Li et al. *Zool. Res.* 2023, 44(5): 967−980
https://doi.org/10.24272/j.issn.2095-8137.2022.449

ZR | Zoological Research

Article Open Access

Deep learning-based activity recognition and fine motor identification using 2D skeletons of cynomolgus monkeys

Chuxi Li[1,#], Zifan Xiao[2,#], Yerong Li[1], Zhinan Chen[1], Xun Ji[3], Yiqun Liu[4], Shufei Feng[5], Zhen Zhang[5], Kaiming Zhang[6], Jianfeng Feng[2], Trevor W. Robbins[2,7], Shisheng Xiong[1,*], Yongchang Chen[5], Xiao Xiao[2,*]

[1] *School of Information Science and Technology Micro Nano System Center*, *Fudan University*, *Shanghai* 200433, *China*
[2] *Department of Anesthesiology*, *Huashan Hospital*; *Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence*, *Ministry of Education*; *Behavioral and Cognitive Neuroscience Center*, *Institute of Science and Technology for Brain-Inspired Intelligence*, *MOE Frontiers Center for Brain Science*, *Fudan University*, *Shanghai* 200433, *China*
[3] *Kuang Yaming Honors School*, *Nanjing University*, *Nanjing*, *Jiangsu* 210023, *China*
[4] *Shanghai Key Laboratory of Intelligent Information Processing*, *School of Computer Science*, *Fudan University*, *Shanghai* 200433, *China*
[5] *State Key Laboratory of Primate Biomedical Research*; *Institute of Primate Translational Medicine*, *Kunming University of Science and Technology*, *Kunming*, *Yunnan* 650500, *China*
[6] *New Vision World LLC.*, *Aliso Viejo*, *California* 92656, *USA*
[7] *Behavioural and Clinical Neuroscience Institute*, *University of Cambridge*, *Cambridge*, *CB2 1TN*, *UK*

ABSTRACT

Video-based action recognition is becoming a vital tool in clinical research and neuroscientific study for disorder detection and prediction. However, action recognition currently used in non-human primate (NHP) research relies heavily on intense manual labor and lacks standardized assessment. In this work, we established two standard benchmark datasets of NHPs in the laboratory: MonkeyinLab (MiL), which includes 13 categories of actions and postures, and MiL2D, which includes sequences of two-dimensional (2D) skeleton features. Furthermore, based on recent methodological advances in deep learning and skeleton visualization, we introduced the MonkeyMonitorKit (MonKit) toolbox for automatic action recognition, posture estimation, and identification of fine motor activity in monkeys. Using the datasets and MonKit, we evaluated the daily behaviors of wild-type cynomolgus monkeys within their home cages and experimental environments and compared these observations with the behaviors exhibited by cynomolgus monkeys possessing mutations in the *MECP2* gene as a disease model of Rett syndrome (RTT). MonKit was used to assess motor function, stereotyped behaviors, and depressive phenotypes, with the outcomes compared with human manual detection. MonKit established consistent criteria for identifying behavior in NHPs with high accuracy and efficiency, thus providing a novel and comprehensive tool for assessing phenotypic behavior in monkeys.

Keywords: Action recognition; Fine motor identification; Two-stream deep model; 2D skeleton; Non-human primates; Rett syndrome.

INTRODUCTION

Action recognition and object phenotype recognition are critical skills essential for human survival and evolutionary progress, with a profound connection to cognitive function, emotional expression, and social communication. Within the field of computer vision, the study of action and phenotype recognition has become a focal point in intelligent surveillance (Ben Mabrouk & Zagrouba, 2018), criminal investigation (Hossain et al., 2013), human-computer interaction (Ahmad & Khan, 2020), video prediction (Vyas et al., 2020), and healthcare (Venkataraman et al., 2013). In clinical applications, action recognition has been widely utilized in stroke rehabilitation using dynamical analysis of motion

Received: 08 March 2023; Accepted: 14 September 2023; Online: 15 September 2023
Foundation items: This work was supported by the National Key R&D Program of China (2021ZD0202805, 2019YFA0709504, 2021ZD0200900), National Defense Science and Technology Innovation Special Zone Spark Project (20-163-00-TS-009-152-01), National Natural Science Foundation of China (31900719, U20A20227, 82125008), Innovative Research Team of High-level Local Universities in Shanghai, Science and Technology Committee Rising-Star Program (19QA1401400), 111 Project (B18015), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), and Shanghai Center for Brain Science and Brain-Inspired Technology
[#]Authors contributed equally to this work
*Corresponding authors, E-mail: sxiong@fudan.edu.cn; xiaoxiao@fudan.edu.cn

(Venkataraman et al., 2013) and in assessing parkinsonism severity through gait-based characteristic recognition (Ricciardi et al., 2019). Studies on action recognition in non-human animals were first applied in pigeons (Dittrich & Lea, 1993), cats (Blake, 1993), and dogs (Delanoeije et al., 2020). Monkeys boast a greater range of motor behaviors compared to other experimental animals, as each of their body joints exhibits multiple degrees of freedom, enabling the production of a diverse array of postures. At present, however, action analysis in non-human primates (NHPs) requires labor-intensive manual observation and lacks standardized assessment.

