Yang et al. *Zool. Res.* 2023, 44(6): 1026−1038
https://doi.org/10.24272/j.issn.2095-8137.2022.485

ZR Zoological Research

Article Open Access

BARN: Behavior-Aware Relation Network for multi-label behavior detection in socially housed macaques

Sen Yang[1,2,3], Zhi-Yuan Chen[2,3], Ke-Wei Liang[2,3], Cai-Jie Qin[2,3], Yang Yang[2,3], Wen-Xuan Fan[1,2,3], Chen-Lu Jie[2,3,4], Xi-Bo Ma[1,2,3,*]

[1] *College of Medicine and Biological Information Engineering*, *Northeastern University*, *Shenyang*, *Liaoning* 110169, *China*
[2] *MAIS*, *State Key Laboratory of Multimodal Artificial Intelligence Systems*, *Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*
[3] *School of Artificial Intelligence*, *University of Chinese Academy of Sciences*, *Beijing*, 100049, *China*
[4] *School of Automation*, *Harbin University of Science and Technology*, *Harbin*, *Heilongjiang* 150080, *China*

ABSTRACT

Quantification of behaviors in macaques provides crucial support for various scientific disciplines, including pharmacology, neuroscience, and ethology. Despite recent advancements in the analysis of macaque behavior, research on multi-label behavior detection in socially housed macaques, including consideration of interactions among them, remains scarce. Given the lack of relevant approaches and datasets, we developed the Behavior-Aware Relation Network (BARN) for multi-label behavior detection of socially housed macaques. Our approach models the relationship of behavioral similarity between macaques, guided by a behavior-aware module and novel behavior classifier, which is suitable for multi-label classification. We also constructed a behavior dataset of rhesus macaques using ordinary RGB cameras mounted outside their cages. The dataset included 65 913 labels for 19 behaviors and 60 367 proposals, including identities and locations of the macaques. Experimental results showed that BARN significantly improved the baseline SlowFast network and outperformed existing relation networks. In conclusion, we successfully achieved multi-label behavior detection of socially housed macaques with both economic efficiency and high accuracy.

Keywords: Macaque behavior; Drug safety assessment; Multi-label behavior detection; Behavioral similarity; Relation network

INTRODUCTION

The study of macaque behavior is crucial in many scientific domains, including pharmacology, yielding valuable data for drug safety assessment, cocaine abuse medications (Wit, 2011), stimulants (Volkow, 2012), sedatives (Kops et al., 2021), and anti-inflammatory medications (Singh et al., 1996), as well as the fields of neuroscience, psychology (Bala et al., 2020), and ethology (Defler, 2000). The process of monitoring and analyzing macaque behavior is instrumental in enhancing our understanding of their health and specific requirements (Bala et al., 2020). However, traditional monitoring methods, which primarily rely on manual observation, are labor-intensive and prone to inaccuracies (Kim et al., 2017). The emergence of video-based methods has provided an automated alternative, facilitating the quantitative prediction of behaviors (Liu et al., 2022), with approaches that focus on either individual or multiple animals. Given their social nature (Ballesta et al., 2014), behavioral analysis of socially housed macaques has attracted increasing attention (Bala et al., 2020; Ballesta et al., 2014).

Interactions among socially housed macaques can influence individual behaviors, necessitating an understanding of the behavior of the target macaque in the context of others (Morimoto & Fujita, 2011; Yu, 2016). Thus, a model accounting for the relationships between socially housed macaques is required (Sun et al., 2018). In recent years, several indirect behavior recognition approaches have been proposed, including detecting the position of macaques using traditional image processing (Liu et al., 2022) and pose estimation (Bala et al., 2020; Negrete et al., 2021; Li et al., 2023), with social interactions then determined based on inter-macaque distance. In addition, Marks et al. (2022) developed SIPEC:BehaveNet, a direct behavior recognition network that classifies three types of social interaction from videos without using pose estimation. However, these techniques fall short in capturing the relationships among macaques and in classifying multiple concurrent behaviors within individual macaques, such as eating while walking, a crucial aspect

This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright ©2023 Editorial Office of Zoological Research, Kunming Institute of Zoology, Chinese Academy of Sciences

Received: 23 March 2023; Accepted: 11 September 2023; Online: 12 September 2023

Foundation items: This work was supported by the Major Project of the National Natural Science Foundation of China (82090051, 81871442) and Outstanding Member Project of Youth Innovation Promotion Association of the Chinese Academy of Sciences (Y201930)

*Corresponding author, E-mail: xibo.ma@nlpr.ia.ac.cn

reflecting the true habits and characteristics of macaques (Glander, 1975; Röder & Timmermans, 2002). When dealing with complex relationships between target classes or a large number of classes, a multitude of different behavior combinations may emerge. In such cases, multi-label methods are typically employed to classify simultaneous classes rather than utilizing single-label techniques (Gu et al., 2018). To date, however, such approaches remain poorly studied.

Multi-label behavior detection, involving the localization, identification, and classification of individual or multiple simultaneous behaviors within a group (Zhang et al., 2019), presents a complex challenge in high-level video recognition (Pan et al., 2021). For application in macaques, existing multi-label behavior detection methods face two main issues. Firstly, due to the complexity of real-world scenarios, multi-label behavior detection in humans usually involves modeling relationships between humans and contextual objects, including first-order (human-human, human-context) and higher-order relationships (human-context-human), which require multiple modeling steps, large models with millions of trainable parameters (Pan et al., 2021), and large-scale datasets (e.g., AVA dataset with 1.6 million behavior labels) (Gu et al., 2018). However, a corresponding dataset for macaques is not yet available and would be very expensive to create. Furthermore, training large models on small-scale datasets may lead to over-fitting (Xu et al., 2019). Secondly, most recent methods typically generate only one output feature (Zhang et al., 2021), converted into behavior predictions using the Sigmoid activation function (Holman, 1948). Although they provide multi-label predictions, these approaches do not offer specific designs for multi-label classification, such as the coexistence of certain behaviors and the exclusion of others.

To address the above issues, we developed the Behavior-Aware Relation Network (BARN), which primarily consists of the Behavior-Aware Module (BAM) and two-branched Behavioral Similarity Reasoning Module (BSRM). First, an original dataset is reorganized into a behavior-aware dataset by dividing the behavior labels into three categories: i.e., behaviors with Apparent Displacement (AD), behaviors without AD (NAD), and foraging behaviors. The behavior-aware dataset contains only AD and NAD behaviors (which cannot co-exist), but not foraging behaviors (which can coexist with both AD and NAD behaviors). BAM is then pre-trained on the reorganized dataset to acquire prior behavioral knowledge. In the training stage, the proposed network first generates the proposals (identities and bounding boxes) and three-dimensional (3D) features of all monkeys using the backbone network SlowFast (Feichtenhofer et al., 2019) and ROI Align (He et al., 2017). In parallel, BAM generates two types of behavior-specific information pertaining to AD and NAD behaviors, which are combined with the bounding boxes to form behavior guidance data. The obtained 3D features of all monkeys and behavior guidance data are then taken as BSRM inputs to model the similarity of AD or NAD behaviors between macaques. The behavior guidance data enables BSRM to focus on the simple but important relationship of behavioral similarity, thus improving network performance. Finally, the proposed network employs a novel behavior classifier to generate three different output features, which is more suitable for multi-label behavior prediction.

To evaluate BARN performance, we established a macaque behavior dataset using several ordinary RGB cameras mounted outside the monkey cages. The dataset contained the daily-life records of socially housed macaques, as well as data annotations, including 65 913 labels for 19 macaque behaviors and 60 367 proposals. We conducted extensive experiments on the proposed dataset. Results showed that BARN achieved significant improvements to the baseline network of SlowFast (Feichtenhofer et al., 2019) and outperformed existing relation networks.

The contributions of our research include the following: (1) We accomplished multi-label behavior detection of socially housed macaques for the first time using the proposed BARN model. (2) We designed BSRM to model simple but important relationships of behavioral similarity among macaques. (3) We proposed a dataset reconstruction approach, and designed BAM to acquire prior behavioral knowledge from the reconstructed dataset. (4) We designed a novel behavior classifier based on the characteristics of behaviors and their various combinations, which was more suitable for multi-label classification.

MATERIALS AND METHODS

Data acquisition

This study complied with international standards for the care and use of non-human primates and was approved by the Animal Management and Use Committee of Joinn Laboratories (China) Co., Ltd. (B-ACU22-H-NHP-001). All cages (length 1.4 m×width 1.1 m×height 2.2 m) were equipped with a viewing platform, two shelves for perching, a water pipe, an externally welded food trough, and several regularly changed toys. A total of 20 male macaques (aged 3–5 years) were divided into four groups and placed in the cages. The concrete floor of the cage was covered with regularly renewed wood shavings to encourage foraging. Given the spatial constraints of a single camera covering the entire cage, three HIK VISION DS-2CD3T47EWDV3-L cameras (1 920×1 080 pixels, 30 FPS, 4 mm focal length) were used. The overall cage environment and camera setup are shown in Figure 1. Cameras #1 and #3 were positioned near the cage to preclude any obstruction by the fence in front of the camera lenses. As the macaques frequently occupied the viewing platform within the enclosures, capturing images of this specific area was essential for observations. To mitigate any interference by the metal railings, Camera #2 was located proximal to the monkey cage, although this camera arrangement posed challenges in capturing the entire viewing platform. Upon careful consideration, the distance between Camera #2 and the viewing platform was set to 15 cm.

