositional low-shot animal and behavior recognition on trimmed videos, and 3) Behavior detection on untrimmed videos. In both the classification and detection tasks, we baseline several state-of-the-art models [19, 30, 53] that have been successfully applied to human action recognition and detection. In the following, we describe the formulation of each challenge and provide baselines and analysis.

4.1. Standard Animal and Behavior Classification on Trimmed Videos

This task explores classification of both the primary behavior that occurs in a trimmed video and the animal that performs the behavior. We report the top-1 per-example and per-class accuracy for all the baseline models.

Many, medium, few splits. To capture the effect of the long-tailed nature of the MammalNet dataset, we group the animal, behavior, and their composition classes into *many*, *medium*, and *few* based on their frequency, and report the average

Baselines	Animal Classification				Behavior Classification				Joint Classification			
	Many	Medium	Few	All	Many	Medium	Few	All	Many	Medium	Few	All
	12	28	133	173	4	4	4	12	33	180	823	1036
SlowFast [19]	**49.6**	**35.6**	9.5	17.2	39.0	27.6	**14.9**	27.2	27.2	19.4	4.4	8.6
C3D [42]	48.6	35.3	10.0	17.5	38.2	27.8	11.7	25.9	28.1	18.8	4.5	8.6
I3D [11]	48.8	34.9	10.5	17.8	39.5	27.2	14.8	27.2	**29.6**	**20.5**	4.4	8.9
MViT V2 [30]	48.5	35.5	**12.9**	**19.7**	**42.3**	**29.2**	11.6	**27.7**	29.5	19.6	**4.8**	**9.0**
SlowFast*	58.3	43.1	16.6	24.5	45.1	32.7	14.8	30.9	38.6	23.5	7.3	12.1
C3D*	58.3	45.4	19.1	26.8	44.6	36.0	15.9	32.2	38.0	26.4	8.4	13.5
I3D*	58.6	42.9	16.9	24.8	46.3	35.0	14.8	32.1	38.3	24.5	8.6	13.3
MViT V2*	**66.7**	**56.0**	**23.4**	**32.5**	**50.9**	**42.4**	**20.0**	**37.8**	**46.2**	**33.0**	**11.8**	**17.8**

Table 2. Per-class Top-1 accuracy for animal, behavior and their joint prediction.* denotes the initialization from the model pretrained on Kinetics 400 [24]. Transfer learning from the human action to the animal behavior recognition receives considerable performance gain. Best performance for each split has been highlighted in **bold**.

Baselines	Animal	Behavior	Joint
SlowFast [19]	35.4	34.2	17.4
C3D [42]	35.0	33.5	17.1
I3D [11]	35.2	34.3	17.9
MViT V2 [30]	**35.6**	**36.8**	**18.0**
SlowFast*	43.0	39.4	22.8
C3D*	44.4	40.3	24.6
I3D*	43.4	41.2	24.0
MViT V2*	**52.6**	**46.6**	**30.6**

Table 3. Per-example accuracy for animal, behavior and their joint prediction. * denotes the initialization from the pretrained model.

per-class accuracy bands chosen based on the frequency percentiles. For animal categories this is broken down as *many*: top 7% frequent classes, *medium*: middle 16% classes, and *few*: the remaining 77% classes. For behavior, *many*: top 33% frequent classes, *medium*: middle 33%, and *few*: the remaining 33% classes. For joint classification, *many*: top 3% frequent classes, *medium*: middle 17% classes, *few*: the remaining 80% classes. We show the number of classes per each split in Table 2.

Dataset setup. We randomly split the examples from each animal-behavior category into 70% for training, 10% for validation, and 20% for testing, and it results in 14,554 training, 1,638 validation, and 3,841 testing videos, respectively.