Deep learning-based algorithms for human action recognition have demonstrated high accuracy and stability (Li et al., 2018; Simonyan & Zisserman, 2014; Tran et al., 2015). However, few deep learning tools are available for NHP applications. DeepLabCut provides unlabeled two-dimensional (2D) posture estimation with supervised learning and has been widely applied across various species, including flies, worms, rodents, and monkeys (Mathis et al., 2018; Nath et al., 2019). However, its accuracy for multiple actions and postures remains unsatisfactory (Supplementary Figure S1). The Kinect device is a powerful tool for automatic recognition of human bone points (Li et al., 2021b; Tran et al., 2017), but cannot easily provide information about the bone joints of monkeys. OpenMonkeyStudio harnesses a multi-view camera setup to generate three-dimensional (3D)-based estimations of the posture of unlabeled monkeys, demonstrating good accuracy (Bala et al., 2020). However, its optimal functioning requires 62 precisely arranged high-resolution video cameras, resulting in considerable resource costs, while a reduction to eight cameras yields a performance rate of only 80% accompanied by impaired accuracy (Bala et al., 2020). While Liu et al. (2022) developed MonkeyTrail, a deep learning-based approach for determining movement trajectories of caged macaques, there remains a need for a tool dedicated to phenotypic behavior recognition and fine action identification.

Rett syndrome (RTT), which predominantly affects girls, is one of the most severe neurodevelopmental disorders worldwide, primarily arising from mutations in the gene encoding methyl-CpG-binding protein 2 (*MECP2*) located on the X chromosome and subsequent downstream gene expression (Amir et al., 1999; Shah & Bird, 2017). Characterized by neurological regression, RTT profoundly affects motor abilities, especially mobility, hand skills, and gait coordination, accompanied by stereotyped features. Patients with RTT also experience anxiety, depression, and cognitive abnormalities (Chahrour & Zoghbi, 2007). Compared to the most widely used rodent model of RTT, NHP models offer the advantage of evolutionary homology with humans. Cynomolgus monkeys and humans share similar brain connectivity patterns with advanced cognitive function and behavioral characteristics, potentially providing a superior translational model involving *MECP2* mutant monkeys (Chen et al., 2017; Qin et al., 2019). In this study, we used this model to measure daily behaviors and detect pathological states, utilizing a deep learning-based algorithm for automatic action recognition.

We first created a benchmark dataset, called MonkeyinLab (MiL), which included 13 categories of NHP actions and postures, as an experimental laboratory-based model, with longitudinal video recordings (2 045 videos). Consequently, based on MiL, we established the MiL2D dataset containing 15 175 annotated images of 2D skeleton data and 15 bone points. Furthermore, we developed the MonkeyMonitorKit (MonKit) toolbox leveraging advanced deep learning techniques for video- and skeleton-based action recognition, enabling precise identification of phenotypic behaviors in NHPs. Finally, using the two datasets and toolbox, we detected daily behavior and estimated fine motor activities potentially related to RTT symptoms in five *MECP2* mutant RTT monkeys and 11 age-matched wild-type (WT) monkeys. Results showed that MonKit performance was comparable to human observations. By providing a validated, automatic, and objective behavioral analysis in NHPs, our toolkit and datasets hold promise for both experimental and clinical studies.

MATERIALS AND METHODS

Animals

Video recordings were obtained from 12 cynomolgus monkeys (*Macaca fascicularis*) to create the dataset and from 16 cynomolgus monkeys for testing both the dataset and MonKit. All monkeys were aged 6–8 years (5.5–12 kg) and were fed separately in single cages (see Table 1). Five *MECP2* mutant RTT monkeys and 11 age-matched WT monkeys were used in the study. Five male WT monkeys were video-recorded in their home cages (0.8 m×0.8 m×0.8 m or 1.0 m×0.8 m×0.8 m) as a baseline control (C1–C5). The term “home cage” refers to the daily living environment where monkeys typically resided and engaged in basic physiological activities, such as sleeping, eating, and drinking (Supplementary Figure S2). For behavioral observations and other experiments, six female WT monkeys (T1–T5) and five sex- and age-matched *MECP2* mutant monkeys (M1–M5) were moved to a standardized test cage (1.0 m×1.0 m×1.0 m) without access to food and water (Supplementary Figure S2). All home and test cages were situated in a controlled environment (temperature: 22±1 °C; relative humidity: 50%±5%) under a 12 h light/12 h dark cycle (lights off at 2000h and on at 0800h). The monkey facility where the experiments were conducted is accredited by AAALAC International and all experimental protocols were approved by the Institutional Animal Care and Use Committee of Yunnan Key Laboratory of Primate Biomedical Research (approval ID: LPBR201903003, to Prof. Yongchang Chen).

