Deep Dive into KABR: A Dataset for Understanding Ungulate Behavior from In-Situ Drone Video

Maksim Kholiavchenko[1*], Jenna Kline[2], Maksim Kukushkin[3,4], Otto Brookes[5,11], Sam Stevens[2], Isla Duporge[9], Alec Sheets[2], Reshma R. Babu[2], Namrata Banerji[2], Elizabeth Campolongo[2], Matthew Thompson[2], Nina Van Tiel[6], Jackson Miliko[7], Eduardo Bessa[8], Majid Mirmehdi[5], Thomas Schmid[4,10], Tanya Berger-Wolf[2], Daniel I. Rubenstein[9], Tilo Burghardt[5], Charles V. Stewart[1]

[1*]Rensselaer Polytechnic Institute, Troy, NY, United States.
[2]The Ohio State University , Columbus, OH, United States.
[3]Leipzig University, Leipzig, Germany.
[4]Martin Luther University Halle-Wittenberg, Halle (Saale), Germany.
[5]University of Bristol, Bristol, United Kingdom.
[6]Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland.
[7]Mpala Research Centre, Nanyuki, Kenya.
[8]University of Brasilia, Brazil, Brasilia.
[9]Princeton University, Princeton, NJ, United States.
[10]Lancaster University, Leipzig, Germany.
[11]Wild Chimpanzee Foundation, Leipzig, Germany.

*Corresponding author(s). E-mail(s): kholim@rpi.edu;
Contributing authors: kline.377@osu.edu;
kukushkin@informatik.uni-leipzig.de; otto.brookes@bristol.ac.uk;
stevens.994@osu.edu; isla.duporge@princeton.edu; sheets.256@osu.edu;
rameshbabu.3@osu.edu; banerji.8@osu.edu; campolongo.4@osu.edu;
thompson.4509@osu.edu; nina.vantiel@epfl.ch; jacksonmiliko@gmail.com;
profbessa@unb.br; majid.mirmehdi@bristol.ac.uk;
thomas.schmid@medizin.uni-halle.de; berger-wolf.1@osu.edu;
dir@princeton.edu; tb2935@bristol.ac.uk; stewart@rpi.edu;

Abstract

In this paper, we extend the dataset statistics, model benchmarks, and performance analysis for the recently published KABR dataset, an in situ dataset for ungulate behavior recognition using aerial footage from the Mpala Research Centre in Kenya. The dataset comprises video footage of reticulated giraffes (lat. *Giraffa reticulata*), Plains zebras (lat. *Equus quagga*), and Grévy's zebras (lat. *Equus grevyi*) captured using a DJI Mavic 2S drone. It includes both spatiotemporal (i.e., mini-scenes) and behavior annotations provided by an expert behavioral ecologist. In total, KABR has more than 10 hours of annotated video. We extend the previous work in four key areas by: (i) providing comprehensive dataset statistics to reveal new insights into the data distribution across behavior classes and species; (ii) extending the set of existing benchmark models to include a new state-of-the-art transformer; (iii) investigating weight initialization strategies and exploring whether pretraining on human action recognition datasets is transferable to in situ animal behavior recognition directly (i.e., zero-shot) or as initialization for end-to-end model training; and (iv) performing a detailed statistical analysis of the performance of these models across species, behavior, and formally defined segments of the long-tailed distribution. The KABR dataset addresses the limitations of previous datasets sourced from controlled environments, offering a more authentic representation of natural animal behaviors. This work marks a significant advancement in the automatic analysis of wildlife behavior, leveraging drone technology to overcome traditional observational challenges and enabling a more nuanced understanding of animal interactions in their natural habitats. The dataset is available at https://kabrdata.xyz.

Keywords: Ungulates Behavior Recognition, Behavior Recognition from Drone Footage, Zebra Behavior Recognition, Giraffe Behavior Recognition

1 Introduction

Behavior, in the context of animal studies, is broadly defined as the way an animal acts or reacts in response to certain stimuli or situations [1]. Animal behavior encapsulates a wide range of activities and interactions that take place in an animal's life. Understanding animal behavior is vital not only for ecological and conservation reasons [2], but also because it provides insights into how different species adapt to their environment, how they communicate, and how they socialize [3]. These insights into animal behavior have implications for a variety of fields, from wildlife management and conservation to agriculture and veterinary medicine.

Studying animal behavior in natural habitats (i.e., in situ) is undeniably important yet presents significant challenges [4]. The primary difficulty lies in locating animals and positioning oneself to observe their behaviors unobscured and clearly. Traditionally, two manual methods have been employed to observe animal behavior: focal sampling, which involves recording the behavior of a selected individual for a fixed period of time, and scan sampling, which entails recording the behavior of multiple individuals within a time interval as the observer gradually sweeps their line of sight through a defined field of view, documenting the behaviors [5]. These methods,

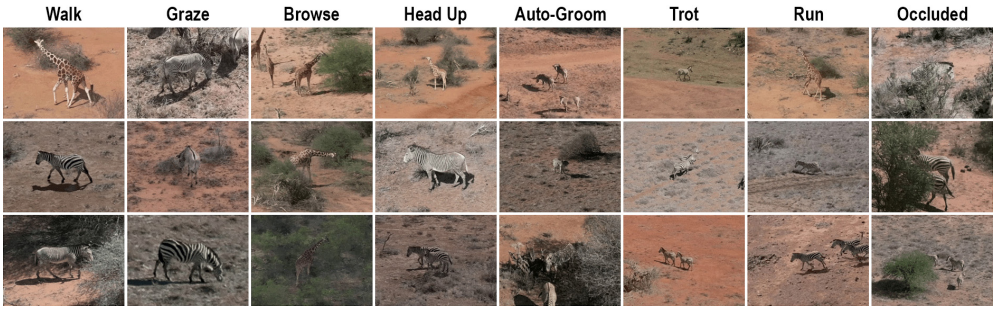

Fig. 1 Examples of the behavior of giraffes, Plains zebras, and Grévy's zebras from our dataset. The dataset includes eight distinct categories: "Walk", "Graze", "Browse", "Head-Up", "Auto-Groom", "Trot", "Run", and "Occluded".

however, capture only a limited fraction of the full behavioral repertoire. The dual challenges of restricted access and limited observations can potentially be mitigated through remote monitoring methods. Increasingly, ethologists have been experimenting with drone-based monitoring and computer vision-aided techniques to locate and track individual animals, thereby enhancing the recording of behavior [6].

The development of modern computer vision technologies for studying animal behavior critically depends on the construction of well-curated datasets. Recently, several large-scale datasets have been proposed for animal behavior recognition [7, 8]. These datasets are typically sourced from online platforms such as YouTube, enabling the collection of a diverse array of species and behaviors. However, there remains a distinct need for behavior recognition datasets collected in situ (i.e., observed and recorded directly in the natural habitat of the species) to provide a more natural representation of behaviors. To improve the applicability of using drones to monitor animal behavior in the wild, it is important to develop experimental datasets.

This work represents an initial step towards addressing these critical needs. By introducing a novel dataset collected from drone videos of Kenyan wildlife in their natural habitats, we aim to enhance the current resources available for the study of animal behavior. This dataset, meticulously designed to reflect in situ scenarios, marks a pioneering effort to capture the complexities of real-world animal behavior. Specifically focused on Kenyan wildlife, it encompasses behaviors of giraffes, Plains zebras, and Grévy's zebras, though the methodology is applicable to other species and environments. The current dataset includes a total of eight categories that describe various animal behaviors. Examples of selected behaviors are shown in Figure 1.