Baselines. We compare SlowFast [19], I3D [11], C3D [42] and MViT V2 [30] models on our tasks. These models are evaluated in two versions: 1) Training with random initialization 2) Initializing with weights from a model pretrained on Kinetics 400 [24]. These methods were originally designed for human action recognition and hence do not have an ability to predict both an action and a subject by default. To accommodate them into our joint prediction setting, we have two task heads, one for animal category recognition and one for behavior recognition. We compute the joint loss, $\mathcal{L}_{joint}$, for both animal and behavior classification as shown in Fig. 1. We tune all the hyper-parameters on the validation data. The final hyper-parameters for each model are provided in the supplement. The loss is defined as:

$$\mathcal{L}_{joint} = -\frac{1}{M}\sum_{i}^{M} y_i^a \log(p_i^a) - \frac{1}{N}\sum_{j}^{N} y_j^b \log(p_j^b) \quad (1)$$

where M is the number of animal classes, and N is the number of behavior classes, p_i^a and y_i^a denote the animal prediction probability and ground truth label for the category i, p_y^b and y_j^b denote the behavior prediction probability and ground truth label for the category j.

Experimental results. The results for per-class and per-example classification are summarized in Tables 2 and 3, respectively. We find that MViT v2 is able to achieve competitive results for all the splits. The best top-1 joint per-class accuracy is 17.8 and per-example accuracy is 30.6, which points to significant room for improvement on these challenging tasks. We also observe that transfer learning from the model that is pretrained on Kinetics 400, a human action recognition dataset, improves both the animal and behavior classification accuracy (with MViT v2, this corresponds to a per-class accuracy gain from 19.7 to 32.5 for animal classification, and from 27.7 to 37.8 for behavior classification). Additionally, we find that performance gain from pretraining is higher for frequently occurring animals and behaviors, indicated by the results shown in $many$, $medium$, and few splits. Accurately predicting low-frequency animal and behavior categories remains a significant challenge.

4.2. Compositional Low-shot Animal and Behavior Classification on Trimmed Videos

It is hard to find a sufficient number of labeled behavior samples for all the animals in our collected taxonomy. For example, *numbat* and *florida panther* have very few behavior

Baselines	Compositional Low-Shot Behavior Classification					
	0-shot		1-shot		5-shot	
	Per-example A / B	Per-class A / B	Per-example A / B	Per-class A / B	Per-example A / B	Per-class A / B
C3D*	27.4 / 23.7	18.3 / 16.1	29.2 / 25.5	21.1 / 19.9	33.3 / 29.3	25.2 / 23.0
I3D*	25.3 / 23.9	16.7 / 15.2	26.8 / 25.7	16.9 / 18.3	30.5 / 28.3	21.7 / 21.9
SlowFast*	26.2 / 24.8	17.9 / 16.3	26.5 / 26.3	16.7 / 19.0	29.6 / 29.2	22.5 / 22.4
MViT V2*	**32.2 / 26.2**	**20.7 / 18.1**	**33.7 / 28.9**	**22.9 / 22.5**	**39.3 / 31.7**	**31.0 / 26.0**

Table 4. Compositional low-shot animal and behavior recognition. * denotes the initialization from the model pretrained on Kinetics 400 [24]. "A" denotes the animal category and "B" denotes the behavior category. The best performance per each column has been highlighted in **bold**.

Baselines	mAP					
	0.50	0.60	0.70	0.80	0.90	Avg.
CoLA [52]	26.02	22.70	18.98	13.46	3.05	15.81
TAGS [35]	23.09	20.97	19.09	16.98	**12.56**	17.63
ActionFormer [53]	**28.48**	**26.14**	**23.17**	**18.69**	10.48	**20.07**

Table 5. The results for behavior detection. We report mAP at the IoU thresholds of [0.5:0.1:0.9]. Average mAP is computed by averaging different tIoU thresholds.

annotations. However, it is plausible to imagine that behaviors can be transferred among different animals that have similar appearances and movement styles. Also the animal recognition under different behaviors (hunting, fighting) can be mutually transferred. To investigate these phenomena, we design the compositional low-shot animal and behavior classification task.

