Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding (Supplementary Materials)

Xun Long Ng‡ Kian Eng Ong‡ Qichen Zheng‡ Yun Ni‡ Si Yong Yeo Jun Liu*
Information Systems Technology and Design, Singapore University of Technology and Design, Singapore
{xunlong_ng, kianeng_ong}@mymail.sutd.edu.sg {qichen_zheng, ni_yun, siyong_yeo, jun_liu}@sutd.edu.sg

1. Details of Data Collection and Verification

The team consisted of 23 members, including biology experts with knowledge of biodiversity.

We manually identify and provide framewise annotations of both animal and action descriptions for over 50 hours of videos that were collected from YouTube videos. We process them into a total of 30,100 video clips, each ranging from 1 second to 117 seconds (average 6 seconds). The title, description or captions in the videos contains the name of the animals. We manually check and tally the animals in the video clips with the animal names provided. As for the actions, we identify them using a defined set of ethological terms [1, 2, 4, 5]. In order to minimize discrepancies in our annotations, we first identify the commonly used terms that describe the same type of action in different classes of animals (*e.g.*, grooming in insects, but preening in birds refers to the same act of self-maintenance), and redefine ambiguous terms that describe vastly different movements (*e.g.*, snake gliding on land versus eagle gliding in the sky). At the same time, we also include descriptions of the scenes, and note their start and end times for video grounding task. We conduct a total of three rounds of quality checks through various permutations of cross-checks by different individuals to verify the correct start and end time of the video clips for video grounding task, and harmonize the nomenclature of actions for action recognition task.

As for pose estimation task, we label a total of 33K images that are extracted from the video clips, using Label Studio [6]. Annotators are given instructions with reference images on how to label each class of animals. As the animal footages are captured in the wild, the complex backgrounds with various illumination and weather conditions make the annotation challenging. Thus, three rounds of quality checks are performed to ensure that the keypoints are correctly labelled.

2. Diverse Range of Animals

Our dataset contains over 850 species of animals. We group them into 6 major animal classes, and further divide them into sub-classes. The distribution of animals in our dataset is shown in Fig. 1 and examples are shown in Fig. 2.

For action recognition, we label the actions of animals in 6 major classes (*i.e.*, mammals, reptiles, amphibians, birds, fishes, insects). For pose estimation, we label the poses of animals in 5 major classes (*i.e.*, mammals, reptiles, amphibians, birds, fishes). Because different animals can have vastly different anatomical structures across classes (*e.g.*, insects vs mammals), which poses a great challenge for pose annotation, we thus label these five animal classes which share similarities in their anatomical structures.

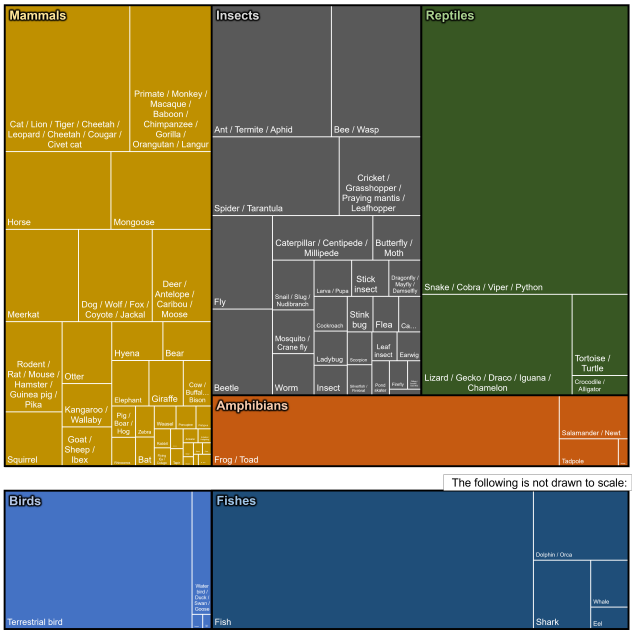

Figure 1. Distribution of clips of over 850 species of animals in 6 major animal classes classified based on their appearance, number of limbs and how they move, and further divided into sub-classes.