For automatic monitoring of macaque behavior, individual identification of each monkey is necessary (Marks et al., 2022). Given that macaques possess thick fur and frequently exit the frame, researchers can employ sensors, jackets, and neck collars to identify monkeys. However, sensors are often prohibitive in cost (Meunier et al., 2018) and jackets can induce significant perturbations in behavior (Bala et al., 2020). Therefore, in the current study, we used different colored neck collars to distinguish the macaques. The obtained videos were categorized by the corresponding monkey cage number, camera number, and timestamp.

Data annotation

As the same behaviors can convey different information when performed in different locations (Defler, 2000), differentiation

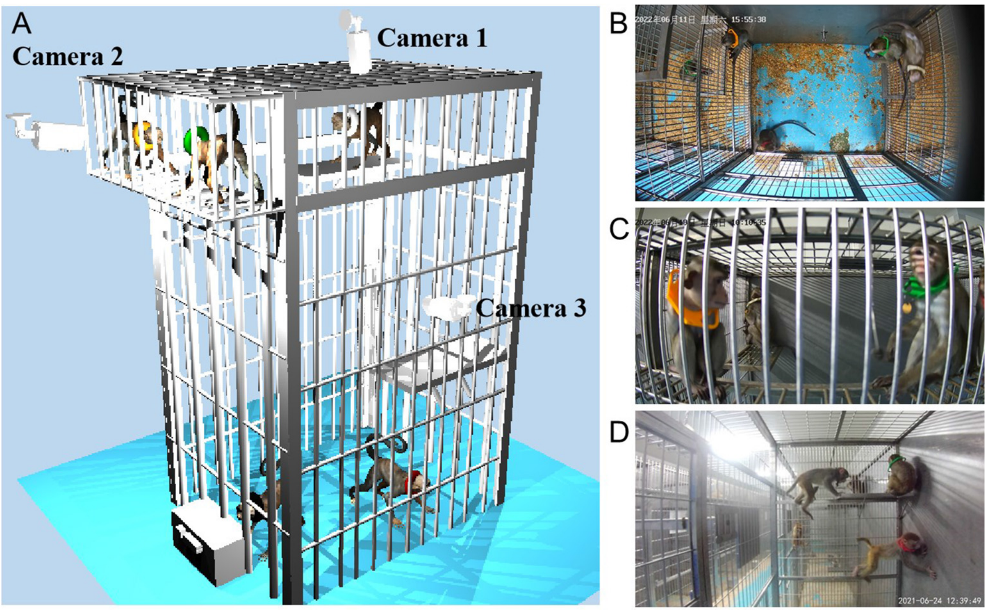

Figure 1 Overall cage environment and camera setup

A: Perspective view of overall environment. Cameras #1 and #3 were placed close to the cage. The distance of Camera #2 to the viewing platform was set to 15 cm. B–D: One video frame recorded by Cameras #1, #2, and #3, respectively.

between these behaviors, such as sitting (high) and sitting (ground), is required. Based on previous research (Marks et al., 2022) and visual analysis of macaque behaviors (Defler, 2000; Westlund et al., 2012), we defined 19 daily behaviors. In addition, five identity labels were also defined based on neck collar colors, including yellow, green, red, black, and white (numbered 0 to 4, respectively). Specific descriptions of each behavior are provided in Table 1.

Due to the long-tail distribution of macaque behaviors, there are fewer "tail" behaviors compared to "head" behaviors in the original videos over the same period. This imbalance may negatively affect behavior detection. To amplify the frequency of "tail" behaviors, the original videos were processed into behavior-intensive video clips of different lengths in the time dimension. Importantly, the model has the capacity to handle original videos of any length automatically.

Jumping was identified as the fastest behavior in the obtained video clips, with a duration of 0.3 to 0.5 s. As a result, one key frame was selected from every 10 video frames, with all target monkeys in the frames then annotated with bounding boxes, identities, and behavior labels using atomic annotation (Gu et al., 2018). For example, if a monkey wearing a green neck collar was observed eating while walking, the bounding box and three labels, including green, eating, and walking, were recorded.

The transition from one behavioral pattern to another manifests as a "transitional behavior", a phenomenon that presents challenges in precise definition and discrimination. In the present study, transitional behavior was identified at the midpoint of the corresponding "transition movement". For example, changing from "sitting" to "bipedal standing" was conceptualized as the transition of the angle between the monkey's thigh and calf from 0° to 180°, with the transition movement thus approximately corresponding to 90°. During the process of data annotation, two specific difficulties were encountered. First, within a single camera view, macaques may be positioned in close proximity to one another, leading to the obscuration of some individuals. Second, spatial constraints related to the housing environment and the wide angle of the camera lens presented difficulties in capturing the entirety of the monkey cage, culminating in a "dead corner" within the field of view of the cameras. Consequently, macaques could enter and exit the camera view, leading to complexities in capturing all behaviors. In contrast to previous studies where these issues remain unaddressed (Ballesta et al., 2014; Marks et al., 2022), we implemented two methods to mitigate these challenges.

Method one: During the recording process, the macaques were detected by different cameras at the same time. Thus, although a macaque may have been blocked from one perspective, it was discernible from others. Therefore, data annotators made the final judgment on the identities and behaviors of the macaques by referencing videos from various perspectives with congruent timestamps. Resolution of this problem was facilitated by the fact that each monkey was equipped with a uniquely colored collar. Thus, the identity and behaviors of the macaques were determined using video timestamps and colored collars.

Method two: A common feature of both identified problems was the reduction in body area of the macaque that could be captured by the camera. Therefore, a threshold was established, whereby the data annotator refrained from labeling the monkey if its body area within view was less than one-third of the original size, and neither its head nor collar was visible.

We formed a special data annotation team of 10 people, with the data annotators divided into two groups. Detailed marking specifications were developed and the VOTT labeling platform (Wbreza, 2021), a public annotation tool, was used for tagging. Upon completion of annotation, two distinct steps were undertaken to improve annotation quality.

Step one: Codes were employed to search for low-level errors, including: (1) instances where a data annotator may label a bounding box with two or more identity tags, despite a one-to-one correspondence rule between bounding boxes and identity tags; (2) instances where a data annotator may label a bounding box with multiple incompatible behaviors, such as "climbing" and "jumping"; and (3) instances where a data

Table 1 Descriptions of macaque behaviors based on visual analysis

Number	Category	Description of behavior
0	Sitting (high)	Sitting on shelf or lookout
1	Sitting (ground)	Sitting on ground
2	Prostrating (high)	Prostrating on shelf or lookout
3	Prostrating (ground)	Prostrating on ground
4	Quadrupedal standing	All limbs are straight, supporting the body
5	Bipedal standing	Hind legs are straight, supporting the body
6	Hang (arm)	Suspending body with forelimbs
7	Hang upside down	Suspending body with hind legs
8	Attaching	Holding and leaning against cage in mid air
9	Walking	Going somewhere on foot
10	Climbing	Holding the cage and going somewhere
11	Jumping	Moving quickly by pushing body with limbs
12	Eating	Holding food and chewing it anywhere
13	Grasping food	Grabbing food from the trough
14	Drinking	Biting the water pipe and drinking
15	Fighting	Fighting with each other
16	Chasing	Monkey running after another monkey
17	Grooming	Cleaning the fur of another monkey or itself
18	Others	Other behaviors, not described above

annotator may only label a bounding box, neglecting to assign an identity or behavior tag.

Step two: After labels were incorporated into the videos, the two data annotator groups conducted reciprocal checks. Any labels identified as ambiguous or suspicious were earmarked for further discussion.

Finally, the identity labels and bounding boxes were employed to create an identity dataset, whereas the behavior labels and bounding boxes were utilized to assemble a behavior dataset.

In total, 19 daily behaviors were classified into three main behavioral categories based on the coexistence (or not) of certain behaviors, i.e., foraging behaviors, behaviors with Apparent Displacement (AD), and behaviors without AD (NAD). Initially, foraging behaviors were classified as eating, grasping food, and drinking, while the remaining 16 behaviors were classified as non-foraging behaviors. Subsequently, based on analysis of significant differences in movement distance, the non-foraging behaviors were divided into NAD and AD behaviors. Typical NAD behaviors included sitting (high), prostrating (high), and quadrupedal standing, while typical AD behaviors included walking, climbing, jumping, and chasing. The NAD and AD behaviors were mutually exclusive categories, whereas foraging behaviors were capable of coexisting with both. Figure 2 provides a detailed illustration of each behavior and its corresponding classification into one of principal categories.

We next reorganized the behavior dataset into a behavior-aware dataset. In detail, the labels corresponding to NAD behaviors were recorded as "NAD" and the labels corresponding to AD behaviors were recorded as "AD". Labels associated with foraging behaviors were not used in the behavior-aware dataset.