Dataset setup. We first select the animal-behavior compositional classes that contain more than 5 examples and allocate 25% of classes to the test set (4,088 videos). For the remaining classes (the other 75% of classes and those classes with ≤ 5 examples), we randomly designate 90% of classes as the training set and 10% as the validation set (898 videos). Under the low-shot scheme, for each compositional class, we randomly sample 5 examples from the test set and move 0, 1, or 5 of them into the training set for the zero-shot, 1-shot, and 5-shot setup, respectively (14,377, 14,511 and 15,047 training videos). The train, val and test sets consist of 983, 53, and 134 compositional classes, respectively.

Baselines. We evaluate the compositional low-shot classification with the SlowFast [19], I3D [11], C3D [42] and MViT V2 [30] models under the joint loss. We initialize these models with the weights pretrained on Kinetics 400 [24].

Experimental results. We summarize the results in Table 4. It shows that MViT v2 consistently achieves the best performance under our low-shot setup. The behavior classification can still achieve 26.2 per-example and 18.1 top-1 accuracy under the zero-shot setting for MViT v2 model. This indicates that behavior classification is transferable from other animals in some cases. Additionally, we observe that its performance is improved to 31.7 per-example and 26.0 per-class top-1 accuracy under 5-shot setup, indicating that training on more videos with the behaviors from the same animal is still necessary. A similar phenomenon is observed to the few-shot animal recognition.

4.3. Behavior Detection on Untrimmed Videos

This task is to detect the behavior in untrimmed videos. The behavior detection algorithm should correctly detect the temporal range for the primary activity presented in the video. We follow previous temporal action localization works [9, 53] to benchmark this task. We report the mean Average Precision (mAP) with different temporal intersections over the union (tIoU) thresholds [0.5:0.1:0.9] as our evaluation metric. We also report the average mAP averaging across different tIoUs.

Dataset setup. Similar to the previous standard animal and behavior classification, we follow the [train:0.7, val:0.1, test:0.2] ratios to split the untrimmed videos at the animal-behavior composition level. This results in 13,318 videos for training, 1,486 for validation and 3,542 for testing. We tune all the hyper-parameters based on the validation set.

Baselines. We evaluate our datasets with the baseline models such as ActionFormer [53], TAGS [35], and CoLA [52]. To produce the features for our MammalNet videos, we first finetune a two-stream I3D [11] model, that is originally pretrained on ImageNet [15] and Kinetics 400 [24], on our dataset, and then extract the RGB and optical flow features for each video. We concatenate these two features together as the model input.

Experimental results. We show the behavior detection results in Table 5. Among all the baselines, ActionFormer demonstrates the most competitive performance with an average mAP of 20.07, and also achieves 28.48 mAP for the threshold of 0.5. Overall, it is clear to see that the behavior detection task is still very challenging for current methods.

5. Analysis

Demonstration of animal and behavior classification. We sample some prediction examples from MViT v2 [30] for

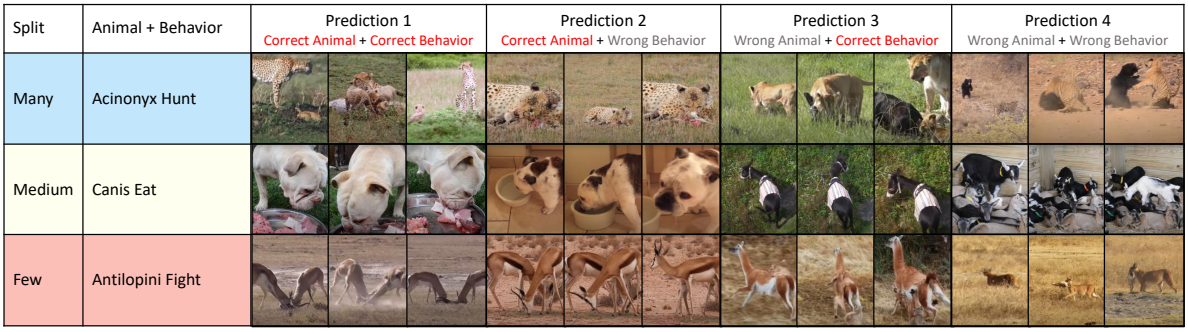

Figure 6. The visualization presents various instances of joint animal and behavior classification. In the third column, accurate predictions are displayed, while the fourth, fifth, and sixth columns showcase mispredicted examples where either the animal or the behavior does not correspond to the correct prediction.