*Corresponding author.

‡Equal contribution to this work.

Figure 2. Examples of animals in 6 main animal classes

3. Diverse Range of Actions

Our dataset contains a diverse range of 140 actions. The collection of actions and behaviors encompasses:

(1) movement, *e.g.*, swimming, running, flying,
(2) transport, *e.g.*, carrying in mouth,
(3) feeding, *e.g.*, eating, biting, drinking,
(4) sensing, *e.g.*, exploring, attending,
(5) resting, *e.g.*, sleeping,
(6) maintenance, *e.g.*, grooming, washing,
(7) communication, *e.g.*, chirping,
(8) aggressive, *e.g.*, attacking, spitting venom,
(9) defensive, *e.g.*, retreating, displaying defensive pose,
(10) social, *e.g.*, playing,
(11) affection, *e.g.*, hugging,
(12) sexual, *e.g.*, sexual display, copulation,
(13) life events, *e.g.*, giving birth, laying eggs, hatching,
(14) other general actions, *e.g.*, panting, flapping ears.

Hence, our dataset covers a broad range of actions seen in nature.

4. More Examples

Here, we provide more examples of our video grounding (Fig. 3), action recognition (Fig. 4), and pose estimation (Fig. 5) tasks. The distribution of animals with its pose annotated is illustrated in Fig. 6.

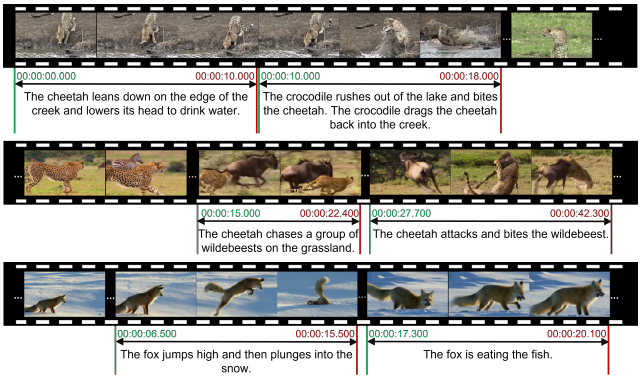

Figure 3. Samples of the video grounding task. Given the language description, we need to detect the corresponding time sequence.

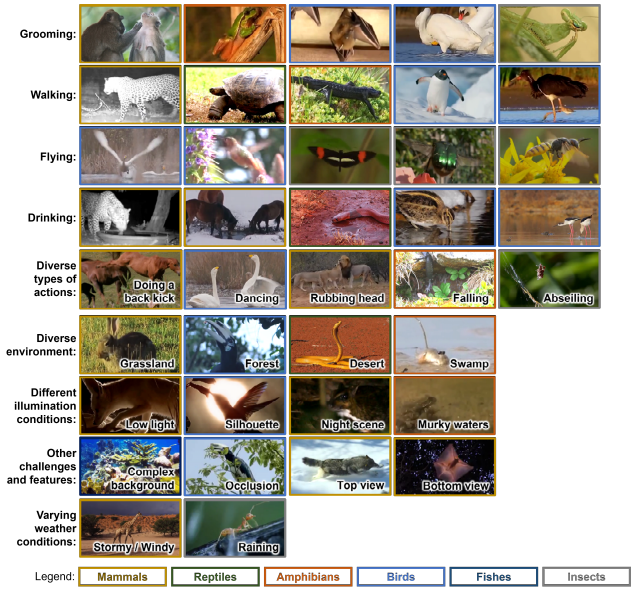

Figure 4. More examples of actions. Rows 1 to 4 show how the same set of actions differ across various animal classes. Row 5 shows examples of various actions in our dataset. Rows 6 to 9 show sample features of our dataset, ranging from diverse environments to varying illumination and weather conditions.