In the final dataset, 21 642 key frames were extracted from 228 video clips of varying duration. These data annotations contained 60 367 identity labels and bounding boxes, 65 913 behavior labels, and 58 964 behavior-aware labels. The above-mentioned labels and bounding boxes were selected as grounding-truth. The number of samples in the three datasets is shown in Figure 3.

Partition of datasets

The dataset was divided into three parts according to the sequence of video timestamps, i.e., training set (earliest part), validation set (middle part), and test set (latest part). Due to the uneven distribution of behaviors within the subsets, e.g., the "Hang upside down" behavior occurred more often in the test set than in the training set, certain videos were adjusted to create similar behavior distributions across the different subsets. Ultimately, 21 642 key frames were split into 13 785 training, 3 706 validation, and 4 151 test key frames, resulting in 41 231 training, 11 482 validation, and 13 200 test bounding boxes with identity and behavior labels (see Supplementary Materials for number of labels in different subsets).

Overall framework of BARN

BARN was designed to generate places (bounding boxes), identities, and multi-label behavior predictions of all macaques in an input video clip (16 continuous video frames in our experiments). As shown in Figure 4, the proposed network contains three primary modules: i.e., the Feature Extraction Module (FEM), Behavior-Aware Module (BAM), and Behavioral Similarity Reasoning Module (BSRM), the latter of which is the key module. Specifically, the network initially employs the FEM and detector to generate the proposals of each monkey, including identities and bounding boxes, along with the 3D features of all monkeys (monkey features). Simultaneously, after training on the behavior-aware dataset, the BAM generates two types of behavior-specific information, which are combined with the bounding boxes to yield behavior guidance information. The BSRM then models the similarity in NAD and AD behaviors between different macaques using both the monkey features and behavior guidance information. The resultant behavior correlation features are converted into three different output features by a novel behavior classifier for final behavior predictions.

The main network modules are described in detail as follows:

(1) Input video frames: Although the input of the proposed model is set to 16 video frames, the model can automatically detect original videos of any length by sequential sampling.

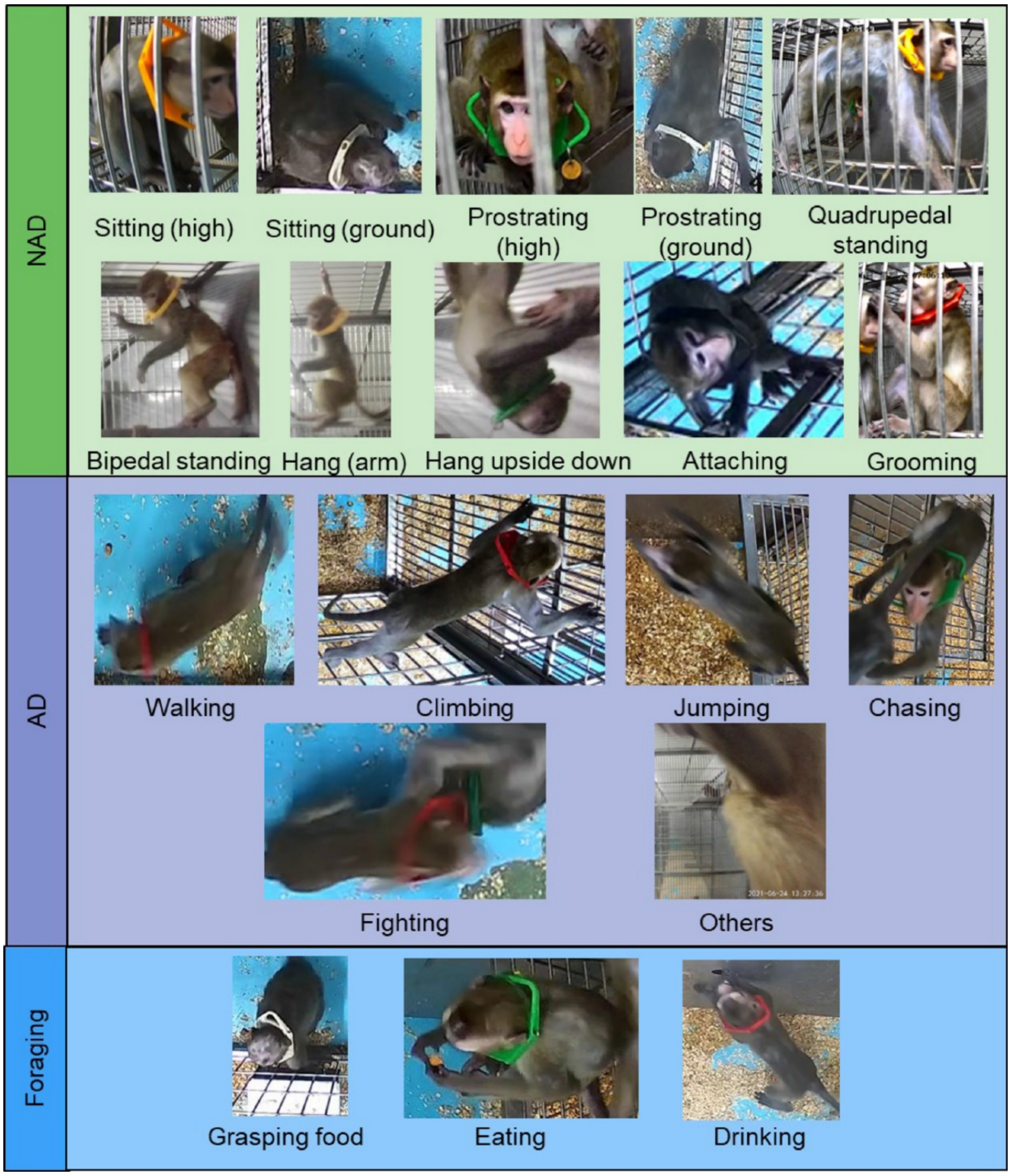

Figure 2 Example of each behavior and corresponding category

NAD, behaviors without apparent displacement. AD, behaviors with apparent displacement. Foraging behaviors can occur simultaneously with other behaviors, while NAD and AD behaviors are mutually exclusive.

Specifically, the original videos and natural behaviors within them are divided into several 16 frame segments automatically and then detected sequentially by the network. If the length of the segment is less than 16 frames, the last frame is copied to ensure a length of 16 frames. In addition, within the study datasets, a frame taken from a single camera view may contain zero to five macaques. We used three macaques as an example in Figure 4.

(2) Monkey detector: The YOLO v7 network (Wang et al., 2022) is employed to generate the proposal for each monkey in each input video frame. The network is pretrained on the proposed identity dataset to detect all monkeys within a single frame. When training the whole network, detection of the monkeys is initially carried out on the middle frame of the input video (e.g., ninth frame of the 16 video frames). The obtained proposals are then duplicated to neighboring frames of the middle frame. The detected bounding boxes (place information) of N monkeys are recorded as $P \in R^{N \times 4}$. It is worth noting that identity information is not used in BARN. Therefore, even if the identity prediction results are incorrect, the prediction for behavior is not influenced.

(3) Feature Extraction Module (FEM): FEM employs the standard SlowFast backbone network (Feichtenhofer et al., 2019) to extract spatiotemporal features (3D features) of input video frames. The backbone network includes both slow and fast pathway. The former uses a 3D convolutional neural network (3D-CNN) at a low frame rate to capture spatial semantics, while the latter operates at a high frame rate to capture motion at a fine temporal resolution. The outputs of the two pathways are then fused into 3D features. Specifically, the SlowFast R-50 4×16 instantiation, pretrained on the AVA v2.1 dataset (Gu et al., 2018), serves as the backbone network. To reduce subsequent calculation costs, the FEM performs average pooling in the time dimension. The obtained 3D features and place information are then converted into monkey features, $M \in R^{N \times C \times H \times W}$, using ROI Align (He et al.,

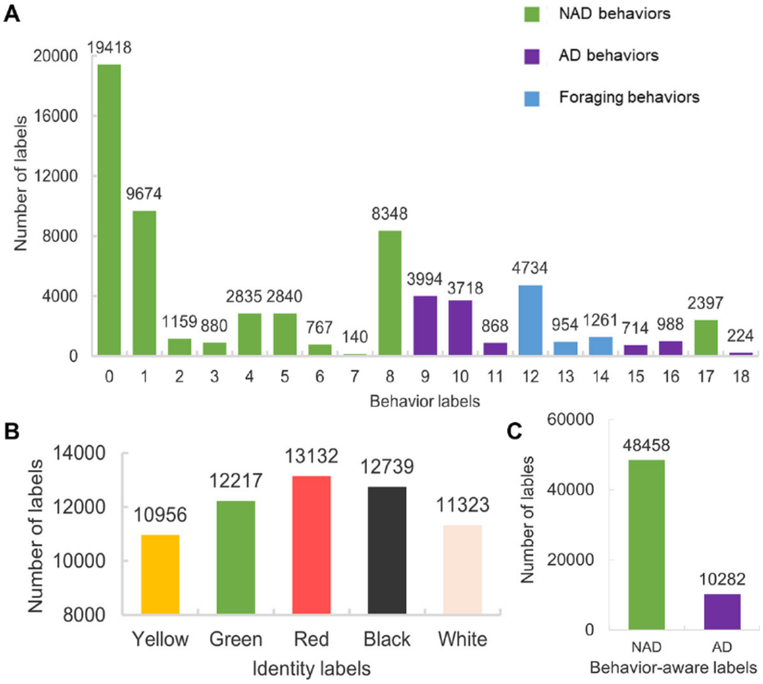

Figure 3 Number of samples in different datasets

A: Behavior dataset. B: Identity dataset. C: Behavior-aware dataset. Behavior dataset contains 19 behaviors and follows long-tail distribution. Identity dataset contains five colored labels, with corresponding to a monkey with a collar of the same color. Behavior-aware dataset is derived from the behavior dataset and contains only two classes. NAD, behaviors without apparent displacement. AD, behaviors with apparent displacement.