This paper presents extended dataset statistics, model benchmarks, and performance analysis for the KABR [9] dataset. Below, we restate the contributions of the original dataset and list those offered by this extended work:

1. We introduce a novel technique for building a dataset for behavior recognition from drone videos; see Figure 2. We detect and track each individual animal in each high-resolution video and link the results into tracklets. For each tracklet, we create a separate video, called a mini-scene, by extracting a sub-image centered on each

detection in a video frame. This allows us to compensate for the movement of the drone and provides a stable and zoomed-in representation of the animal. This also preserves fine-grained details of animal behavior, such as auto-grooming.
2. We present a new dataset for animal behavior recognition collected in situ and from drones, focused specifically on Kenyan wildlife. The dataset, referred to as Kenyan Animal Behavior Recognition (KABR), comprises annotated mini-scenes and provides a natural view of animal behavior in the wild, resulting in 10 hours of annotated image sequences in the Charades [10] format.
3. We provide comprehensive dataset statistics and extensive behavior recognition baselines using state-of-the-art deep learning models for video classification. The transformer model, UniformerV2, is benchmarked on the KABR dataset, and the original models are reevaluated. As part of this, we perform an in-depth analysis of weight initialization strategies, comparing randomly initialized weights to those pre-trained on human action recognition datasets (i.e., Kinetics-400 [11]) and investigate model performance in an end-to-end and zero-shot setting. A detailed examination of model performance across different segments of the data distribution, including class-wise and species-wise analysis, as well as a formal evaluation of head and tail classes, is also provided. We report that the best-performing model demonstrates mean average precision (mAP) of 66.36%, further highlighting the challenging nature of the KABR dataset.

Our contributions provide a valuable resource for researchers studying animal behavior and ecology, particularly in the context of wildlife conservation efforts in Kenya. By accurately categorizing and analyzing animal behaviors, we can better understand their natural patterns and inform conservation strategies to protect endangered animals.

Small Data Statement

Collecting high-quality in situ data for animal behavior recognition presents significant logistical challenges and requires specialized equipment and expertise. Our dataset, meticulously gathered over a specific period and geographic area at the Mpala Research Centre in Kenya is inherently smaller than large-scale datasets sourced from online platforms. This localized and temporally constrained effort, however, offers detailed, behavior-specific annotations crucial for in-depth behavioral analysis, contrasting with the broader but less specific annotations of larger, generalized datasets.

The annotation process for our dataset involved manual labeling of behaviors in each frame by a team of annotators supervised by an expert behavioral ecologist. This labor-intensive process ensures high accuracy and reliability but limits the dataset's size. To address these limitations, we used state-of-the-art computer vision techniques to detect and track individual animals in high-resolution drone videos. By creating mini-scenes centered on each detected animal, we ensured a focused view of the animal's behavior, enhancing data quality and usability. We also utilized interpolation tools to fill in missing detections within tracks, improving continuity and accuracy. This approach ensures that the extracted mini-scenes are robust and reliable despite the limited dataset size.

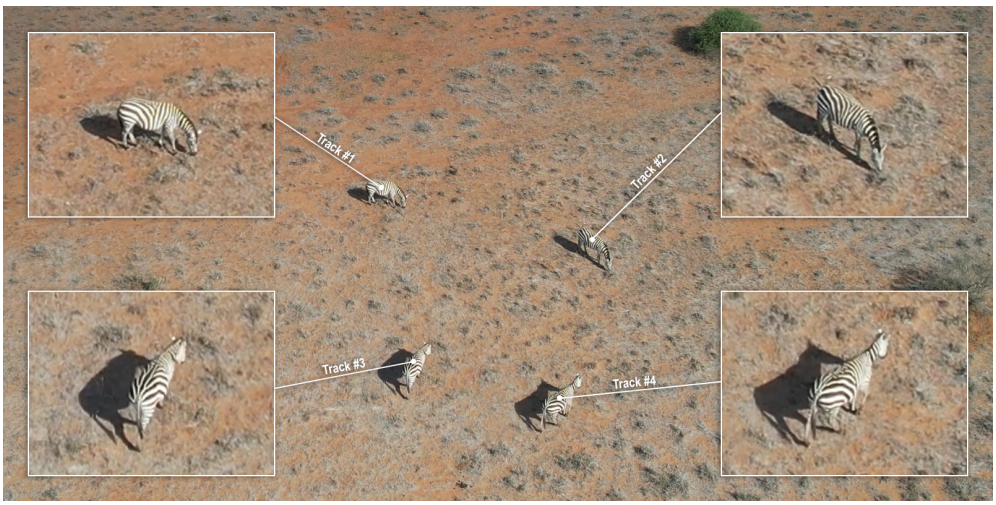

Fig. 2 A mini-scene is a sub-image cropped from the drone video footage centered on and surrounding a single animal. Mini-scenes simulate the camera as well-aligned with each individual animal in the frame, compensating for the movement of the drone and ignoring everything in the large field of view but the animal's immediate surroundings. The KABR dataset consists of mini-scenes and their frame-by-frame behavior annotation.

2 Related Work

Action classification and action detection are distinct tasks within the field of behavior recognition [12]. Although both tasks involve the analysis and understanding of actions, they differ significantly in their objectives and methodologies. The primary goal of action classification is to assign a single category to an entire video, indicating the action being performed [11, 13, 14]. This task focuses on identifying the overall action without specifying its temporal extent or precise location within the video. In contrast, action detection seeks not only to recognize the action category but also to detect and localize the temporal extent of the action within a video [15]. This task involves identifying the specific duration and position of the action. Our concept of mini-scenes integrates both action detection and classification, providing a comprehensive approach to behavior recognition by simultaneously recognizing and localizing actions within a video.

Action recognition datasets, such as Charades [10], UCF [7, 16, 17] and Kinetics [11, 13, 14] have been crucial in advancing the field of behavior recognition. However, these datasets mainly focus on human actions, and the transferability of these datasets to the study of animal behavior is relatively unexplored.

Animal Kingdom [8] and MammalNet [18] are both prominent large-scale datasets for animal behavior recognition. These datasets offer comprehensive collections of annotated video footage featuring a wide range of animal species over 50 and 539 hours, respectively. These datasets primarily rely on videos sourced from online platforms such as YouTube and, therefore, lack the in situ aspect of data collection where observations occur directly in animals' natural habitats. APT-36K [19], also sourced

from YouTube videos, further pushes to bridge the gap between behavior recognition and animal detection with a collection of 80 video clips for each of the 30 species represented. In our paper, we contribute to bridging this gap by introducing a novel in situ dataset specifically centered around Kenyan wildlife.

Prior research has explored the potential of drone videos in addressing challenges related to animal behavior recognition. Notably, Koger et al. [20] introduced a deep learning method focused on reconstructing landscapes from drone videos, enabling the recognition of animal body postures and the ecological context in which they reside. In contrast to the proposed approach, our method is focused on recognizing animal behavior at the individual level rather than understanding the relationship between animals and their landscapes. Additionally, the authors of [21] employed drones to study spatial positioning within groups of feral horses, while [22] used drones to track sharks, unveiling their movement patterns. Furthermore, drone technology was harnessed by [23] for wildlife detection. These diverse applications underscore the potential of drone videos in advancing our understanding of animal behaviors and ecological dynamics.

Several other substantial datasets have been meticulously assembled with a strong focus on recognizing animals [24–26], estimating their poses from images [27, 28], or generating new views of images with animals [29]. For instance, the iNaturalist dataset [30] contains over 859,000 images of more than 5,000 different types of plants and animals. Similarly, the iWildCam [31] dataset contains 263,528 images from 323 locations of camera traps. These datasets provide a plethora of sample images, but they are designed to classify species and count individual animals in images rather than study their behavior.

Some works have proposed targeted solutions for recognizing the behavior of certain animals. These solutions are often based on specific characteristics of the animal's behavior, which may not apply to other species. For instance, a study may focus on recognizing the behavior of primates [32–35], pigs [36–38], goats [39], cows [40, 41], meerkats [42], dogs [43], cats [44], or mice [45–48]. Though these specialized solutions are useful for studying particular animal behaviors, they are typically smaller and may not generalize well to other species or contexts. Therefore, it is important to consider the scope and limitations of these targeted approaches when using them to study animal behavior.