Video data collection

Each monkey was photographed in two distinct time-windows, once in the morning and once in the afternoon, during periods without feeding and foraging. Prior to video data collection, each monkey was placed in the test cage for one day for habituation and adaptation. To ensure accuracy during recording, each monkey was observed for at least 3 days, with recording time controlled from 0800h to 1100h and from 1300h to 1700h. Total recording time for each monkey was 3–10 h (3–6 h in home cage and 6–10 h in test cage). The recording equipment used included a network camera (model: DS-2CD1021FD-IW1, Hikvision, China) anddigital HD camcorder (model: HDR-CX405, Sony, Japan). The camera frame rate was 30 fps, with resolutions of 720×576 pixels in the home cage and 1 920×1 080 pixels in the test cage. The images were compressed to 256×340 pixels in the training and test sets.

Video data annotation

Video annotations were constructed utilizing Python scripts. The behavior recognition video data were set to an accuracy

Table 1 Animal information and recording environment

Serial number	Gender	Age (Year)	Recording time (min)	Shooting environment
WT in cage (C1)	Male	Unknown	268	Home cage with a plank in the middle.
WT in cage (C2)	Male	7	170	Home cage with a plank in the middle.
WT in cage (C3)	Male	7	325	Home cage with a plank in the middle.
WT in cage (C4)	Male	6	165	Home cage with a plank in the middle.
WT in cage (C5)	Male	8	148	Home cage with a plank in the middle.
WT in test (T1)	Female	6	600	Test cage with two rails in the top and a plank in the middle.
WT in test (T2)	Female	8	378	Test cage with two rails in the top and a plank in the middle.
WT in test (T3)	Female	6	381	Test cage with two rails in the top and a plank in the middle.
WT in test (T4)	Female	6	421	Test cage with two rails in the top.
WT in test (T5)	Female	8	204	Test cage with two rails in the top.
WT in test (T6)	Female	7	374	Test cage with two rails in the top.
MECP2 mutant (M1)	Female	7	493	Test cage with two rails in the top.
MECP2 mutant (M2)	Female	8	383	Test cage with two rails in the top and a plank in the middle.
MECP2 mutant (M3)	Female	7	385	Test cage with two rails in the top and a plank in the middle.
MECP2 mutant (M4)	Female	6	623	Test cage with two rails in the top and a plank in the middle.
MECP2 mutant (M5)	Female	7	387	Test cage with two rails in the top.

of 0.033–0.04 s (25–30 fps), while the bone point data were configured with a pixel range circle. To identify bone points accurately and eliminate line-of-sight “jitter”, a self-constructed Python script was used, accounting for context information.

Deep learning experimental setup

All experiments were conducted using PyTorch v1.7.1 for deep models, in conjunction with Nvidia Quadro RTX8000 GPU (memory: 48 Gb), Intel Xeon Gold 5220R CPU (2.2 GHz, 24 Cores), and Ubuntu v18.04 operating system.

For action recognition preprocessing, we regarded RGB and optical flow images obtained via OpenCV and Dense_flow as spatial and temporal features. These images were cropped around their centric pixels and standardized to a size of 224×224 pixels. Before the training phase, multi-site random crop (unified 224×224 pixel scale) and random horizontal flip (triggering probability set to 0.5) were employed for data augmentation. The model was trained for 50 epochs with a batch size of 90. Stochastic gradient descent (SGD) was used as an optimizer, with the initial learning rate, momentum factor, dropout rate, and weight decay set to 0.01, 0.9, 0.8, and 0.0005, respectively. Specifically, cross-entropy loss with learning rate decay was used, setting the step size and multiplicative factor to 20 and 0.1, respectively. The number of samplings for grouping was set to 8. During the testing stage, three patches (left, centered, and right) were cropped from the full pixel image to augment the test data, and the final score was calculated according to the weighted summation of the scores between RGB and optical flow, with their weighted ratio set to 1:1.5.

For keypoint recognition preprocessing, the images were cropped around their centric pixels and standardized to a size of 256×256 pixels. The space coordinates (x and y) of the keypoints, center point, and bounding box of each monkey were saved to a json file. Before the training phase, geometric transformation (e.g., rotation, flip), image processing (e.g., contrast, brightness), and different ChangeColor temperature were combined for data augmentation by IMGAUG toolbox and Pytorch's transformer module. To simulate the environment with iron bars, CageAUG was used to enhance the image data (Li et al., 2021a). The keypoint coordinates were converted into a heatmap and input into the network. The model was trained for 140 epochs, with the batch size set to 64. Adam was employed as an optimizer, and the initial learning rate, momentum factor, and weight decay were set to 0.01, 0.9, and 0.0001, respectively. The mean square error (*MSE*) was chosen as the loss function. During the testing phase, the accuracy rate was derived using PCKh as the benchmark standard of keypoint evaluation.

Statistical analysis

Analyses were conducted in small groups and ongoing power analysis was used to estimate the number of replicates required. The tests were specified to be two-sided, with similar variance observed between groups. Data exceeding three standard deviations (*SD*s) from the mean were excluded. Student’s *t*-test (two-tailed, unpaired) followed by the Mann-Whitney test was applied to identify significant differences using Prism v9.0.0(86). Data are presented as mean±standard error of the mean (*SEM*), and statistical significance was considered at $P<0.05$. For behavior recognition and daily behavior classifications, analyses were performed based on videos of monkeys in the different cages. Each video recording lasted for at least 3 h on different days. Null values (means that did not detect any specific behavior in a video) were removed during analysis.