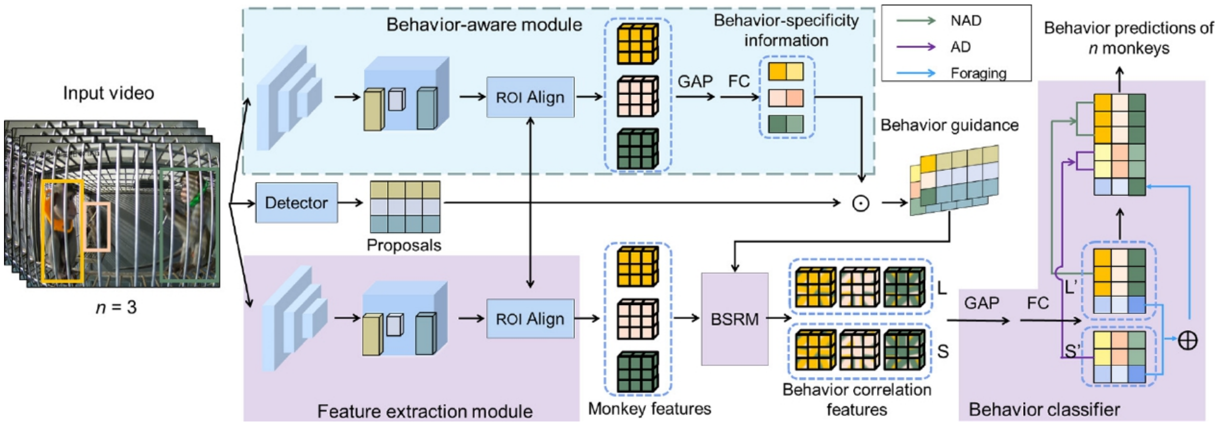

Figure 4 Overview of proposed muti-label behavior detection framework

Videos including several macaques (such as *n*=3) are processed with the FEM to produce spatiotemporal context features. For each macaque proposal (bounding box), monkey features are extracted from the context features by ROI Align. Given the monkey features and behavior guidance from the BAM and proposals, the BSRM models behavioral similarity between macaques. The two obtained behavior correlation features are converted into final behavior predictions for each macaque by the behavior classifier. NAD, behaviors without apparent displacement. AD, behaviors with apparent displacement. GAP, global average pool. FC, fully connected layer. BSRM, Behavioral Similarity Reasoning Module.

2017), where N, C, H, and W represent number of monkeys, channel, height, and width, respectively.

(4) Behavior-Aware Module (BAM): The BAM is designed to acquire prior knowledge on NAD and AD behaviors from the reorganized behavior-aware dataset. As shown in Figure 4, the structure of BAM is similar to that of FEM, except that BAM contains a global average pool (GAP) layer and a fully connected layer (FC) with an output dimension of 2. The output of BAM is recorded as behavior-specificity information $D \in R^{N\times2}$, including $D_l \in R^{N\times1}$ and $D_s \in R^{N\times1}$.

To distinguish between NAD and AD behaviors more effectively, BAM weights are pretrained on the behavior-aware dataset. The learned weights are then used to initialize BAM when training the whole network.

(5) Behavior guidance information: As macaque behaviors are related to their location (Defler, 2000), monkey detector place information is also regarded as important prior knowledge. As shown in Figure 4, place information P is combined with the behavior-specificity information (D_l, D_s) into a prior knowledge of behavior guidance information $G \in R^{2\times N\times5}$, including $G_l \in R^{N\times5}$ and $G_s \in R^{N\times5}$. The process can be expressed as:

$$\begin{aligned} G_l &= concatenate(D_l, P) \\ G_s &= concatenate(D_s, P) \\ G &= \{G_l, G_s\} \end{aligned} \tag{1}$$

(6) Behavioral Similarity Reasoning Module (BSRM): The architecture of BSRM is shown in Figure 5. BSRM uses two

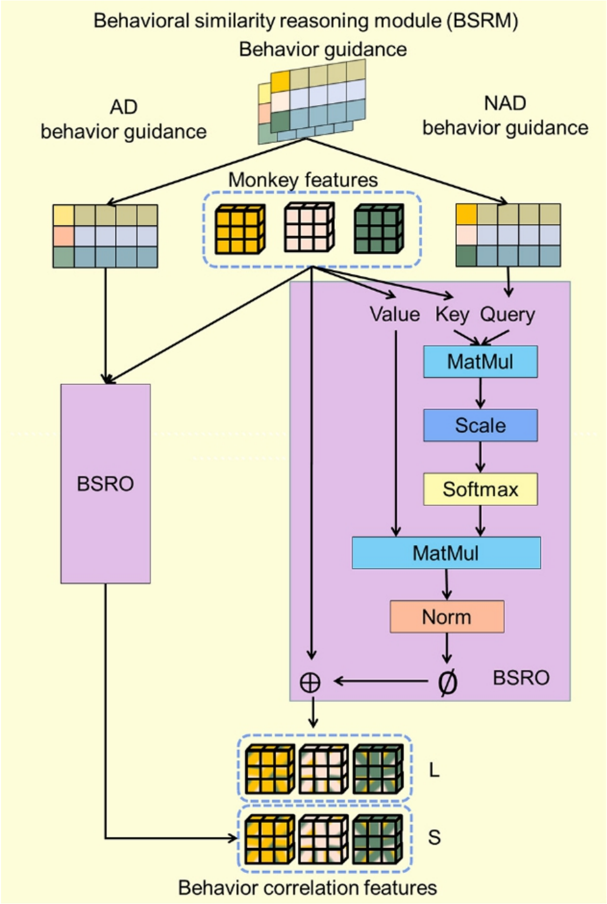

Figure 5 Architecture of BSRM

Monkey features and two types of behavior guidance information are processed with different BSROs to two behavior correlation features. BSRO uses monkey features to produce tensors Key and Value and uses behavior guidance to produce tensor Query. Point multiplication is used to interact different tensors in the MatMul layer. NAD, behaviors without apparent displacement. AD, behaviors with apparent displacement. BSRO, behavioral similarity reasoning operation.

kinds of behavior guidance information (G_l, G_s) and monkey features (M) as inputs, and builds two branches to generate two behavior correlation features (L, S), with each branch performing behavioral similarity reasoning operation (BSRO).

BSRO is inspired by the mechanism of primate emotional contagion (Morimoto & Fujita, 2011). Notably, perception of another's emotional expression can automatically evoke the same emotion in the perceiver, inducing potential performance of similar behaviors (Morimoto & Fujita, 2011). For instance, one-day-old human neonates may begin to cry when they hear the cry of another (Morimoto & Fujita, 2011). Utilizing this conceptual framework, BSRO can generate features sensitive to certain behaviors using the behavior guidance information. As shown in Figure 5, the monkey features and NAD behavior guidance information can be converted into behavior correlation features L.

Based on Pan et al. (2021), BSRO uses convolution to generate three tensors, Query (Q), Key (K), and Value (V), with the following innovations: (1) While Pan et al. (2021) employed a combination of human and context features as inputs to model relationships between humans and contextual objects, BSRO excludes context features from the network to avoid network emphasis on secondary (monkey-context) and complex (monkey-context-monkey) relationships among socially housed macaques. (2) In contrast to Pan et al. (2021), where the generation of the three tensors (Q, K and V) was based solely on certain features, BSRO utilizes the NAD behavior guidance information to generate Q and uses monkey features to generate K and V.