In contrast, our dataset offers a distinctive, valuable contribution to the field of animal behavior recognition, as it focuses specifically on in situ drone videos of Kenyan wildlife. Our innovative approach provides significant benefits over traditional video analysis methods and supplies a valuable resource for researchers studying animal behavior and ecology, particularly within the critical context of wildlife conservation efforts in Kenya.

3 Dataset

3.1 Data Collection

The dataset of drone videos utilized in our study was collected by our research team at the Mpala Research Centre in Kenya. The data collection period extended from

January 6, 2023, to January 21, 2023. Throughout this period, our team undertook multiple expeditions to various locations within the research center's vicinity. The drone flights were meticulously planned to capture the behaviors of giraffes, Plains zebras, and Grévy's zebras. These species were chosen due to their ecological significance and conservation status in the region.

The dataset consists of 1,139,893 individual frames: 488,638 featuring Grévy's zebras, 492,507 of Plains zebras, and 158,748 frames featuring giraffes. In total, there are 14,764 distinct sets of behaviors. To ensure high-quality footage, our team utilized DJI Mavic 2S drones equipped with advanced camera capabilities. The videos were recorded in 5.4K resolution at a speed of 29.97 frames per second, providing a smooth and accurate representation of the animals' behaviors. The drones were flown at varying altitudes and distances from the animals to capture a diverse range of perspectives. The diversity in recording distances allows us to observe behaviors at different scales and will eventually allow us to consider social dynamics within animal groups.

During the flights, the pilot carefully maneuvered the drones to capture the animals' behaviors. Depending on the specific behavior being recorded, the pilot employed a variety of flight paths, including vertical ascents and descents, circular orbits, and linear trajectories. The maneuvers were executed with precision and consideration for the animals' well-being, maintaining a safe and non-intrusive distance.

3.2 Data Curation — Mini-Scenes

The raw drone video data typically contains multiple animals in each frame, with each animal occupying a small fraction of the high-resolution image. In our dataset, the maximum number of animals visible in a single frame is thirteen. Directly analyzing these frames to extract behavior is impractical. Instead, we extract mini-scenes, which are sub-videos of the full-resolution footage. Each mini-scene is centered on an individual animal as it moves through the scene and is cropped to include the animal and its immediate surroundings. This method allows us to compensate for much of the drone's movement and provides a stable, zoomed-in representation of the animal's behavior. This approach facilitates accurate tracking of individual animals within a group. We anticipate that, in future work, this will be particularly useful for studying social dynamics among animals.

To implement our mini-scenes approach, we utilized YOLOv8 [49] to detect the animals in each frame and the SORT [50] tracking algorithm to follow their movement. We then extract a window of size 400 pixels wide and 300 tall, values determined empirically based on the characteristics of the animals observed and the surrounding environment and properties of the drone. We pay special attention to ensuring that the animal fits entirely into the mini-scene based on the dimensions of the bounding box.

We have developed a set of tools to facilitate the data annotation process. One of the tools we used extensively was the interpolation tool, which filled in any missing detections within a track, thereby improving the overall tracking quality. The tool uses a linear interpolation algorithm that estimates an animal's location based on its previous movements, helping fill in gaps where automatic detection may have failed. Our data processing pipeline is illustrated in Figure 3. All mini-scenes must satisfy

a length criterion; if the total length of the behaviors in a mini-scene is less than 90 frames, we filter it out. The processing code is available at [51].

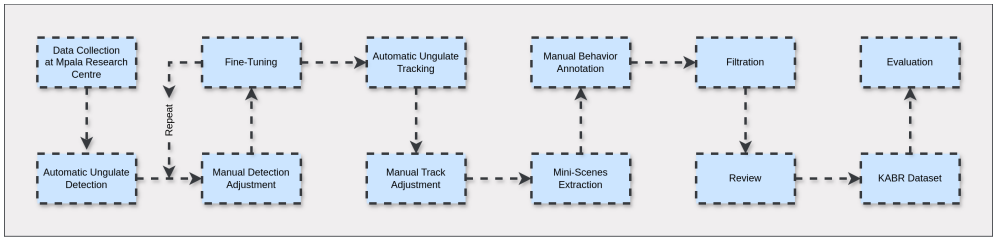

Fig. 3 Overview of the pipeline for KABR dataset preparation.

The mini-scenes we extracted using our pipeline are a crucial component of the manual annotation process for behavior recognition. These mini-scenes provide a zoomed-in and stable view of individual animals' behavior, making it easier for human annotators to accurately identify and label behavior.

3.3 Behaviors and Annotation

Our dataset contains a total of eight behavior categories, including "Walk", "Graze", "Browse", "Head Up", "Auto-Groom", "Trot", "Run", and "Occluded" as determined by our expert behavioral ecologist looking at the properties of the videos. These include three locomotion behaviors, "Walk", "Trot" and "Run", each representing a different gait. "Run" could have been split into canter and gallop, but these were too infrequent and indistinguishable. Two of the other behaviors refer to eating: "Graze" refers to the behavior of an animal when they are eating grass or other vegetation, while "Browse" describes the behavior of animals feeding on trees and bushes. For the remaining categories, "Head Up" refers to the behavior of an animal when it lifts its head to look around or observe its surroundings. Typically, these are different types of vigilance, and "Auto-Groom" describes the behavior of animals when they groom themselves, which can include licking, scratching, or rubbing their bodies. Finally, the category of "Occluded" is used when the animal is not fully visible in the video footage. This can occur due to obstructions such as trees or other animals blocking the view or due to technical limitations of the camera or drone.

To ensure accurate behavior annotation in our dataset, we employed a team of 10 individuals, all of whom were trained in the process. The team was led by an experienced expert behavioral ecologist who oversaw the annotation process. We utilized CVAT [52], a powerful tool for collaborative video annotation, to enable the team to work together remotely and efficiently. Once the initial annotations were complete, we took an additional step to ensure quality control by having all videos manually reviewed by a designated annotator. Finally, we utilized an automatic filtering process to split the annotated videos into convenient training iterations based on their resulting length. This ensured that the training data was properly organized and could

be effectively used in the development of deep learning models. Overall, our comprehensive annotation process and quality control measures ensure that our dataset is accurate, reliable, and suitable for a wide range of research applications.

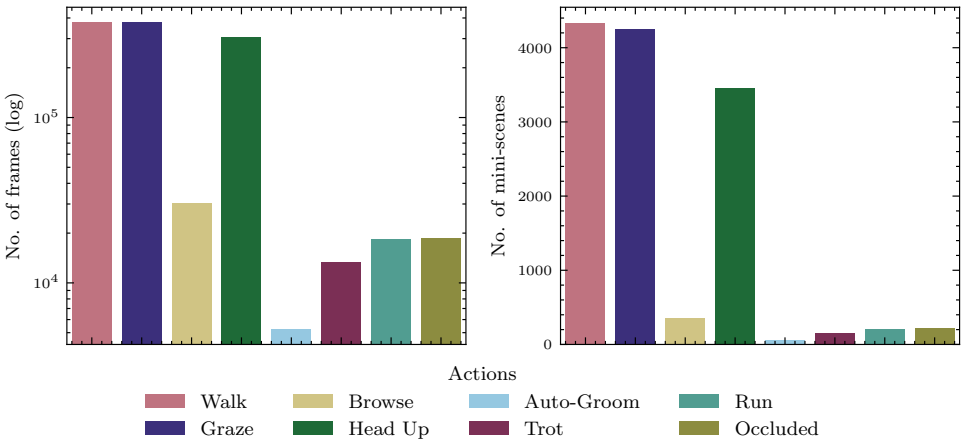

Fig. 4 Distribution of frames (left) and mini-scenes (right) per action in the KABR dataset.

3.4 Data Split

We provide a train-test split of the mini-scenes for evaluation purposes, with 75% for train and 25% for testing. No mini-scene was divided by the split. The splits ensured a stratified representation of giraffes, Plains zebras, and Grévy's zebras.