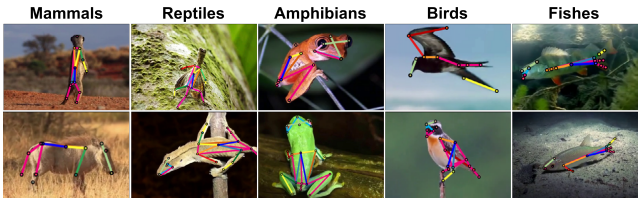

Figure 5. Examples of animal poses in our dataset

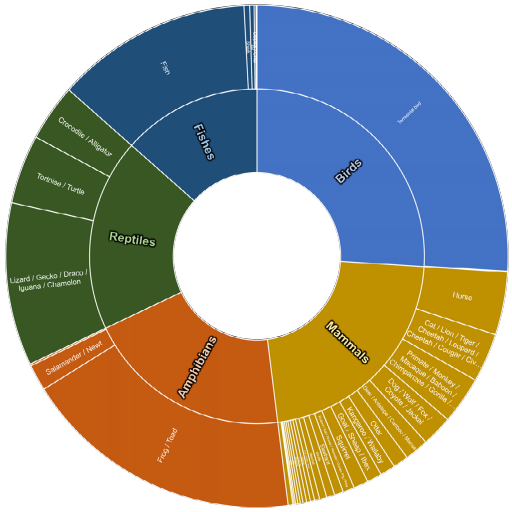

Figure 6. Distribution of the 33K animal pose annotations in the 5 major animal classes

5. Details of the Proposed CARe Model

In this section, we present the implementation details of our CARe model. We discuss in detail the architecture of the model (Fig. 7) and how it is trained (Algorithm 1).

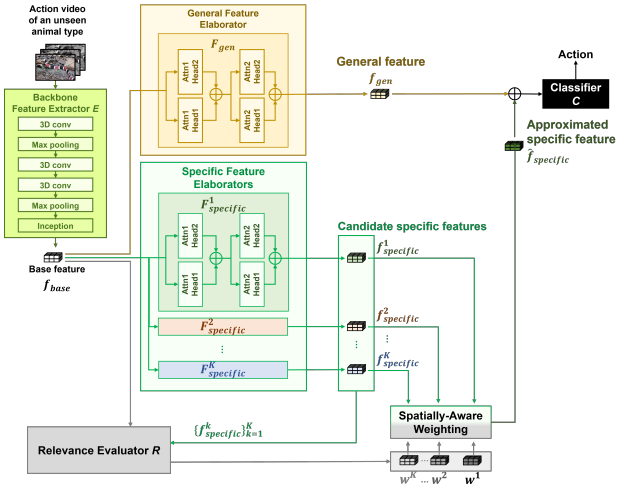

Figure 7. Architecture of our Collaborative Action Recognition (CARe) model.

The backbone feature extractor E, the feature elaborators F_{gen} and $\{F^k_{specific}\}^K_{k=1}$, and the classifier C form the basic version of our model. Below we present the design of each of them.

For the backbone feature extractor E, we adopt the early part of the I3D architecture. It includes three 3D convolutional layers E_conv1, E_conv2, and E_conv3, two max-pooling layers $E\_maxpool1$ and $E\_maxpool2$, and an inception submodule of the I3D architecture E_incept. Given the input video $x \in \mathbb{R}^{ch \times t \times h \times w}$ (whereby the number of channels ch, number of frames t, height h, width w equals 3, 16, 180, and 320 respectively), it will be transformed by E_conv1, $E\_maxpool1$, E_conv2, E_conv3, $E\_maxpool2$, and E_incept sequentially to obtain the base feature $f_{base} \in \mathbb{R}^{ch' \times t' \times h' \times w'}$, with $ch' = 256$, $t' = 8$, $h' = 23$, and $w' = 40$.

Regarding the subsequent feature elaborators $\{F^k_{specific}\}^K_{k=1}$ and F_{gen}, each of them consists of two lightweight two-head self-attention layers [3]. Given the base feature, the specific features $\{f^k_{specific}\}^K_{k=1}$ and general feature f_{gen} computed by the feature elaborators have the size of $ch'' \times h'' \times w''$, where ch'', h'', and w'' equal 4, 12, and 20.

The final classifier C consists of one linear layer C_linear that maps the flattened and concatenated general and specific features to action likelihoods $y \in \mathbb{R}^M$ for the M possible actions.