Specifically, monkey features (M) are converted to the tensors K and V by convolution. The NAD behavior guidance information is initially resized to the dimensions of the monkey features using the unsqueeze and repeat functions within the Torch software library (Collobert et al., 2002), and subsequently converted into tensor Q. The operation can be described as follows:

$$\begin{gathered} K, V = conv2D(M) \\ Q = conv2D(repeat(unsqueeze(G_l))) \end{gathered} \tag{2}$$

where $M, Q, K, V \in R^{N \times C \times H \times W}$ and $G_l \in R^{N \times 5}$. BSRO then adjusts the size of the three tensors using the *unsqueeze* function. The obtained tensors are converted into attention vectors Att and behavioral similarity features Z based on the relational reasoning operation of HR^{2}O (Pan et al., 2021). It is worth noting that point multiplication is used for the interaction between different tensors. The operation can be described by:

$$\begin{gathered} Q^{'}, K^{'}, V^{'} = unsqueeze(Q, K, V) \\ Att = unsqueeze\left(soft\max\left(sum\left(\frac{Q^{'} \cdot K^{'}}{\sqrt{d}}\right)\right)\right) \\ Z = sum\left(Att \cdot V^{'}\right) \end{gathered} \tag{3}$$

where $Q^{'} \in R^{N \times 1 \times C \times H \times W}$, $K^{'} \in R^{1 \times N \times C \times H \times W}$, $V^{'} \in R^{1 \times N \times C \times H \times W}$, $Att \in R^{N \times N \times 1 \times H \times W}$, and $Z \in R^{N \times C \times H \times W}$. Following Wu et al. (2019), the normalization layer, ReLU activation function, two-dimensional (2D) convolution, and dropout layer are then employed to generate the NAD behavior correlation features L. The process can be represented as:

$$\begin{gathered} Z^{'} = Dropout\,(conv2D\,(ReLU\,(norm\,(Z)))) \\ L = M + Z^{'} \end{gathered} \tag{4}$$

The process of generating AD behavior correlation features S is similar to that of NAD. Finally, the behavior correlation features are passed to the proposed behavior classifier.

(7) Behavior classifier: As foraging behaviors may coexist with NAD or AD behaviors, behavior correlation features (L and S) are also used to predict foraging behaviors. As shown in Figure 4 (right), the behavior correlation features are converted into three different output features using the GAP and FC layers, as follows:

$$\begin{gathered} L^{'} = \left\{L_l^{'}, L_f^{'}\right\} = FC_1(GAP(L)) \\ S^{'} = \left\{S_s^{'}, S_f^{'}\right\} = FC_2(GAP(S)) \\ O = \{O_l, O_s, O_f\} = \left\{L_l^{'}, S_s^{'}, L_f^{'} + S_f^{'}\right\} \end{gathered} \tag{5}$$

where $L_l^{'} \in R^{N \times 10}$ and $S_s^{'} \in R^{N \times 6}$ represent the predictions of 10 NAD behaviors and six AD behaviors of N monkeys in the input video frames, respectively. $L_f^{'} \in R^{N \times 3}$ and $S_f^{'} \in R^{N \times 3}$ are converted into predictions for the three foraging behaviors by an addition operation.

RESULTS

The proposed BARN necessitated the preliminary training of both the monkey detector and BAM, prior to implementation of

BARN itself. Thus, the monkey detector and BAM were trained in advance. Various comparative experiments were then conducted on the proposed macaque behavior dataset, followed by extensive ablation experiments to study the effects of the different BARN modules. Finally, BARN was used to analyze the behaviors of macaques to show the wide application foreground.

Training of monkey detector and BAM

Mean average precision (mAP) (Everingham et al., 2010) was selected as the evaluation metric in the object and multi-label behavior detection tasks. Specifically, given the predictions with different classification confidences and ground-truth, we computed precision and recall at different classification confidences of the model on each class and obtained the P-R curve, with the area under the P-R curve representing average precision (AP) of each class. The mAP of the model was then produced by averaging the AP of the model across all classes. The calculation process compared the prediction results of the model with the ground-truth and generated a quantitative evaluation result for model performance. We used mAP with an IoU threshold of 0.5 and a classification confidence threshold of 0.002 to evaluate the models.

For the monkey detector, we trained YOLO v7 on the proposed identity dataset using default setting. After 50 epochs, we achieved 93.1% precision, 91.1% recall, and 95.1% mAP on the validation set, and achieved 94.4% precision, 91.3% recall, and 95.1% mAP on the test set.

For evaluation of behavior detection, the area under the curve (AUC) metric was added, calculated similarly to mAP. The true positive rate (TPR) and false positive rate (FPR) were first calculated at different classification confidence thresholds of the model on each class to generate receiver operating characteristic (ROC) curves for each class. The AUC of the model was generated by averaging the areas under the ROC curves of all classes.

BAM was trained to classify two classes in single-label form on the behavior-aware dataset using synchronized stochastic gradient descent (SGD) across three NVIDIA TITAN RTX GPUs. The batch size was set to 16. BAM was trained end-to-end for 33k iterations with a base learning rate of 0.1125. During the first 4k iterations, linear warm-up was performed (Goyal et al., 2017) with a weight decay of 10^{-5} and Nesterov momentum of 0.9. The base learning rate was then multiplied by 0.1 at iterations 8k and 12k. After 30 epochs, BAM achieved 89.3% mAP and 63.3% AUC on the validation set and 77.8% mAP and 54.3% AUC on the test set. When training BARN, we initialized BAM with the weights learned on the behavior-aware dataset and froze the parameters of its backbone network. This maintained the module's perception of NAD and AD behaviors.

Comparison experiments on behavior dataset

In the context of limited existing methods for multi-label behavior detection in macaques, we applied state-of-the-art networks from the human AVA dataset (Gu et al., 2018) to the macaque behavior dataset. As detailed in Table 2, the "SlowFast" baseline network (Feichtenhofer et al., 2019) achieved a mAP of 58.8%. The "ACAR (no bank)" approach (Pan et al., 2021), as a high-order (monkey-context-monkey) relation network without a feature bank, was less effective than the baseline network. "ACRN" (Sun et al., 2018), as a first-order (monkey-monkey, monkey-context) relation network, showed an improvement of 0.5% in mAP over the baseline network. The proposed BARN, modeling similarities in NAD and AD behaviors among macaques, achieved the highest mAP of 64.3%. Similar results were achieved on the test set. These findings validate the effectiveness of the proposed model in modeling behavioral similarity. The "SlowOnly" network (Feichtenhofer et al., 2019), which is a lighter network than "SlowFast", also improved upon the baseline network, implying that lightweight backbone networks may be more suitable for small-scale behavior datasets.

In addition, AP values of the baseline network, ACRN, and BARN for each specific behavior were determined and displayed in Figure 6. The proposed BARN model achieved the highest AP for 13 behaviors (marked by red rectangles) and relatively comparable AP values for the other behaviors.

Ablation study

To verify the effectiveness of the proposed improved module, we conducted detailed ablation experiments on the behavior dataset. The experimental results are described below.

(1) Behavior-Aware Module: The BAM backbone was frozen, and the behavior-specificity information generated was used for subsequent BSRM. Ablation experiments were then performed to verify the effectiveness of these modifications. First, experiments were conducted to investigate the effect of freezing different layers of BAM on the results. As shown in Table 3, "Freeze-Non", "Freeze-Backbone", and "Freeze-All" indicate freezing no parameters, freezing only parameters of the backbone network, and freezing parameters of all layers during the training process, respectively. Results showed that freezing the backbone network of BAM achieved the best performance (64.3% val mAP, 88.1% val AUC, 60.1% test mAP, and 83.1% test AUC).

We subsequently analyzed the effects of behavior-specificity and place information on the performance of the proposed network. As shown in Table 4, "Guide-Non" signifies that neither behavior-specificity nor place information were used, and the tensor Query of BSRM was generated solely

Table 2 Comparison of BARN (ours) with state-of-the-art methods (e.g., SlowFast (Feichtenhofer et al., 2019), SlowOnly (Feichtenhofer et al., 2019), ACAR (no bank) (Pan et al., 2021), ACRN (Sun et al., 2018)) using AVA dataset

Model	SlowFast (baseline) (%)	SlowOnly (%)	ACAR (no bank) (%)	ACRN (%)	**BARN (ours)** (%)
Backbone	SlowFast	**SlowOnly**	SlowFast	SlowFast	SlowFast
Pre-train	AVA v2.1	AVA v2.1	AVA v2.1	AVA v2.1	AVA v2.1
Relational Relationing	No	No	High-Order	First-Order	**Behavioral Similarity**
Val mAP	58.8	59.4	57.8	59.3	**64.3**
Val AUC	85.1	86.3	83.8	87.2	**88.1**
Test mAP	45.8	52.6	46.1	47.7	**60.1**
Test AUC	79.0	81.9	78.5	80.6	**83.1**

Bold font represents best performance achieved in the experiments.