RESULTS

MiL dataset

To establish uniform standards for identifying activities and behaviors in NHPs, we introduced the MiL benchmark dataset and used Python scripts to analyze each video. To increase sample diversity for training and testing, different sized sliding windows (20 to 110 frames) were set for the videos for data augmentation, ensuring that a specific action appeared several times in the dataset with different durations and combinations. According to the original standards for manual identification in previous studies (Chen et al., 2017; Feng et al., 2011; Harlow & Suomi, 1971; Hirasaki et al., 2000; Ma et al., 2017; Richter, 1931; Sun et al., 2017), the daily behaviors of monkeys were divided into 13 categories, including 10 action and three posture categories. Based on observations of long-term videos for each monkey in a single cage, the categories were identified in a total of 2 045 videos, forming the MiL dataset (Figure 1). The action categories

included: climb, hang, turn, walk, shake, jump, move down, lie down, sit down, and stand up. Postures were detected during low activity periods and included: stand, sit, and huddle. The definitions of each category are shown in Table 2. The MiL dataset of the 13 categories covered nearly all daily behaviors of NHPs housed under single-cage conditions.

Two-stream model based on temporal shift and split attention (TSSA) for action detection

The specific network block diagram is shown in Figure 2A. Initially, the category tag was input into each action video, with video durations of approximately 1–4 s, covering all daily monkey actions. Utilizing a random sampling approach coupled with sparse temporal grouping, eight representative frames depicting each discrete action were extracted from each individual short video. The corresponding optical flow picture and RGB flow information were input into the temporal shift (TS) (Lin et al., 2019) and split attention (SA) (Zhang et al., 2022) (TSSA) network, respectively through the parts of RGB net and optical flow net for training. In each net module, residual neural network (ResNet)-50 was used as the backbone network, with the addition of a self-attention mechanism module to the feature map position of each layer. The feature map was subsequently partitioned into different cardinals, and a series of transformations were applied with different weights to enhance the importance of certain cardinals and overall network performance. Finally, the corresponding action category was output through softmax

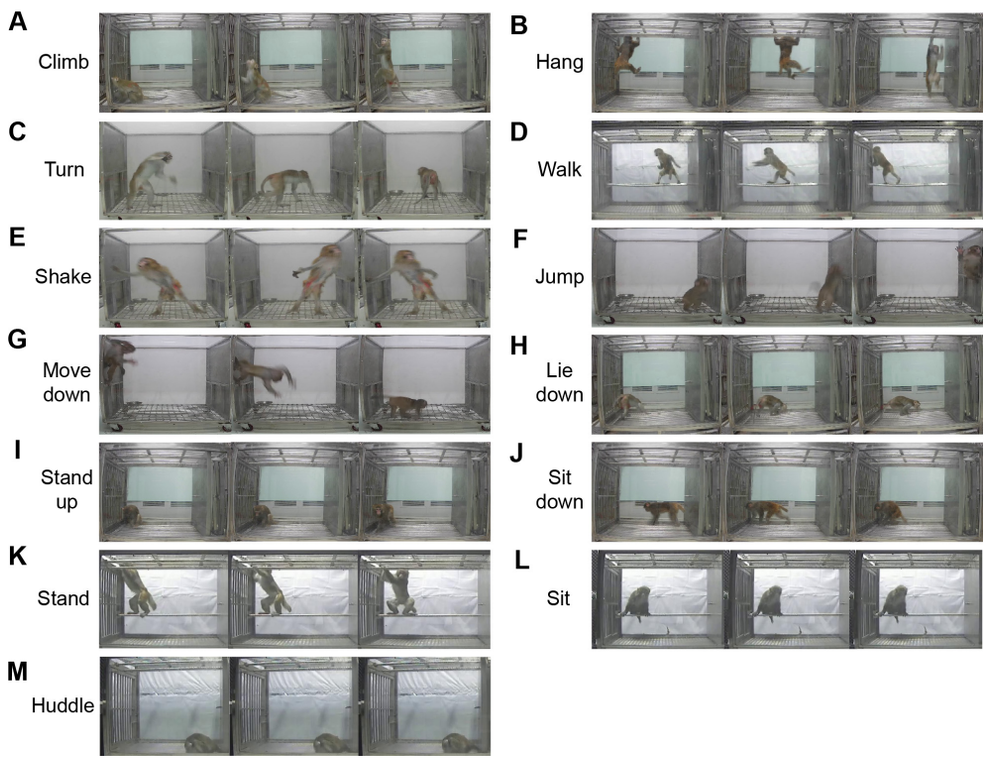

Figure 1 Examples of MiL dataset, with videos corresponding to actions and postures (labels)

A–J: Ten action categories. A: Climb; B: Hang; C: Turn; D: Walk; E: Shake; F: Jump; G: Move down; H: Lie down; I: Stand up; J: Sit down. K–M: Three posture categories. K: Stand; L: Sit; M: Huddle. Each row represents non-contiguous frames randomly sampled in the corresponding video. Video lengths range from 20 to 110 frames.