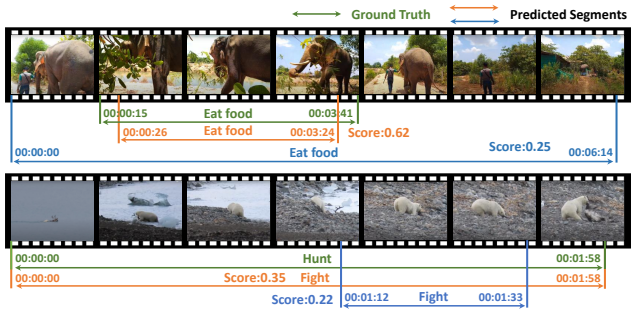

Figure 7. The visualization for behavior detection examples.

the joint categories *Acinonyx hunt*, *Canis eat* and *Antilopini fight*. We demonstrate one correct prediction and three mispredicted examples where the model mis-recognizes the animal or behavior or both of them in Fig. 6.

Demonstration of behavior detection: We visualize behavior detection results in Fig. 7 from ActionFormer [53]. The top example shows correctly predicted behavior with a proposal closely aligned with the ground-truth. The bottom example shows a misclassified behavior, and it mistakenly predicts the *hunt* behavior as *fight*.

Joint animal and behavior recognition vs. separate recognition. We also train the model for recognizing the animal and behavior separately and provide the results in the Table 1 of the supplementary. Comparing with the joint recognition, we find that training a system to recognize the animal and behavior together can improve the behavior recognition under separate training by $\sim$2.2 per-class accuracy. This indicates that being capable of understanding the animal types can benefit behavior prediction. However, the results also indicate that predicting animal only can be more useful in recognizing animal types.

Dataset bias. Our dataset is downloaded from YouTube channels with diverse backgrounds and video quality, and it might differ from video datasets collected for behavioral or ecological studies. The videos on YouTube might also exhibit potential biases:

1) Over-representation of some behaviors and under-representation of others. For example, people might prefer watching videos with *fight* or *eat food* activities, and these two behaviors are more likely to be over-represented, while *urinate* and *defecate* behaviors are less interesting to humans and hence are underrepresented on YouTube. However, they are still important in informing conservation-related actions to protect the environment.

2) Bias towards captive animals or wild animals that are habituated to humans. We find that many videos are shot at zoos, farms, and homes, etc. These animals may display different behaviors than wild or non-habituated animals.

6. Conclusion

We introduced MammalNet, a large-scale video dataset for mammal recognition and behavior understanding. We have collected videos for hundreds of different mammals and structured them by following the scientific mammal taxonomy. MammalNet consists of 18,346 untrimmed videos covering 173 mammal categories and 12 common behaviors. We established three challenges: standard animal and behavior classification, compositional low-shot animal and behavior classification, and behavior detection. Through our experiments, we found that the accurate recognition of animals at scale and their common behaviors is very challenging even with current state-of-the-art models, especially when the dataset has a long-tail distribution. We also found that learning to recognize the behavior of unseen animals is possible via transfer from the other seen animals. To promote further research and development in the field of animal behavior study, we have open-sourced all of our data and code to the research community.

Acknowledgement: This work is supported by KAUST BAS/1/1685-01-01 and KAUST FCC/1/1973-58-01 (Red Sea Research Center), DARPA's SemaFor and PTG programs, the Caltech Resnick Sustainability Institute, and Germany's Excellence Strategy–'Centre for the Advanced Study of Collective Behaviour' EXC 2117-422037984.