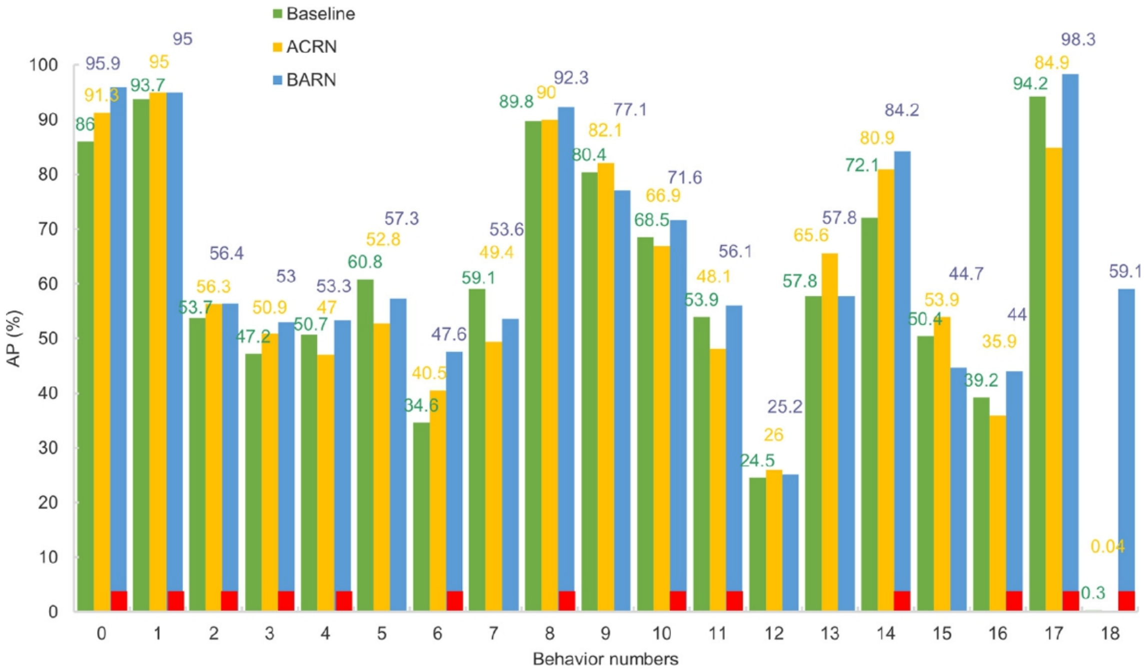

Figure 6 Comparative analysis of BARN (ours) with baseline network SlowFast and ACRN (Sun et al., 2018) on the validation set

Abscissa is number of each behavior; ordinate is average precision (AP) of each behavior. Red rectangle signifies that BARN achieved the highest AP for a given behavior.

Table 3 Ablation results after freezing different layers of BAM

Model	Freeze-Non (%)	**Freeze-Backbone** (%)	Freeze-All (%)
Pre-train	AVA v2.1	AVA v2.1	AVA v2.1
Val mAP	55.5	**64.3**	61.4
Val AUC	84.6	**88.1**	87.4
Test mAP	44.5	**60.1**	56.6
Test AUC	78.8	**83.1**	84.4

Bold font represents best performance achieved in the experiments.

Table 4 Ablation results based on behavior guidance information

Model	Guide-Non (%)	Guide-Bs (%)	Guide-Place (%)	**Guide-Bs&Place** (%)
Behavior Guidance	No	Behavior-Specificity	Place	**Both**
Val mAP	56.6	61.1	52.5	**64.3**
Val AUC	85.7	88.5	87.5	**88.1**
Test mAP	57.0	58.0	49.1	**60.1**
Test AUC	83.5	84.5	82.8	**83.1**

Guide-Non, no behavioral guidance. Guide-Bs, using behavior-specificity information as behavioral guidance. Guide-Place, using place information as behavioral guidance. Guide-Bs&Place, using both as behavioral guidance. Bold font represents best performance achieved in the experiments.

with monkey features. The terms "Guide-Bs" and "Guide-Place" indicate that the network applied behavioral specificity and place information as behavioral guidance information, respectively, while "Guide-Bs&Place" indicates that the network concatenated both into behavioral guidance information. Results showed that "Guide-Bs" achieved a higher mAP value than "Guide-Non", while "Guide-Bs&Place" achieved the highest performance overall (64.3% val mAP, 88.1% val AUC, 60.1% test mAP, and 83.1% test AUC). These outcomes support the hypothesis that merging both types of information is beneficial for network performance, playing complementary roles.

(2) Behavioral Similarity Reasoning Module: BARN modeled behavioral similarity based on the baseline network and achieved obvious improvements (see comparison experiments), proving the BSRM is beneficial for the network. Thus, we next investigated whether it was beneficial for generating three different output features. As seen in Table 5, "BSRM-NoGroup" represents the generation of one output feature using one-branched BSRM, where all behavior-specificity information ($D \in R^{N\times2}$) and place information ($P \in R^{N\times4}$) are concatenated into behavior guidance information ($G \in R^{N\times6}$). The term "BSRM-Group" represents the generation of three different output features using two-branched BSRM and the proposed behavior classifier. Results showed that "BSRM-Group" achieved a higher mAP than "BSRM-NoGroup", thereby proving the effectiveness of generating different output features.

(3) Behavior Classifier: The integration of various features is frequently achieved through addition and concatenation operations (Feichtenhofer et al., 2019), with the FC layer required for concatenation. In the current study, we conducted

Table 5 Ablation results after grouping operations in BSRM

Model	BSRM-NoGroup (%)	BSRM-Group (%)
Branches	One	**Two**
Val mAP	59.3	**64.3**
Val AUC	87.4	**88.1**
Test mAP	56.4	**60.1**
Test AUC	84.0	**83.1**

Bold font represents best performance achieved in the experiments.

several experiments on three distinct combinations of the methods for fusing prediction results of foraging behaviors $\{L'_{fi}\}_{i=1}^{3}$ and $\{S'_{fi}\}_{i=1}^{3}$. These three combinations, "Add", "Add-FC", and "Cat-FC", are represented by the following equations, respectively:

$$
\begin{aligned}
O_f &= \left\{L'_{fi} + S'_{fi}\right\}, i = 1,2,3 \\
O_f &= \left\{FC(L'_{fi} + S'_{fi})\right\}, i = 1,2,3 \\
O_f &= \left\{FC(cat(L'_{fi}, S'_{fi}))\right\}, i = 1,2,3
\end{aligned}
\tag{6}
$$

Table 6 shows the experimental results of the above three combinations. "Add" achieved the best performance and was thus taken as the default setting.

Application of monkey detector to analyze motion of socially housed macaques

Implementation of the proposed BARN required the employment of a monkey detector (YOLO V7 network) to generate bounding boxes. Here, to evaluate the application of the monkey detector, we determined the movement distance of the macaques using the monkey detector on the test set of the identity dataset. Specifically, the monkey detector was run with a classification confidence threshold of 0.8 on the test set to generate proposals of the macaques, achieving a mAP of 74.3%. Subsequent calculations were computed to determine the Euclidean distance traversed by the center points of the bounding boxes in adjacent key frames, as well as total movement distance of each macaque in the test set. Comparative analysis was performed between the generated results and ground-truth (see Materials and Methods for details). As illustrated in Figure 7, the movement distance generated by the monkey detector was lower than that of ground-truth, which may be due to occlusions and the convergence of motion within the field of view of a single camera.

Application of BARN to analyze behaviors of socially housed macaques

To evaluate the application of the proposed network on macaque behavior analysis, BARN was applied with a classification confidence threshold of 0.8 on the test set of the proposed macaque behavior dataset to generate behavior predictions. Behavior predictions with a classification confidence above 0.8 were regarded as correct classifications. The generated results were visualized in several videos. As seen in Figure 8, the bounding boxes and movement trajectories of the macaques are drawn based on their collar colors. The identities $\{e_i\}_{i=0}^{4}$ and behaviors $\{f_j\}_{j=0}^{18}$ of each macaque were drawn in white in the format $e_i\#f_{j1} - f_{j2}$ (Supplementary Videos S1, S2). The ground-truth proposals were used to evaluate the behavior predictions of BARN to enhance the visualization process.

Table 6 Ablation results based on application of behavior classifier

Model	Add (%)	Add-FC (%)	Cat-FC (%)
Fusing Method	Addition	Addition	Concatenation
FC	No	one FC	one FC
Val mAP	64.3	**64.9**	62.6
Val AUC	**88.1**	87.7	87.8
Test mAP	**60.1**	56.8	56.6
Test AUC	**83.1**	84.3	83.7

Bold font represents best performance achieved in the experiments.

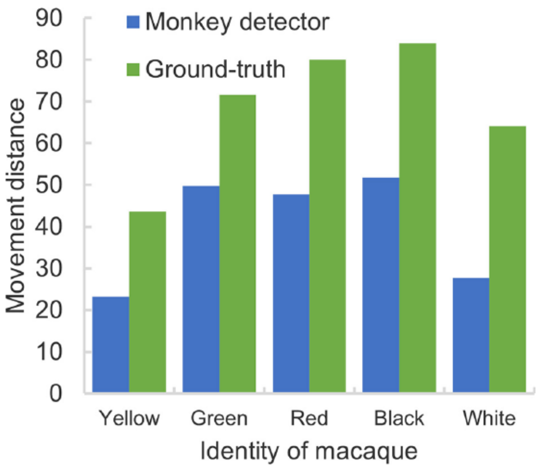

Figure 7 Movement distance of each macaque generated by ground-truth and monkey detector on the test set of the proposed macaque identity dataset

Test set contains a series of key frames at different camera views, with each frame containing zero to five macaques.

We then computed the duration of each behavior and compared the BARN prediction results with that of ground-truth and ACRN. As the interval duration between adjacent key frames was 0.33 s in the study, we multiplied the number of correctly classified behaviors by 0.33 to generate behavior duration. Figure 9A shows the duration of each behavior for yellow macaques, Figure 9B shows the total duration of all behaviors for yellow macaques, Figure 9C shows the total duration of all behaviors for all macaques, and Figure 9D shows the duration of each behavior for all macaques (see Supplementary Materials for the durations of other macaques).