Table 2 Definitions of 10 actions and three postures in MiL dataset

Actions	
Climb	Move slowly from the ground to the side wall or from side wall to the top of the cage by using all arms and legs.
Hang	Grab the levers in the top of the cage or move from one lever to another.
Turn	Bend over from standing position or turn around using all arms and legs.
Walk	Stand mainly use legs only and walk on the ground or on the horizontal pole in the cage.
Shake	Body shaking with two feet still.
Jump	Move quickly from ground into the air or to the side wall by mainly using legs only.
Move down	Move quickly from the side wall to the ground.
Lie down	From standing on four legs to laying prone.
Sit down	Move from other posture to sitting or squatting.
Stand up	Move from sitting or squatting posture to standing.
Postures	
Stand	Remain standing on two legs.
Sit	Remain sitting of the ground or the horizontal pole.
Huddle	Crawl or lay on the ground and curl up body.

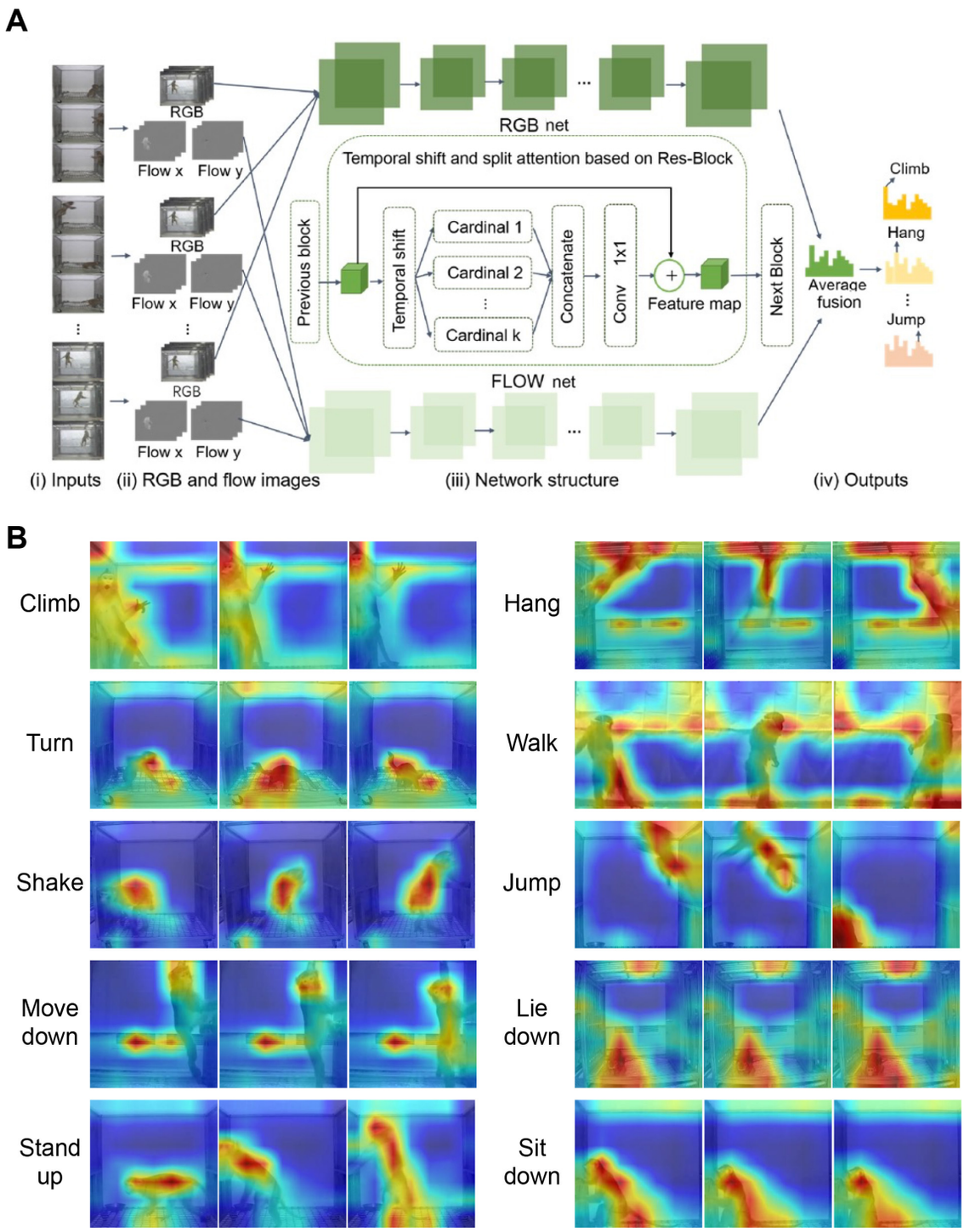

Figure 2 Two-stream action recognition model and heatmaps of feature visualization