3.5 Class Distribution

Our dataset exhibits a long-tailed distribution, signifying a considerable disparity in the count of samples across the behavior categories. This is expected since certain behaviors are considerably more frequent in animals' natural settings compared to other behaviors. The distribution of classes is shown in Figure 4. Similar imbalances occur in recent larger datasets [8, 18, 19] scraped from YouTube.

To characterize the imbalance in our dataset, classes are categorized into head and tail segments as defined in [53, 54]. Specifically, we consider classes that contribute 50% of the training samples as head classes and the rest as tail classes. As shown in Figure 4, the two head classes, "Walk" and "Graze", dominate the class distribution, although "Head Up" also contributes a significant proportion of samples.

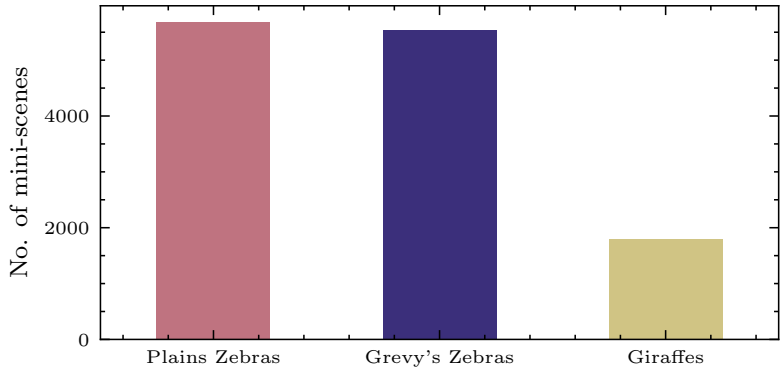

Fig. 5 Distribution of mini-scenes per species in the KABR dataset.

3.6 Species Distribution

Figure 5 shows that the dataset is also somewhat imbalanced with respect to species. While the number of samples for both Zebra species is relatively close, there are significantly fewer samples for Giraffes. Additionally, Figure 6 highlights that the classes are not uniformly distributed *across* species. For example, the most common and rare classes for each species are different. Furthermore, examples of each class are not present for every species; the "Graze", "Auto-Groom", "Trot", and "Occluded" behaviors are observed exclusively in the Zebra species, while "Browse" is only observed for Giraffes.

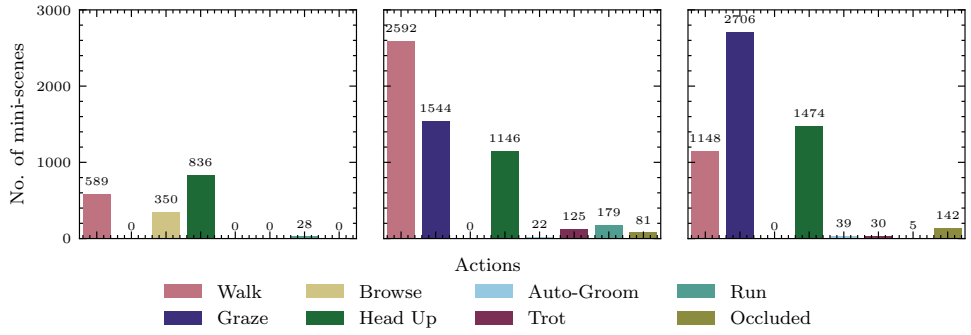

Fig. 6 Distribution of mini-scenes per action for Giraffes (left), Plains zebras (center), and Grévy's Zebras (right).

4 Experiments

To comprehensively assess the performance of different models on our dataset, we conducted evaluations using four well-known architectures: I3D [55], SlowFast [56],

X3D [57], and UniFormerV2 [58]. Models were selected based on SOTA performance achieved for human action recognition datasets and for comparison with other animal behavior recognition benchmarks [8, 18, 35]. The key computational and training details are described in the following section and reported in Table 1.

4.1 Setup

All models were trained for 120 epochs. During training, we used a batch size of 64. To improve the model's performance and reduce the risk of overfitting, we applied data augmentation techniques during training. Specifically, we used horizontal flipping to randomly mirror the input frames horizontally and color augmentations to randomly modify the brightness, contrast, and saturation of the input frames. To address the long-tailed class distribution, we employed the EQL [59] loss function described in Equation 1, which selectively ignores gradients for rare categories, enabling the learning of rare categories during network parameter updates.

$$\mathcal{L}_{\text{EQL}} = -\sum_{j=1}^{C} w_j \log(\hat{p}_j) \tag{1}$$

C is the total number of categories; $\hat{p}_j$ is the predicted probability for class j; w_j is the weighting factor applied to the loss of class j, aimed at reducing the negative gradient contribution from frequent categories for rare categories:

$$w_j = 1 - E(r) \cdot T_\lambda(f_j) \cdot (1 - y_j) \tag{2}$$

$E(r)$ is an indicator function that equals 1 when the region proposal r belongs to the foreground, and 0 otherwise; $T_\lambda(f_j)$ is a threshold function based on the frequency f_j of class j, where λ is a frequency threshold distinguishing rare categories; y_j is the ground truth label for class j.

Table 1 The total number of parameters, gigaflops (GFLOPs), choice of optimizer, sample rate (SR), batch size (BS), and weight initialization (WI) strategy for each model are reported. K-400 indicates Kinetics-400 pre-trained weights and 400M+K710 indicates CLIP-based pretraining on CLIP-400M [60] followed by post pretraining on Kinetics-710.

Model	Optimizer	SR	BS	WI	Params (M)	GFLOPs
X3D-L [9]	SGD	16x5	5	Random	5.35	17.74
I3D	SGD	16x5	64	K-400	27.24	116.47
SlowFast	SGD	16x5, 4x5	64	K-400	33.57	32.82
X3D-L	SGD	16x5	64	K-400	5.35	17.74
UniformerV2-B	AdamW	16x5	64	400M+K710	114.25	148.27

I3D and X3D were trained with 16 input frames with a sampling rate of 5. For SlowFast, the Slow branch was trained with 16 input frames with a sampling rate of 5, and the Fast branch was trained with 4 input frames with a sampling rate of 5.

Models were evaluated using mAP, precision, recall, and F1-score. As described in Section 3.5, behavior classes were grouped, based on class frequency, into head and tail segments, and mAP is reported for each segment. Results for each model are reported

at the best performing epoch (see Figure 7 to compare model performance over all epochs).

4.2 Results

As shown in Table 2, the X3D model initialized with Kinetics-400 pre-trained weights outperforms the other models with respect to mAP. It achieves the highest overall mAP of 66.36%, narrowly surpassing SlowFast at 66.10%. The X3D model also demonstrates superior results for both head and tail classes, with mAP scores of 96.96% and 56.16%, respectively. The SlowFast model shows strong overall performance, achieving the highest recall (65.28%) and F1-score (65.82%), while the I3D model achieves the highest precision (67.17%). The difference in overall mAP between the best (X3D) and worst (I3D) performing models is approximately 1.5%, indicating that while there are differences in model performance, they are relatively small. Additionally, all models perform exceptionally well on head classes (mAP $>$ 95% for all models), but there is a significant drop in performance for tail classes (mAP ranging from 50.58% to 56.16%). This highlights the ongoing challenge of recognizing less frequent behaviors, a common issue in long-tailed recognition tasks.

Table 2 Behavior recognition benchmarks. The precision (P), recall (R), and F1 score (F1) are reported for the I3D, SlowFast, X3D, and UniformerV2 models. Each model is evaluated in an end-to-end (E2E) and zero-shot (ZS) setting. The batch size (BS) and weight initialization (WI) strategy are also shown, where K-400 indicates Kinetics-400 pre-trained weights and 400M+K710 indicates CLIP-based pretraining on [60] followed by post-pretraining on Kinetics-710 [58]. The results from the previous SOTA model are highlighted in gray, and the highest scores across all metrics are shown in bold for each evaluation setting.