DISCUSSION

In this study, we introduced the Behavior-Aware Relation Network (BARN), which was developed to detect the locations and identities of socially housed macaques and provide multi-label behavior predictions for each animal. The newly proposed network functions by the acquisition of prior behavioral knowledge through the reorganization of original behavior datasets, followed by the construction of simple but important relationships of behavioral similarity among macaques for final behavior predictions. To the best of our knowledge, the proposed network is the first model to successfully accomplish multi-label behavior detection in socially housed macaques.

The proposed network faces several limitations. First, BARN automatically splits the original videos into several 16 frame segments, enabling processing of videos of arbitrary length. However, during the segmentation process, the inherent

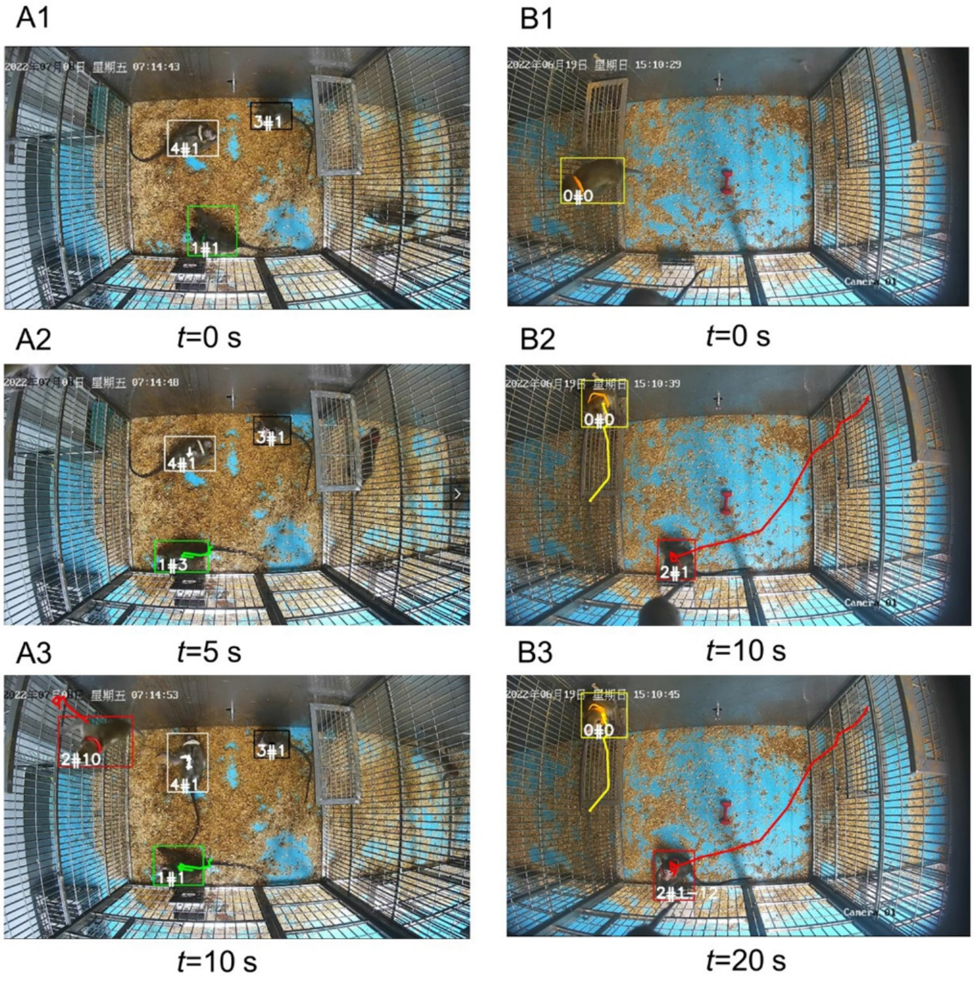

Figure 8 Visualization results of bounding boxes, identities, movement trajectories, and behaviors of macaques in two videos

Bounding boxes and movement trajectories of each macaque are drawn in the color of its collar. Identities and behaviors of macaques are separated by "#", and simultaneous behavior is separated by "-".

dependencies between the individual 16 frame segments and adjacent natural behaviors are not efficiently exploited. A feasible solution to this issue may be the incorporation of a long-term behavioral memory bank (Pan et al., 2021). In addition, application of an unsupervised time-warping method may also be beneficial for capturing natural behavior structural patterns (Han et al., 2022). Second, the efficiency of BARN is dependent upon the ability of the detector to generate the bounding boxes of the macaques. Although the YOLO v7 detector achieved 95.1% val and test mAP values on the identity dataset, its performance may be compromised under certain situations, such as during occlusions or the intersection of multiple macaques while performing opposite movements. Merging the prediction results from various camera perspectives, utilizing both the timestamp and collar color as synchronization parameters, may prove beneficial for circumventing the issue of occlusion, as macaques are unlikely to be obstructed in the view of every camera.

Our experimental results revealed a dependency of relation network performance on environmental complexity. In the human AVA dataset (Gu et al., 2018), modeling complex relationships yielded better results compared to modeling simple relationships, whereas the opposite trend was observed in our macaque dataset. This discrepancy may be attributed to the relative simplicity of the environment in our dataset compared to the AVA dataset. Given that BARN removes environmental information through the application of ROI Align, it may be more suitable for simple environments, such as laboratories and zoos. In contrast, in locations characterized by more complex environments, such as sanctuaries, the integration of monkey and context features within the BSRM framework may achieve better performance.

Behavior predictions generated by BARN hold potential for application in the analysis of macaque behavior. BARN operates by sequentially sampling 16 consecutive frames from videos of arbitrary length, then generating behavior predictions for each macaque in the input video frames end-to-end. This process allows the extraction of information about the behaviors of each macaque, start and end times of each behavior, and frequency of each behavior based on the timestamps of the input videos. The construction of an objective function may enable the derivation of other behavioral metrics of interest (Jafrasteh & Suárez, 2021). For behaviors that lack annotation, complementary unsupervised approaches may prove useful (Hsu & Yttri, 2019; Wiltschko et al., 2015). Moreover, the utilization of published model weights from this study may facilitate network initialization and training on new datasets targeting specific behaviors. Although BARN was trained using rhesus macaque datasets, its application to other monkey species, such as cynomolgus macaques, is also feasible. For monkeys with large differences in appearance from rhesus macaques, such as

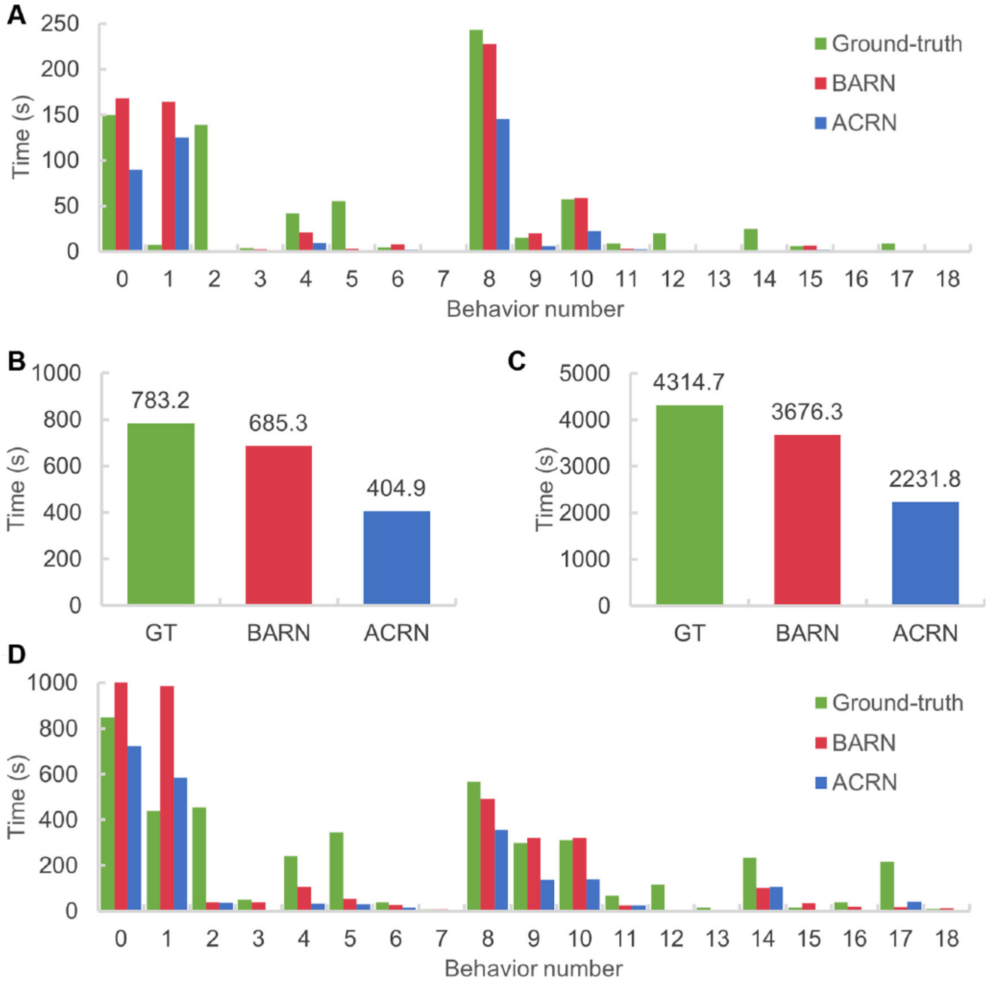

Figure 9 Duration of behaviors generated by ground-truth, BARN (ours), and ACRN (Sun et al., 2018) on the test set of the proposed macaque behavior dataset

A: Duration of each behavior for yellow macaques. B: Total duration of all behaviors for yellow macaques. C: Total duration of all behaviors for all macaques. D: Duration of each behavior for all macaques.

squirrel monkeys, employing the model weights published in this study as initial weights for pre-training may be an effective approach. Furthermore, high-level events can be obtained by combining the situations of all macaques.