A: Overview of TSSA Network architecture. (i) Video segments are randomly sampled as input; (ii) Two video modalities, RGB and optical flow, serve as inputs in the two-stream model; (iii) Separate networks with the same architecture, each containing Res-blocks as backbone and shift and split attention modules in blocks; (iv) Output from previous block is used as input for feature extraction in the next block. Single-stream net predicts action scores using average fusion, and class scores are combined for the final prediction. B: Grad-Cam++ heat maps of action recognition. Heat maps obtained from test videos classified under the trained model. Colors represent different weights (ranging from 0–1, blue to red) signifying the importance of the area related to the prediction result. Red area in the frames provided the most important discriminative features used by the model in the final predictions.

and onehot encoding. The evaluation index was top-1 and the final accuracy rate achieved using the two-stream neural network based on TSSA fusion (Xiao et al., 2022) was 98.99% (RGB 89.83%, Flow 93.05%, RGB+Flow 98.99%). The outcomes of action recognition, represented by the confusion matrix, are provided in Supplementary Figure S3. Furthermore, the Grad-Cam++ approach (Chattopadhay et al., 2018), a method used for feature visualization of CNN model predictions, was applied for heatmap construction (Figure 2B). The heatmap indicated the varied contributions of the attention

module on distinct areas in the classification results. The weighted combination of positive partial derivatives from the final convolutional layer in the feature map was used to provide a specific class score, as shown from red to blue. We observed distinct postures resulting from a single action, thus providing discriminative features for action predictions. For instance, the network effectively captured the “Hang” category based on specific position characteristics (e.g., the angle between the monkey's body and ceiling was almost 90°, and the arms held the rails from the ceiling). Similarly, the “Jump” category” was characterized by a change in vertical position. The algorithm also captured characteristics such as head drooping and hip raising for the "Lie down" category, while for the "Stand up" category, the attention module focused on tracking the movement of the head and body. Through feature visualization, we determined the localizations of each object and weights of features. Most of the attention module areas corresponded to the monkey’s body.

Action recognition and daily behavior classifications

Eleven WT monkeys and five *MECP2* mutant monkeys were video recorded in their home or test cages, allowing for the observation of spontaneous daily behaviors. Using the proposed two-stream action recognition model and the MiL dataset, different daily behavior patterns were observed among the three groups of monkeys. As expected, WT monkeys in their home cages spent more time engaged in low-activity behaviors and less time in other categories compared to both their behavior and that of *MECP2* mutants in the test cages (Figure 3A; Supplementary Figure S4). The *MECP2* mutant monkeys exhibited decreased activity duration and patterns in comparison to WT monkeys within the confines of the test cages (Figure 3A). Specifically, WT monkeys in the test cages spent significantly more time in the Climb (*P*=0.078 compared to *MECP2* mutants), Turn, and Lie down categories (Figure 3B, D, H). The groups in the test cages demonstrated significantly higher percentages of Hang and Walk compared to the WT monkeys in the home cages (Figure 3C, E), while exhibiting lower levels of low-activity behavior (Figure 3J). Notably, the *MECP2* mutant monkeys spent markedly less time performing Jump and Move down