	Method	BS	WI	mAP (%)			*P*	*R*	F1
				Overall	Head	Tail			
	X3D-L [9]	5	Random	61.94	96.53	50.40	62.46	61.87	61.53
E2E	I3D	64	K-400	65.06	96.81	54.48	**67.17**	62.94	64.52
	SlowFast	64	K-400	66.10	96.72	55.90	67.05	**65.28**	**65.82**
	X3D-L	64	K-400	**66.36**	**96.96**	**56.16**	66.44	63.65	64.70
	UniformerV2-B	64	400M+K710	61.78	95.38	50.58	64.37	54.82	57.41
ZS	I3D	64	K400	**15.74**	**37.45**	**8.51**	14.10	**72.72**	**21.16**
	SlowFast	64	K400	13.69	36.16	6.19	14.00	26.94	9.03
	X3D-L	64	K400	14.50	34.76	7.74	**14.81**	49.32	19.26
	UniformerV2-B	64	400M+K710	12.74	31.93	6.34	11.01	11.39	10.48
ZS	I3D	64	Random	**18.41**	**50.58**	7.69	4.66	62.50	7.63
	SlowFast	64	Random	11.95	30.13	5.89	11.41	61.81	**16.01**
	X3D-L	64	Random	12.49	32.87	5.70	10.41	**65.86**	15.66
	UniformerV2-B	64	Random	17.80	43.14	**9.36**	**17.17**	13.53	8.97

Furthermore, Table 2 shows that all models initialized with Kinetics-400 pre-trained weights outperform the best-performing baseline reported earlier in [9]. The same model architecture (i.e., X3D) achieves a $\sim$ 4% improvement in mAP under this training setup. Additionally, while performance is comparable for head classes, significant improvement is observed for tail classes.

In the zero-shot setting, performance drops significantly across all models. When compared to models initialized with either Kinetics-400 or 400M+K710 weights, the I3D model performs best, achieving the highest mAP scores for overall (15.74%), head

(37.45%), and tail (8.51%) classes, as well as the best recall (72.72%) and F1 score (21.16%). Interestingly, when initialized with random weights, the I3D outperforms its Kinetics pre-trained counterpart, improving on overall (18.41%) and head (50.58%) mAP scores. This trend is repeated by the UniformerV2-B model, which outperforms its pre-trained counterpart across all metrics except the F1 score. These results suggest that while pre-training on large-scale human action recognition datasets, such as Kinetics-400 [11], clearly provide a stronger initialization for end-to-end training, they may not transfer well *directly* to animal behavior, as evidenced in the zero-shot behavior recognition results.

The mAP curves displayed in Figure 7 show that most models maintain mAP above 60% (mAP $>$ 0.60) throughout the training process. Generally, the best performance is achieved by the X3D and SlowFast models. The SlowFast model achieves consistently higher mAP earlier in training, outperforming other models by approximately 2-3% (mAP=0.02-0.03) on average. However, in the later epochs (epoch $>$ 50), X3D demonstrates superior performance, exceeding SlowFast by 1-2% mAP. By the final epoch, X3D achieves the highest mAP, followed by SlowFast, I3D, and UniformerV2, respectively. This convergence behavior suggests that while SlowFast and X3D have largely stabilized, I3D may benefit from extended training, and UniformerV2 requires further optimization to prevent overfitting.

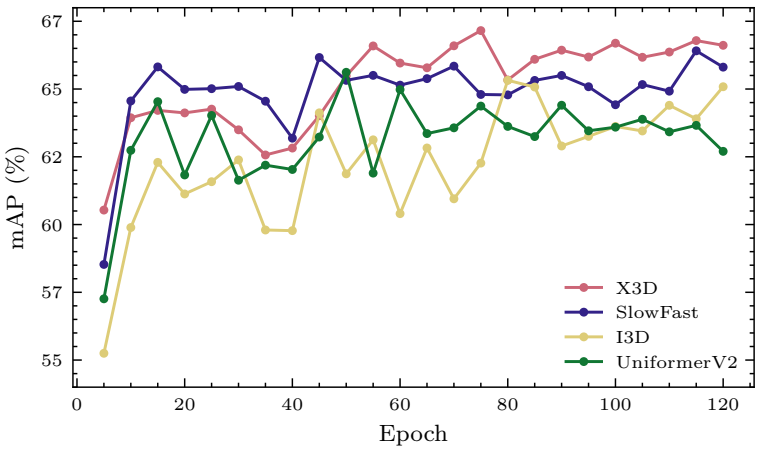

Fig. 7 A comparison of mAP across 120 training epochs on the validation split for the I3D, SlowFast, X3D, and UniformerV2 models.

4.3 Species-wise evaluation

While the differences in overall performance are relatively small, class-wise performance varies across models and species. As shown in Table 3, the SlowFast model demonstrates the highest mean performance for Giraffes and Plains zebras at 83.5%

and 66.3%, respectively, while the I3D model achieves the best results for Grévy's zebras at 70.6%.

Table 3 Species-wise action recognition benchmarks. Results for the I3D, SlowFast, X3D, and UniformerV2 models for Giraffes (G), Plains zebra (ZP), and Grévy's zebras (ZG) for each available action class. The actions are ordered as follows: Walk , Graze, Browse, Head up, Auto-Groom, Trot, Run, Occluded. The head classes, Walk and Graze, are located on the first two rows of the table. Dashes (-) indicate classes that are not observed for a particular species.

I3D			SlowFast			X3D			UniformerV2			Mean	Std
G	ZP	ZG	G	ZP	ZG	G	ZP	ZG	G	ZP	ZG		
96.5	94.0	99.1	96.8	94.3	98.7	97.3	95.3	98.9	92.7	93.9	97.6	96.25	2.0
-	96.6	97.6	-	97.2	95.3	-	96.7	95.6	-	96.4	95.8	96.40	0.7
40.5	-	-	55.2	-	-	43.1	-	-	54.4	-	-	48.30	6.5
95.8	93.2	94.0	94.1	92.8	95.0	94.3	93.6	94.6	89.9	93.1	94.3	93.72	1.4
-	16.9	9.1	-	10.5	0.3	-	15.6	7.1	-	15.9	7.3	10.31	5.3
-	23.7	73.7	-	54.0	78.4	-	31.0	76.8	-	43.7	76.7	57.25	20.8
72.5	69.3	89.5	87.8	97.0	89.1	88.6	81.9	91.3	55.9	92.2	82.1	83.10	11.2
-	15.9	31.2	-	18.4	21.9	-	18.7	21.1	-	27.7	23.8	22.33	4.7
76.3	58.5	70.6	83.5	66.3	68.3	83.0	61.8	69.3	73.2	66.1	68.2		

For head classes (highlighted in gray), all models perform strongly. For the "Walk" action, X3D attains the best score for Giraffes (97.3%) and Plains zebras (95.3%), while I3D performs best for Grévy's zebras (99.1%). Similarly, the "Graze" action, applicable only to zebra species, is well-recognized, with SlowFast achieving the highest accuracy for Plains zebras (97.2%) and I3D for Grévy's zebras (97.6%). It is worth noting that all models also perform strongly on the "Head up" action since, although it is categorized as a tail class, it has a comparable number of samples to both head classes (see Figure 4 and Figure 6). However, performance on tail classes (non-highlighted) shows significant variation. The "Browse" action, exclusive to Giraffes, is best recognized by the SlowFast model (55.2%). For Plains zebras, I3D and UniformerV2 achieve the best performance on "Auto-Groom" (16.9%) and "Occluded" (27.7%), respectively, while SlowFast achieves the best performance on "Trot" (54.0%). However, for Grévy's zebras, a different model achieves the best performance on each of the zebra-specific actions, with the highest scores for "Auto-Groom", "Trot", and "Occluded" being achieved by I3D (9.1%), SlowFast (78.4%), and UniformerV2 (27.7%), respectively. The "Run" action, observed across all species, shows high variability across models and species. X3D achieves the best performance for Giraffes (88.6%) and Grévy's zebras (91.3%), while SlowFast leads for Plains zebras (97.0%).