In addition to behavior predictions and identities, the places and movement trajectories of each macaque can be generated. Places encompass 2D coordinates corresponding to the center point of the bounding box, and research that is interested in the movement trajectories of macaques can benefit from this network. It is important to note that places reflect whole-level movement rather than subtle movements of small body parts (Liu et al., 2022). Moreover, for 3D movement trajectories, one economical approach is to model the actual environment and map 2D positions into 3D space (Marks et al., 2022). Alternatively, using MouseVenue3D to generate 3D positions of markerless animals may be another viable approach (Han et al., 2022).

In the context of multi-label behavior detection of socially housed macaques, modeling the relationships among individual macaques is crucial. Existing methods generally model relationships between entities based on large models and large-scale annotated datasets and are therefore difficult to apply to macaques. To overcome these challenges, we developed the BARN model and a macaque behavior dataset. Experimental results demonstrated the effectiveness of the different modules and showed that the proposed network outperformed many state-of-art methods. Notably, BARN successfully accomplished multi-label behavior detection of socially housed macaques and can be easily used to analyze macaque behaviors.

DATA AVAILABILITY

Our code and macaque datasets are freely available online (https://github.com/BertonYang18/BARN-monkey) along with the publication of this study. We hope to receive feedback on any potential bugs or issues. Supplementary Videos S1 and S2 can also be found online (https://drive.google.com/drive/folders/1v2ZcXlrAR7rB0Pws4SWUqZPTKupVQIw7?usp=share_link).

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

S.Y. completed the main experiment, wrote the manuscript draft, and conducted the literature review. S.Y., Z.Y.C., K.W.L., W.X.F., and C.L.J. participated in the design of data acquisition and annotation scheme. X.B.M., Y.Y., and C.J.Q discussed the study, provided suggestions to improve the experimental scheme, and edited the manuscript. All authors read and approved the final version of the manuscript.

REFERENCES

Bala PC, Eisenreich BR, Yoo SBM, et al. 2020. Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. *Nature Communications*, **11**(1): 4560.

Ballesta S, Reymond G, Pozzobon M, et al. 2014. A real-time 3D video tracking system for monitoring primate groups. *Journal of Neuroscience Methods*, **234**: 147−152.

Collobert R, Bengio S, Mariéthoz J. 2002. Torch: a modular machine learning software library. REP_WORK (30 October, 2002). Idiap, https://os.unil.cloud.switch.ch/tind-customer-epfl/5ea06583-58ae-4f33-bcc5-ae8feb746af1?response-content-disposition=attachment%3B%20filename%2A%3DUTF-8%27%27rr02-46.pdf&response-content-type=application%2Fpdf&AWSAccessKeyId=ded3589a13b4450889b2f728d54861a6&Expires=1682421084&Signature=gej7yCqVtYuJiwgttsO0YPoiqXo%3D.

Defler TR. 2000. Locomotion and posture in *Lagothrix lagotricha*. *Folia Primatologica*, **70**(6): 313–327.

Everingham M, Van Gool L, Williams CKI, et al. 2010. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, **88**(2): 303−338.

Feichtenhofer C, Fan HQ, Malik J, et al. 2019. Slowfast networks for video recognition. *In*: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 6201–6210.

Glander KE. 1975. Habitat description and resource utilization: a preliminary report on mantled howling monkey ecology. *In*: Tuttle RH. Socioecology and Psychology of Primates. Berlin: De Gruyter Mouton*,* 37–58.

Goyal P, Dollár P, Girshick R, et al. 2017. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv*: 1706.02677.

Gu CH, Sun C, Ross DA, et al. 2018. AVA: a video dataset of spatio-temporally localized atomic visual actions. *In*: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 6047–6056.

Han YN, Huang K, Chen K, et al. 2022. MouseVenue3D: a markerless three-dimension behavioral tracking system for matching two-photon brain imaging in free-moving mice. *Neuroscience Bulletin*, **38**(3): 303−317.

He KM, Gkioxari G, Dollár P, et al. 2017. Mask R-CNN. *In*: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2980–2988.

Holman CC. 1948. Hæmangioma of the sigmoid colon. Report of a case. *British Journal of Surgery*, **36**(142): 210.

Hsu AI, Yttri EA. 2019. B-SOiD: an open source unsupervised algorithm for discovery of spontaneous behaviors. *BioRxiv*, 770271.

Jafrasteh B, Suárez A. 2021. Objective functions from Bayesian optimization to locate additional drillholes. *Computers & Geosciences*, **147**: 104674.

Kim NY, Kim SJ, Jang SY, et al. 2017. Behavioral characteristics of Hanwoo (*Bos taurus coreanae*) steers at different growth stages and seasons. *Asian-Australasian Journal of Animal Sciences*, **30**(10): 1486−1494.

Kops MS, Pesic M, Petersen KU, et al. 2021. Impact of concurrent remifentanil on the sedative effects of remimazolam, midazolam and propofol in cynomolgus monkeys. *European Journal of Pharmacology*, **890**: 173639.

Li C, Xiao Z, Li Y, et al. 2023. Deep learning-based activity recognition and fine motor identification using 2D skeletons of cynomolgus monkeys. *Zoological Research*, **44**(5): 967−980.

Liu MS, Gao JQ, Hu GY, et al. 2022. MonkeyTrail: a scalable video-based method for tracking macaque movement trajectory in daily living cages. *Zoological Research*, **43**(3): 343−351.

Marks M, Jin QH, Sturman O, et al. 2022. Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nature Machine Intelligence*, **4**(4): 331−340.

Meunier B, Pradel P, Sloth KH, et al. 2018. Image analysis to refine measurements of dairy cow behaviour from a real-time location system. *Biosystems engineering*, **173**: 32−44.

Morimoto Y, Fujita K. 2011. Capuchin monkeys (*Cebus apella*) modify their own behaviors according to a conspecific's emotional expressions. *Primates*, **52**(3): 279−286.

Negrete SB, Labuguen R, Matsumoto J, et al. 2021. Multiple monkey pose estimation using OpenPose. *bioRxiv:* 428726.

Pan JT, Chen SY, Shou MZ, et al. 2021. Actor-context-actor relation network for spatio-temporal action localization. *In*: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 464–474.

Röder EL, Timmermans PJA. 2002. Housing and care of monkeys and apes in laboratories: adaptations allowing essential species-specific behaviour. *Laboratory Animals*, **36**(3): 221−242.

Singh GB, Bani S, Singh S. 1996. Toxicity and safety evaluation of Boswellic acids. *Phytomedicine*, **3**(1): 87−90.

Sun C, Shrivastava A, Vondrick C, et al. 2018. Actor-centric relation network. *In*: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 335–351.

Volkow ND. 2012. Long-term safety of stimulant use for ADHD: findings from nonhuman primates. *Neuropsychopharmacology*, **37**(12): 2551−2552.

Wang CY, Bochkovskiy A, Liao HYM. 2022. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv*: 2207.02696.

Wbreza. 2021. VOTT.https://github.com/microsoft/VoTT.

Westlund K, Fernström AL, Wergård EM, et al. 2012. Physiological and behavioural stress responses in cynomolgus macaques (*Macaca fascicularis*) to noise associated with construction work. *Laboratory Animals*, **46**(1): 51−58.

Wiltschko AB, Johnson MJ, Iurilli G, et al. 2015. Mapping sub-second structure in mouse behavior. *Neuron*, **88**(6): 1121−1135.

Wit HD. 2011. Sex hormones: a new treatment for cocaine abuse?. *Neuropsychopharmacology*, **36**(11): 2155−2156.

Wu CY, Feichtenhofer C, Fan HQ, et al. 2019. Long-term feature banks for detailed video understanding. *In*: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 284–293.

Xu Q, Zhang M, Gu ZH, et al. 2019. Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs. *Neurocomputing*, **328**: 69−74.

Yu S. 2016. New challenge for bionics —brain-inspired computing. *Zoological Research*, **37**(5): 261−262.

Zhang YB, Tokmakov P, Hebert M, et al. 2019. A structured model for action detection. *In*: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 9967–9976.

Zhang YY, Li XY, Marsic I. 2021. Multi-label activity recognition using activity-specific features and activity correlations. *In*: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 14620–14630.