Besides components of the basic version of the CARe model, the relevance evaluator R is included to compute similarity scores for unseen animals. It consists of two layers, R_conv1 and R_conv2, which will be used to further transform the base feature, and K layers $\{R\_linear^k\}^K_{k=1}$ to compute the set of similarity scores $\{w^k\}^K_{k=1}$. More specifically, f_{base} will be first transformed by R_conv1 and R_conv2. To obtain the similarity score for the k-th specific feature elaborator w^k, we flatten and concatenate the transformed base feature and the specific feature $f^k_{specific}$, which will be fed into its respective layer R_linear^k to compute w^k.

Training and testing. Since the base feature extractor follows the I3D architecture, we initialize its parameters with the pre-trained I3D. The parameters of the base feature extractor will be optimized with an initial learning rate of 0.001. Other components of our CARe model will be randomly initialized and given a higher initial learning rate of 0.01. SGD optimizers are used to update all components. A summary of the training scheme can be found in Algorithm 1.

Algorithm 1: Training Procedures of CARe Model

Input : Data of K animal types $\{D_k\}^K_{k=1}$
Learning rates α, β
Hyper-parameter λ
Output: Backbone feature extractor E
General feature elaborator F_{gen}
Specific feature elaborators $\{F^k_{specific}\}^K_{k=1}$
Classifier C
Relevance evaluator R

while *not converged* **do**
  1. Update E, F_{gen}, $\{F^k_{specific}\}^K_{k=1}$, C:
  Calculate the cross-entropy losses $\{\ell^k\}^K_{k=1}$ for $\{D^k\}^K_{k=1}$ using E, F_{gen}, their respective $\{F^k_{specific}\}^K_{k=1}$, and C;
  Update the parameters for E, F_{gen}, $\{F^k_{specific}\}^K_{k=1}$ and C;
  2. Update R**:**
  Sample 1 type of animal as the meta-test D_{mtest} and the other K-1 types of animal as meta-train D_{mtrain};
  2.1 Meta-train:
  Calculate cross-entropy losses ℓ_{mtrain} for D_{mtrain};
  Update the parameters for R using:
  $\phi' \leftarrow \phi - \alpha \nabla_\phi \ell_{mtrain}(\phi)$;
  2.2 Meta-test:
  Calculate the cross-entropy loss ℓ_{mtest} for D_{mtest};
  2.3 Meta update:
  Update the parameters ϕ of R using:
  $\phi \leftarrow \phi - \beta((1-\lambda)\nabla_\phi \ell_{mtrain}(\phi) + \lambda \nabla_\phi \ell_{mtest}(\phi'))$;
end

In our experiments, 16 frames from each video sequence are randomly selected and used as the input. The model was trained for 40 epochs with a 2-GPU implementation. The initial learning rates were used for the first 30 epochs. The reduced learning rates that are 10% of the initial ones were used to train the model for 10 additional epochs.

References

[1] National Centre for the Replacement Refinement and Reduction of Animals in Research. General ethograms. Accessed November 11, 2021 [Online]. Available: https://nc3rs.org.uk/sites/default/files/documents/EvaluatingEnvironmentalEnrichment/General%20ethograms.pdf. 1

[2] Joseph Garner. Ethogram comparison. Accessed November 11, 2021 [Online]. Available: https://mousebehavior.org/ethogram-comparison/. 1

[3] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, pages 244–253, 2019. 3

[4] Paul E Rose and Lisa M Riley. Conducting behavioural research in the zoo: A guide to ten important methods, concepts and theories. *Journal of Zoological and Botanical Gardens*, 2(3):421–444, 2021. 1

[5] Richard Stafford, Anne E Goodenough, Kathy Slater, William S Carpenter, Laura Collins, Heather Cruickshank, Sarah Downing, Sally Hall, Katie McDonald, Heather McDonnell, et al. Inferential and visual analysis of ethogram data using multivariate techniques. *Animal Behaviour*, 83(2):563–569, 2011. 1

[6] Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2021. Open source software available from https://github.com/heartexlabs/label-studio. 1