5 Discussion

The benchmark results using state-of-the-art video classification algorithms indicate that the dataset is both interesting and challenging. Though it is necessarily smaller than recent Animal Kingdom [8] and MammalNet [18] datasets and captures a more focused set of behaviors, it represents an important step in the evolution of animal behavior data collection and analysis because the videos were collected in situ and from drones. As such, it is closer to, and more representative of, how behavioral analysis can be carried out in the field in the future.

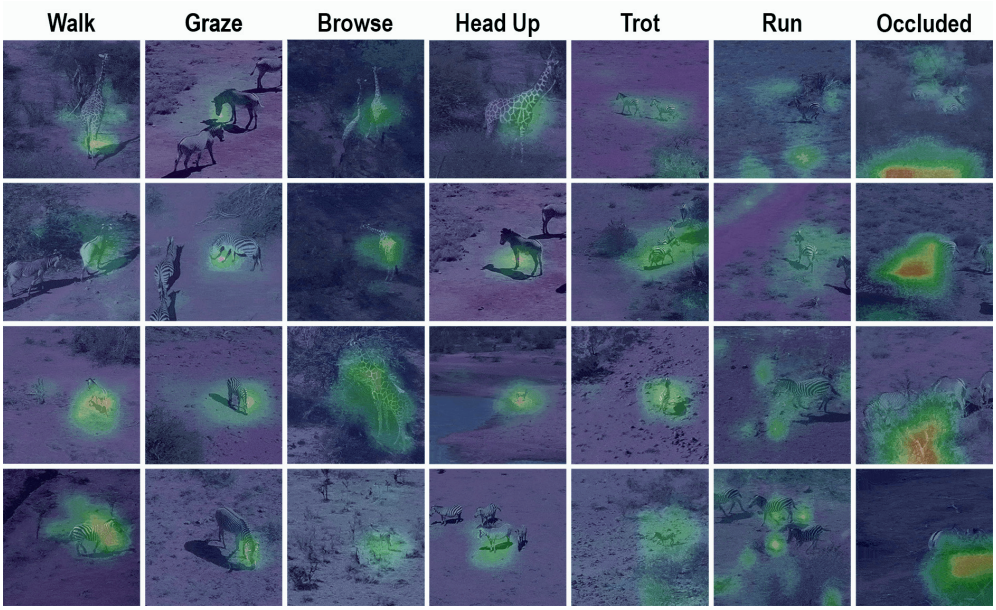

Fig. 8 Grad-CAM visualization for different behaviors in the dataset.

One limitation of the dataset, as it currently exists, is that some rare behaviors are captured infrequently or not at all. The complete set of tools for KABR that we have developed and shared openly form a powerful framework to support searching for examples of these behaviors. The mini-scenes approach provides a means of rapidly processing high-resolution videos into a form that can be analyzed for individual behaviors. The next step would be to augment the behavior classification approaches to facilitate anomaly detection. An interesting question is the potential integration of KABR with MammalNet or Animal Kingdom for exactly this purpose.

The proposed pipeline has several important advantages. By applying detection and tracking algorithms, we can extract zoomed-in footage that is stabilized on the animal of interest. Consequently, the animal remains consistently centered in the frame throughout the mini-scene, enhancing the accuracy of subsequent analysis. This is unlike typical action recognition, where the animal could be moving across a fixed frame. Consequently, if an object moves from one side of the frame to the opposite side, the resulting bounding box may fail to accurately reflect the object's actual position. In contrast, our approach avoids this issue by maintaining the animal of interest at the center of the frame throughout the extracted mini-scene, allowing for more precise localization of the moving object over a longer period of time.

Another important future step is using the mini-scenes approach to analyze complex social behaviors, such as dominance, aggression, mating, and grooming. Behaviors can be analyzed in isolation within each mini-scene, in the overlap between the bounding regions of mini-scenes, and in a graphical representation of a neighborhood of mini-scenes.

A final justification of the efficacy of the mini-scenes approach can be seen in a Grad-CAM [61] analysis of the mini-scene classification activation, as shown in Figure 8. This demonstrates that the neural network indeed prioritizes the region covered by the animal in the center of the frame and even the body part. In the case of the Occluded category, where the animal is not visible within the frame, the network shifts its attention to focus on other objects present. In the case of Run, the background changes very rapidly, especially in the region that is being newly occluded in each frame as the animal moves. This allows the network to identify it as Run.

6 Conclusion

KABR is an in situ dataset designed for animal behavior recognition from drone videos, focusing on Kenyan wildlife, including giraffes, Plains zebras, and Grévy's zebras. It encompasses eight categories that describe various animal behaviors, offering a comprehensive view of animal behavior in their natural habitat. In this paper, we present extended dataset statistics, model benchmarks, and performance analysis. We also revisit our dataset construction technique, which compensates for drone movement and enables the capture of fine-grained details of animal behavior. The benchmark models demonstrate the dataset's effectiveness for training conventional deep-learning models for animal behavior recognition from drone footage. Our contributions provide a valuable resource for researchers studying animal behavior and ecology, particularly within the context of wildlife conservation efforts in Kenya. This work represents a significant advancement in the field of animal behavior recognition and establishes a solid foundation for future research.

Declarations

Funding and/or Conflicts of interests/Competing interests: We declare no competing interests. This work is supported by the National Science Foundation under Award No. 2118240 and Award No. 2112606 and the UKRI CDT in Interactive AI under grant EP/S022937/1.
Availability of data and materials: The KABR dataset official webpage can be found at https://kabrdata.xyz. The dataset is publicly available on Hugging Face under *CC0 1.0 Universal* license — https://doi.org/10.57967/hf/1010. The data processing scripts are available at https://github.com/Imageomics/kabr-tools.
Consent for publication: All authors consent that the publisher has the author's permission to publish research findings.
Ethical considerations: Two important categories of ethical considerations were addressed in our work. First, no humans appeared in the videos, and all participants were faculty, students, or employees of the Mpala Research Centre. Second, our research was conducted under the authority of a Nacosti Research License (No. NACOSTI/P/22/18214). This license confirms our adherence to the regulations in place and allows us to collect drone footage of animals in their natural habitats. We followed a data collection protocol that strictly complies with the guidelines set forth by the Institutional Animal Care and Use Committee (No. IACUC 1835F). These guidelines are designed to ensure the ethical and humane treatment of animals involved

in research activities. We also followed the guidelines laid out in [62]. One particular instance of this is that we consistently approached the animals from downwind, allowing the noise to dissipate before reaching the animals.

Author contributions: M. Kholiavchenko, J. Kline, M. Kukushkin, and O. Brookes are responsible for the methodology and evaluation; C. Stewart, T. Berger-Wolf, D. Rubenstein, T. Burghardt, and M. Mirmehdi, T. Schmid are responsible for the supervision of the research and project administration; J. Kline, S. Stevens, D. Rubenstein, I. Duporge, and J. Miliko are responsible for data collection; M. Kholiavchenko, J. Kline, A. Sheets, R. Babu, N. Banerj, N. Tiel, and E. Bessa are responsible for data annotation; D. Rubenstein is responsible for the supervision of the data annotation team; E. Campolongo and M. Thompson are responsible for data hosting and maintenance.