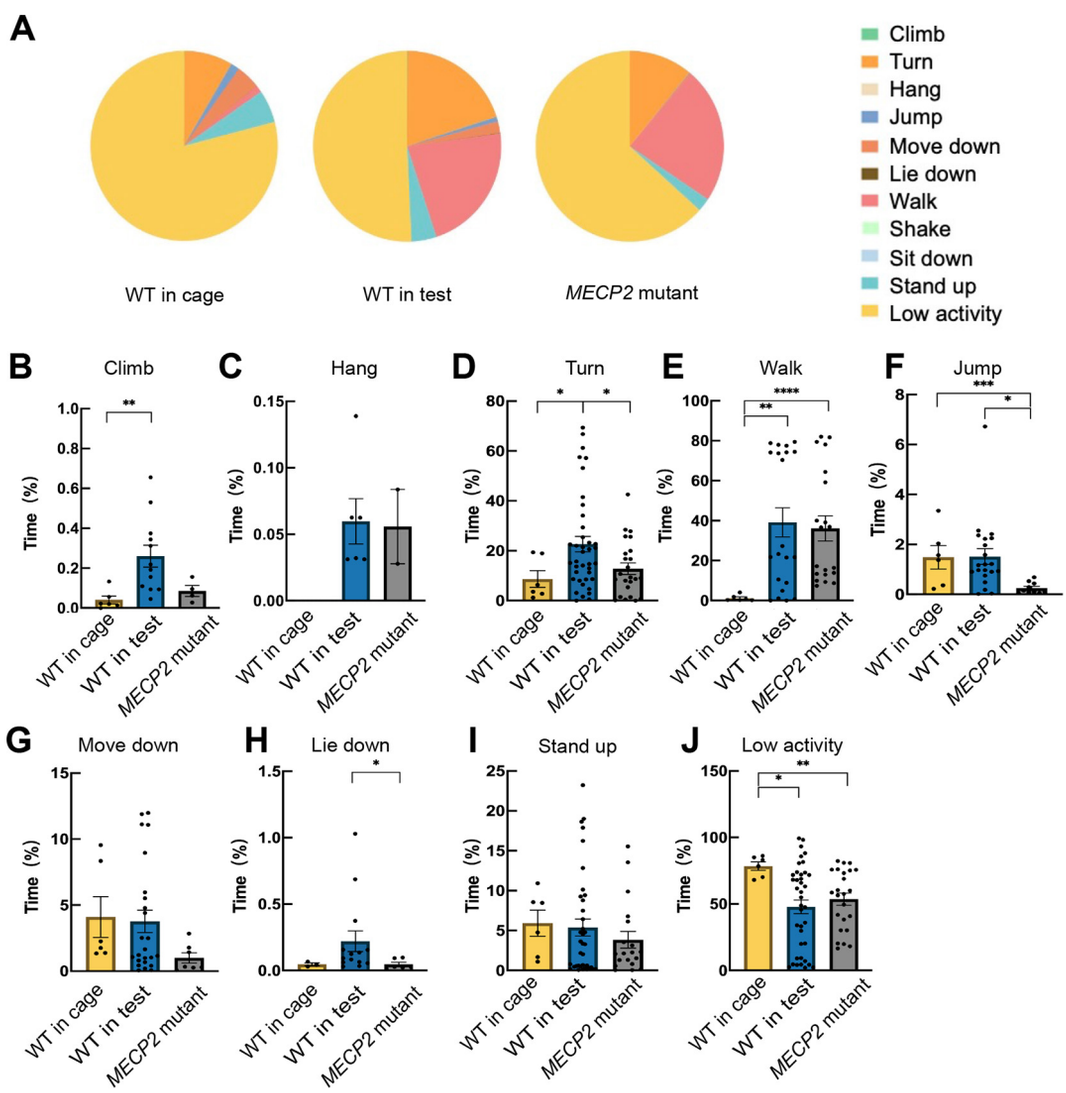

Figure 3 Action recognition of daily performance in WT monkeys in home cage, WT monkeys in test cage, and *MECP2* mutant monkeys in test cage

A: Percentage of time spent in detected action categories. B–J: Time spent in B: Climb, C: Hang, D: Turn, E: Walk, F: Jump, G: Move Down, H: Lie Down, I: Stand Up, and J: Low Activity. Each dot represents a video clip more than 3 h in length in one day. *: *P*<0.05; **: *P*<0.01; ***: *P*<0.001; ****: *P*<0.0001.

(Figure 3F, G). Moreover, the *MECP2* mutant monkeys showed reduced activity in high intensity and challenging actions, consistent with RTT patients experiencing severe motor disabilities. As expected, the behavioral performance of the WT monkeys differed in the home and test cages.

MonKit toolbox and keypoint prediction

Based on the MiL dataset, we established the MiL2D dataset of images with 2D skeleton and key bone points. The MiL2D dataset consisted of 15 175 annotated images spanning a large variation of poses and positions seen in the 13 MiL categories. In total, 15 skeleton keypoints were marked in detail (Figure 4A). MaskTrack R-CNN (Yang et al., 2019) was used to track the positions of the monkeys (Figure 4B). The dataset included diverse configurations of cage environments, and monkeys with corresponding skeleton points were