References

[1] A-z animals. https://a-z-animals.com/. 3

[2] National geographics. https://www.nationalgeographic.com/. 3

[3] John Alcock. *Animal behavior: An evolutionary approach*. Sinauer Associates, 2009. 2, 4

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

[5] 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

[6] Oded Berger-Tal, Daniel T Blumstein, Scott Carroll, Robert N Fisher, Sarah L Mesnick, Megan A Owen, David Saltz, Colleen Cassady St Claire, and Ronald R Swaisgood. A systematic survey of the integration of animal behavior into conservation. *Conservation Biology*, 30(4):744–753, 2016. 2

[7] Cigdem Beyan and Robert B Fisher. Detection of abnormal fish trajectories using a clustering based hierarchical classifier. In *BMVC*, 2013. 2

[8] Michael D Breed and Janice Moore. *Animal behavior*. Academic Press, 2021. 2, 4

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 4, 7

[10] 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, 2019. 2

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6, 7

[12] 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. 2

[13] Yu-Chen Chen, Shintami C Hidayati, Wen-Huang Cheng, Min-Chun Hu, and Kai-Lung Hua. Locality constrained sparse representation for cat recognition. In *International Conference on Multimedia Modeling*, pages 140–151. Springer, 2016. 2

[14] Daniel Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977. 4

[15] 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. 4, 7

[16] Rod K Dishman, Andrew S Jackson, and Molly S Bray. Self-regulation of exercise behavior in the tiger study. *Annals of Behavioral Medicine*, 48(1):80–91, 2014. 2

[17] Lee Alan Dugatkin. *Principles of animal behavior*. University of Chicago Press, 2020. 2, 4

[18] 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

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 5, 6, 7

[20] Liqi Feng, Yaqin Zhao, Yichao Sun, Wenxuan Zhao, and Jiaxi Tang. Action recognition using a spatial-temporal network for wild felines. *Animals*, 11(2):485, 2021. 2, 3

[21] 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. 2

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

[23] Fabian Caba Heilbron and Juan Carlos Niebles. Collecting and annotating human activities in web videos. In *Proceedings of International Conference on Multimedia Retrieval*, pages 377–384, 2014. 4

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6, 7

[25] Roland Kays, Monica Lasky, Maximilian L Allen, Robert C Dowler, Melissa TR Hawkins, Andrew G Hope, Brooks A Kohli, Verity L Mathis, Bryan McLean, Link E Olson, et al. Which mammals can be identified from camera traps and crowdsourced photographs? *Journal of Mammalogy*, 2022. 5

[26] Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, and Georgios Tzimiropoulos. Animalweb: A large-scale hierarchical dataset of annotated animal faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6939–6948, 2020. 2

[27] 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. 2

[28] Nicole Laws, Andre Ganswindt, Michael Heistermann, Moira Harris, Stephen Harris, and Chris Sherwin. A case study: fecal corticosteroid and behavior as indicators of welfare during relocation of an asian elephant. *Journal of Applied Animal Welfare Science*, 10(4):349–358, 2007. 2

[29] Weining Li, Sirnam Swetha, and Mubarak Shah. Wildlife action recognition using deep learning. 2020. 2, 3

[30] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 4804–4814, 2022. 5, 6, 7

[31] 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. 2

[32] 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. 2

[33] Joy Mench. Why it is important to understand animal behavior. *ILAR journal*, 39(1):20–26, 1998. 2, 4

[34] 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

[35] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Temporal action detection with global segmentation mask learning. *European Conference on Computer Vision*, 2022. 7

[36] 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

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

[38] 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. 2

[39] Maheen Rashid, Xiuye Gu, and Yong Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6894–6903, 2017. 2

[40] 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

[41] Moira Shooter, Charles Malleson, and Adrian Hilton. Sydog: A synthetic dog dataset for improved 2d pose estimation. *arXiv preprint arXiv:2108.00249*, 2021. 2

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6, 7

[43] 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. 2

[44] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015. 2

[45] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2

[46] Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie. Exploring fine-grained audiovisual categorization with the ssw60 dataset. 2

[47] Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie. Exploring fine-grained audiovisual categorization with the ssw60 dataset, 2022. 2

[48] 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. 2

[49] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International journal of computer vision*, 101(1):184–204, 2013. 4

[50] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 2

[51] Heng Yang, Renqiao Zhang, and Peter Robinson. Human and sheep facial landmarks localisation by triplet interpolated features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. 2

[52] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16010–16019, June 2021. 7

[53] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 2022. 5, 7, 8