Appendix A Learning rate analysis

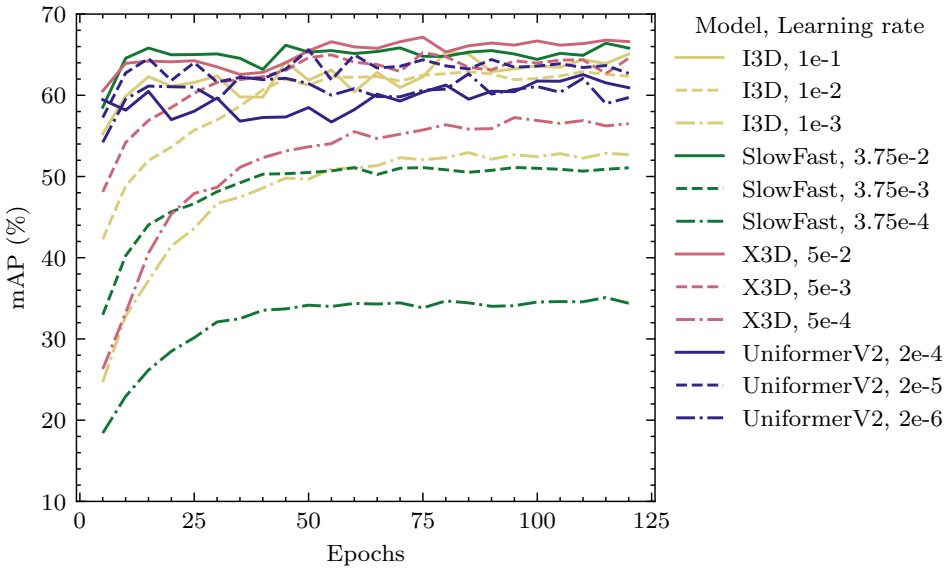

Fig. A1 Training history for I3D, SlowFast, X3D, UniformerV2 and using selected learning rates.

As shown in Table A1, the performance of various action recognition models is significantly influenced by the choice of learning rate. Across all models, the highest overall mAP is observed with the largest learning rate, with the exception of UniformerV2. SlowFast achieves a mAP score of 34.61% with a learning rate of 3.75×10^{-4} which increases significantly to 66.07% when the learning rate is increased by a factor of 100 to 3.75×10^{-2}. In contrast, I3D and X3D show less variability across learning rates. The mAP for I3D improves from 53.10% to 64.33%, while X3D increases from

56.39% to 66.36% as the learning rate rises. UniformerV2 achieves an mAP of 60.30% with a learning rate of 2×10^{-6}, which increases to 64.57% when the learning rate is raised to 2×10^{-5}. In all models, with the exception of SlowFast, the performance of head classes is not significantly impacted, whereas a consistent decrease in tail class performance is observed.

I3D and SlowFast achieve the best overall, head, and tail class performance at the highest tested learning rate. The X3D model follows a similar trend, except the head class performance peaks at a lower learning rate of 5×10^{-3} (97.12%). The UniformerV2 model shows a slight deviation from this pattern. The highest overall (64.57%) and tail (54.69%) mAP is achieved with a learning rate of 2×10^{-5}, while the best head class performance (50.37%) is observed at the higher rate of 2×10^{-4}.

Table A1 Learning rate analysis. Results of (a) I3D, (b) SlowFast, (c) X3D, and (d) UniformerV2 using various learning rates.

(a) I3D				**(b) SlowFast**			
Learning rate	**mAP (%)**			**Learning rate**	**mAP (%)**		
	Overall	Head	Tail		Overall	Head	Tail
1×10^{-1}	**64.33**	**96.58**	**53.58**	3.75×10^{-2}	**66.07**	**96.74**	**55.85**
1×10^{-2}	62.92	96.42	51.75	3.75×10^{-3}	50.56	95.69	35.52
1×10^{-3}	53.10	96.42	38.66	3.75×10^{-4}	34.61	84.84	17.87
(c) X3D				**(d) UniformerV2**			
5×10^{-2}	**66.36**	96.91	**56.17**	2×10^{-4}	61.85	**95.35**	50.68
5×10^{-3}	64.46	**97.12**	53.58	2×10^{-5}	**64.57**	94.21	**54.69**
5×10^{-4}	61.75	95.12	50.62	2×10^{-6}	60.30	94.70	48.83

As shown in Figure A1 the most significant fluctuations in model performance are observed for UniformerV2 and I3D models, particularly at the highest learning rates of 2×10^{-4} and 1×10^{-1}, respectively. In contrast, computationally smaller models like X3D and SlowFast exhibit less pronounced fluctuations.

Appendix B Dataset format

The proposed KABR dataset adopts the dataset format introduced by [10] in the Charades dataset. It is structured into two main directories: “images” and “annotation”. The “images” directory contains subdirectories for each video. Each video subdirectory stores the individual frames of the video as sequentially numbered image files. For example, ”video_1” includes files such as “image_1.jpg”, “image_2.jpg”, ..., “image_n.jpg”. This structure is repeated for all videos, where each video is stored in its subdirectory (“video_2”, ..., “video_n”), allowing for easy access to the frames of each video individually. The “annotation” directory contains metadata and annotations necessary for training and evaluating models. The file “classes.json” lists the behavior classes to be recognized in the dataset. The annotations for the training and validation sets are stored in “train.csv” and “val.csv” respectively, which link the

image sequences to the corresponding class labels. Following this format, KABR maintains a clear and scalable organization of the video data and annotations, consistent with the Charades dataset format.

References

[1] Mackintosh, N.J.: Animal Learning and Cognition. Academic Press, San Diego, California (2013)

[2] Greggor, A.L., Blumstein, D.T., Wong, B., Berger-Tal, O.: Using animal behavior in conservation management: a series of systematic reviews and maps. Springer (2019)

[3] Snowdon, C.T.: Animal signals, music and emotional well-being. Animals **11**(9), 2670 (2021)

[4] Kline, J., Stewart, C., Berger-Wolf, T., Ramirez, M., Stevens, S., Babu, R.R., Banerji, N., Sheets, A., Balasubramaniam, S., Campolongo, E., *et al.*: A framework for autonomic computing for in situ imageomics. In: 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS), pp. 11–16 (2023). IEEE

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

[6] Duporge, I., Kholiavchenko, M., Harel, R., Wolf, S., Rubenstein, D., Crofoot, M., Berger-Wolf, T., Lee, S., Barreau, J., Kline, J., Ramirez, M., Stewart, C.: Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos. arXiv preprint arxiv:2405.17698 (2024)

[7] Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arxiv 2012. arXiv preprint arXiv:1212.0402 (2012)

[8] Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19023–19034 (2022)

[9] Kholiavchenko, M., Kline, J., Ramirez, M., Stevens, S., Sheets, A., Babu, R., Banerji, N., Campolongo, E., Thompson, M., Van Tiel, N., *et al.*: Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 31–40 (2024)

[10] Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The

Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 510–526 (2016). Springer

[11] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)

[12] Cao, L., Tian, Y., Liu, Z., Yao, B., Zhang, Z., Huang, T.S.: Action detection using multiple spatial-temporal interest point features. In: 2010 IEEE International Conference on Multimedia and Expo, pp. 340–345 (2010). IEEE

[13] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. arXiv preprint arXiv:1808.01340 (2018)

[14] Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)

[15] Li, A., Thotakuri, M., Ross, D.A., Carreira, J., Vostrikov, A., Zisserman, A.: The ava-kinetics localized human actions video dataset. arXiv preprint arXiv:2005.00214 (2020)

[16] Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1996–2003 (2009). IEEE

[17] Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Machine vision and applications **24**(5), 971–981 (2013)

[18] Chen, J., Hu, M., Coker, D.J., Berumen, M.L., Costelloe, B., Beery, S., Rohrbach, A., Elhoseiny, M.: Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13052–13061 (2023)

[19] Yang, Y., Yang, J., Xu, Y., Zhang, J., Lan, L., Tao, D.: Apt-36k: A large-scale benchmark for animal pose estimation and tracking. Advances in Neural Information Processing Systems **35**, 17301–17313 (2022)

[20] Koger, B., Deshpande, A., Kerby, J.T., Graving, J.M., Costelloe, B.R., Couzin, I.D.: Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision. Journal of Animal Ecology (2023)

[21] Inoue, S., Yamamoto, S., Ringhofer, M., Mendonça, R.S., Pereira, C., Hirata, S.: Spatial positioning of individuals in a group of feral horses: A case study using drone technology. Mammal Research **64**, 249–259 (2019)

[22] Raoult, V., Tosetto, L., Williamson, J.E.: Drone-based high-resolution tracking

of aquatic vertebrates. Drones **2**(4), 37 (2018)

[23] Corcoran, E., Winsen, M., Sudholz, A., Hamilton, G.: Automated detection of wildlife using drones: Synthesis, opportunities and constraints. Methods in Ecology and Evolution **12**(6), 1103–1114 (2021)

[24] Blount, D., Gero, S., Van Oast, J., Parham, J., Kingen, C., Scheiner, B., Stere, T., Fisher, M., Minton, G., Khan, C., *et al.*: Flukebook: an open-source ai platform for cetacean photo identification. Mammalian Biology **102**(3), 1005–1023 (2022)

[25] Nepovinnykh, E., Eerola, T., Biard, V., Mutka, P., Niemi, M., Kunnasranta, M., Kälviäinen, H.: Sealid: Saimaa ringed seal re-identification dataset. Sensors **22**(19), 7602 (2022)

[26] Nepovinnykh, E., Chelak, I., Eerola, T., Immonen, V., Kälviäinen, H., Kholiavchenko, M., Stewart, C.V.: Species-agnostic patterned animal re-identification by aggregating deep local features. International Journal of Computer Vision, 1–16 (2024)

[27] Naik, H., Chan, A.H.H., Yang, J., Delacoux, M., Couzin, I.D., Kano, F., Nagy, M.: 3d-pop-an automated annotation approach to facilitate markerless 2d-3d tracking of freely moving birds with marker-based motion capture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21274–21284 (2023)

[28] Shao, H., Pu, J., Mu, J.: Pig-posture recognition based on computer vision: Dataset and exploration. Animals **11**(5), 1295 (2021)

[29] Giebenhain, S., Waldmann, U., Johannsen, O., Goldluecke, B.: Neural puppeteer: Keypoint-based neural rendering of dynamic shapes. In: Proceedings of the Asian Conference on Computer Vision, pp. 2830–2847 (2022)

[30] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8769–8778 (2018)

[31] Beery, S., Agarwal, A., Cole, E., Birodkar, V.: The iwildcam 2021 competition dataset. arXiv preprint arXiv:2105.03494 (2021)

[32] Bain, M., Nagrani, A., Schofield, D., Berdugo, S., Bessa, J., Owen, J., Hockings, K.J., Matsuzawa, T., Hayashi, M., Biro, D., *et al.*: Automated audiovisual behavior recognition in wild primates. Science Advances **7**(46), 4883 (2021)

[33] Ma, X., Kaufhold, S.P., Su, J., Zhu, W., Terwilliger, J., Meza, A., Zhu, Y., Rossano, F., Wang, Y.: Chimpact: A longitudinal dataset for understanding chimpanzee behaviors. arXiv preprint arXiv:2310.16447 (2023)

[34] Lei, Y., Dong, P., Guan, Y., Xiang, Y., Xie, M., Mu, J., Wang, Y., Ni, Q.: Postural behavior recognition of captive nocturnal animals based on deep learning: a case study of bengal slow loris. Scientific Reports **12**(1), 7738 (2022)

[35] Brookes, O., Mirmehdi, M., Stephens, C., Angedakin, S., Corogenes, K., Dowd, D., Dieguez, P., Hicks, T.C., Jones, S., Lee, K., et al.: Panaf20k: a large video dataset for wild ape detection and behaviour recognition. International Journal of Computer Vision, 1–17 (2024)

[36] Li, D., Chen, Y., Zhang, K., Li, Z.: Mounting behaviour recognition for pigs based on deep learning. Sensors **19**(22), 4924 (2019)

[37] Zhang, K., Li, D., Huang, J., Chen, Y.: Automated video behavior recognition of pigs using two-stream convolutional networks. Sensors **20**(4), 1085 (2020)

[38] Cowton, J., Kyriazakis, I., Bacardit, J.: Automated individual pig localisation, tracking and behaviour metric extraction using deep learning. IEEE Access **7**, 108049–108060 (2019)

[39] Jiang, M., Rao, Y., Zhang, J., Shen, Y.: Automatic behavior recognition of group-housed goats using deep learning. Computers and Electronics in Agriculture **177**, 105706 (2020)

[40] Nguyen, C., Wang, D., Von Richter, K., Valencia, P., Alvarenga, F.A., Bishop–Hurley, G.: Video-based cattle identification and action recognition. In: 2021 Digital Image Computing: Techniques and Applications (DICTA), pp. 01–05 (2021). IEEE

[41] Zia, A., Sharma, R., Arablouei, R., Bishop-Hurley, G., McNally, J., Bagnall, N., Rolland, V., Kusy, B., Petersson, L., Ingham, A.: Cvb: A video dataset of cattle visual behaviors. arXiv preprint arXiv:2305.16555 (2023)

[42] Rogers, M., Gendron, G., Valdez, D.A.S., Azhar, M., Chen, Y., Heidari, S., Perelini, C., O'Leary, P., Knowles, K., Tait, I., et al.: Meerkat behaviour recognition dataset. arXiv preprint arXiv:2306.11326 (2023)

[43] Iwashita, Y., Takamine, A., Kurazume, R., Ryoo, M.S.: First-person animal activity recognition from egocentric videos. In: 2014 22nd International Conference on Pattern Recognition, pp. 4310–4315 (2014). IEEE

[44] Feng, L., Zhao, Y., Sun, Y., Zhao, W., Tang, J.: Action recognition using a spatial-temporal network for wild felines. Animals **11**(2), 485 (2021)

[45] Geuther, B.Q., Peer, A., He, H., Sabnis, G., Philip, V.M., Kumar, V.: Action detection using a neural network elucidates the genetics of mouse grooming behavior. Elife **10**, 63207 (2021)

[46] Jhuang, H., Garrote, E., Yu, X., Khilnani, V., Poggio, T., Steele, A.D., Serre, T.: Automated home-cage behavioural phenotyping of mice. Nature communications **1**(1), 68 (2010)

[47] Burgos-Artizzu, X.P., Dollár, P., Lin, D., Anderson, D.J., Perona, P.: Social behavior recognition in continuous video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1322–1329 (2012). IEEE

[48] Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J.J., Perona, P., Anderson, D.J., Kennedy, A.: The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. Elife **10**, 63720 (2021)

[49] Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics. https://github.com/ultralytics/ultralytics

[50] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3464–3468 (2016). IEEE

[51] Kholiavchenko, M.: KABR Tools. https://doi.org/10.5281/zenodo.11288084 . https://doi.org/10.5281/zenodo.11288084

[52] Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., TOsmanov, Kruchinin, D., Zankevich, A., DmitriySidnev, Markelov, M., Johannes222, Chenuet, M., a-andre, telenachos, Melnikov, A., Kim, J., Ilouz, L., Glazov, N., Priya4607, Tehrani, R., Jeong, S., Skubriev, V., Yonekura, S., truong, zliang7, lizhming, Truong, T.: Opencv/cvat: V1.1.0. https://doi.org/10.5281/zenodo.4009388 . https://doi.org/10.5281/zenodo.4009388

[53] Starr, S., Williams, J.: The long tail: a usage analysis of pre-1993 print biomedical journal literature. Journal of the Medical Library Association: JMLA **96**(1), 20 (2008)

[54] Perrett, T., Sinha, S., Burghardt, T., Mirmehdi, M., Damen, D.: Use your head: Improving long-tail video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2415–2425 (2023)

[55] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)

[56] Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)

[57] Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 203–213 (2020)

[58] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y.: UniFormerV2: Spatiotemporal Learning by Arming Image ViTs with Video UniFormer (2022)

[59] Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11662–11671 (2020)

[60] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)

[61] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

[62] Duporge, I., Spiegel, M.P., Thomson, E.R., Chapman, T., Lamberth, C., Pond, C., Macdonald, D.W., Wang, T., Klinck, H.: Determination of optimal flight altitude to minimise acoustic drone disturbance to wildlife using species audiograms. Methods in Ecology and Evolution **12**(11), 2196–2207 (2021)