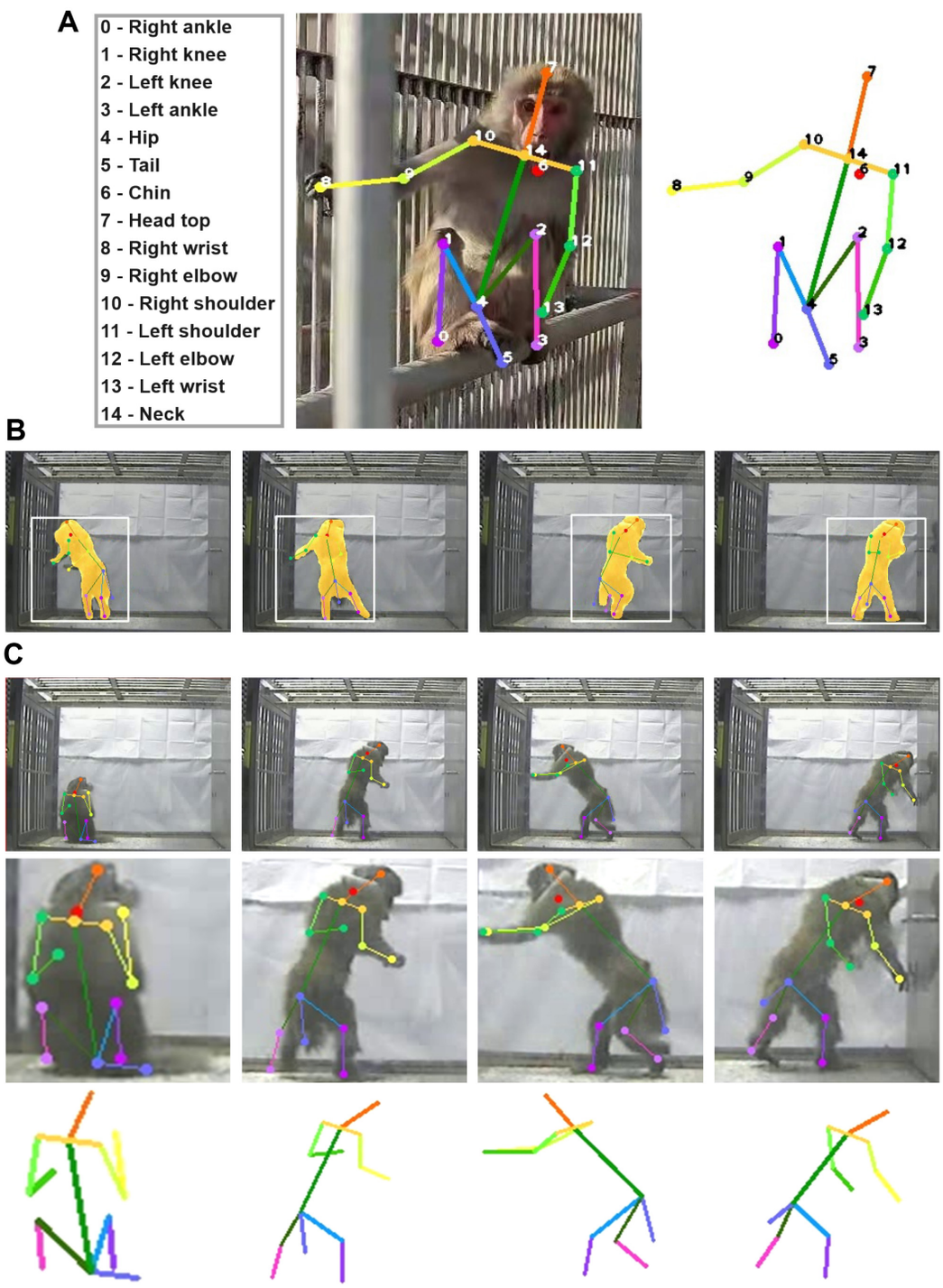

Figure 4 Illustrations of MiL2D dataset with 15 skeleton keypoints

A: Definition and location of 15 bone points. 0, right ankle; 1 right knee; 2, left knee; 3, left ankle; 4, hip; 5, tail; 6, chin; 7, head top; 8, right wrist; 9, right elbow; 10, right shoulder; 11, left shoulder; 12, left elbow; 13, left wrist; 14, neck. B: Representative images with bounding boxes using MaskTrack R-CNN tracking and 2D skeleton. C: Representative images of monkey panorama, corresponding skeleton point diagram, and partial enlargement of monkey.

detected clearly in the MiL2D dataset (Figure 4C).

Using the MiL2D dataset, we conducted a monkey bone recognition task to train and test the MonKit toolbox based on a high-resolution network (HRNet) (Wang et al., 2021) (Figure 5). Images from the original input video were processed to 256× 340 pixels. MaskTrack R-CNN was used to track the position of the monkeys (Figure 5B). Subsequently, the rectangle information representing the position of each monkey was intercepted and input into the HRNet to generate a heatmap (Figure 5D). The *MSE* loss function was applied to compare the target and calculate the loss. Finally, the 15 bone points were transformed into x and y space coordinates (Figure 5E; Supplementary Figure S5). The achieved accuracy of 98.8% with the MiL2D and OpenMonkeyStudio datasets is consistent with our previous study, which only used the OpenMonkeyStudio dataset for training (Li et al., 2021a).

Posture recognition and estimates of fine motor activities

We detected and predicted three postures (huddle, sit, and stand) observed in the daily life of monkeys using accurate height information obtained from bone points. The y-axis coordinates of four skeleton points (0, right ankle; 3, left ankle; 4, hip; 7, head top) were acquired through HRNet, and ($y_{max}y_{min}$) was calculated to determine height information (Figure 6A), while excluding interference from the monkey's tail. Based on analysis, the three groups of monkeys (WT in home cage, WT in test cage, and *MECP2* mutant) showed no significant differences in time spent in the three postures (Figure 6B–D; Supplementary Figure S6).

Utilizing the MonKit dataset and keypoint prediction, we also detected fine motor activities characterized by stereotyped patterns and head-down behaviors with relatively small motion amplitudes. Stereotyped behavior patterns mainly refer to repetitive and purposeless body movements at a fixed frequency, often observed in RTT patients as a feature of autism. In monkeys, stereotyped behaviors include turning over, circling, pacing, and cage shaking. In the current study, we determined bone point recognition by HRNet to estimate stereotyped behaviors in monkeys. The coordinates of the center point of the monkeys were determined by calculating the sum of the 15 bone points with their respective vector directions. The formula used was:

$$C_{center}(x,y) = \left(\frac{1}{N}\sum_{i=0}^{N} x_i, \frac{1}{N}\sum_{i=0}^{N} y_i\right) \quad (1)$$

where N is the number of bone points (n=15 in this study). The average spatial position values (x and y) of the bone points were calculated. Stereotyped behavior is characterized by repetitive movements, implying that the vector direction of a specific action should be equivalent to zero, indicating that animals move in reciprocal motion patterns (Figure 7A). The

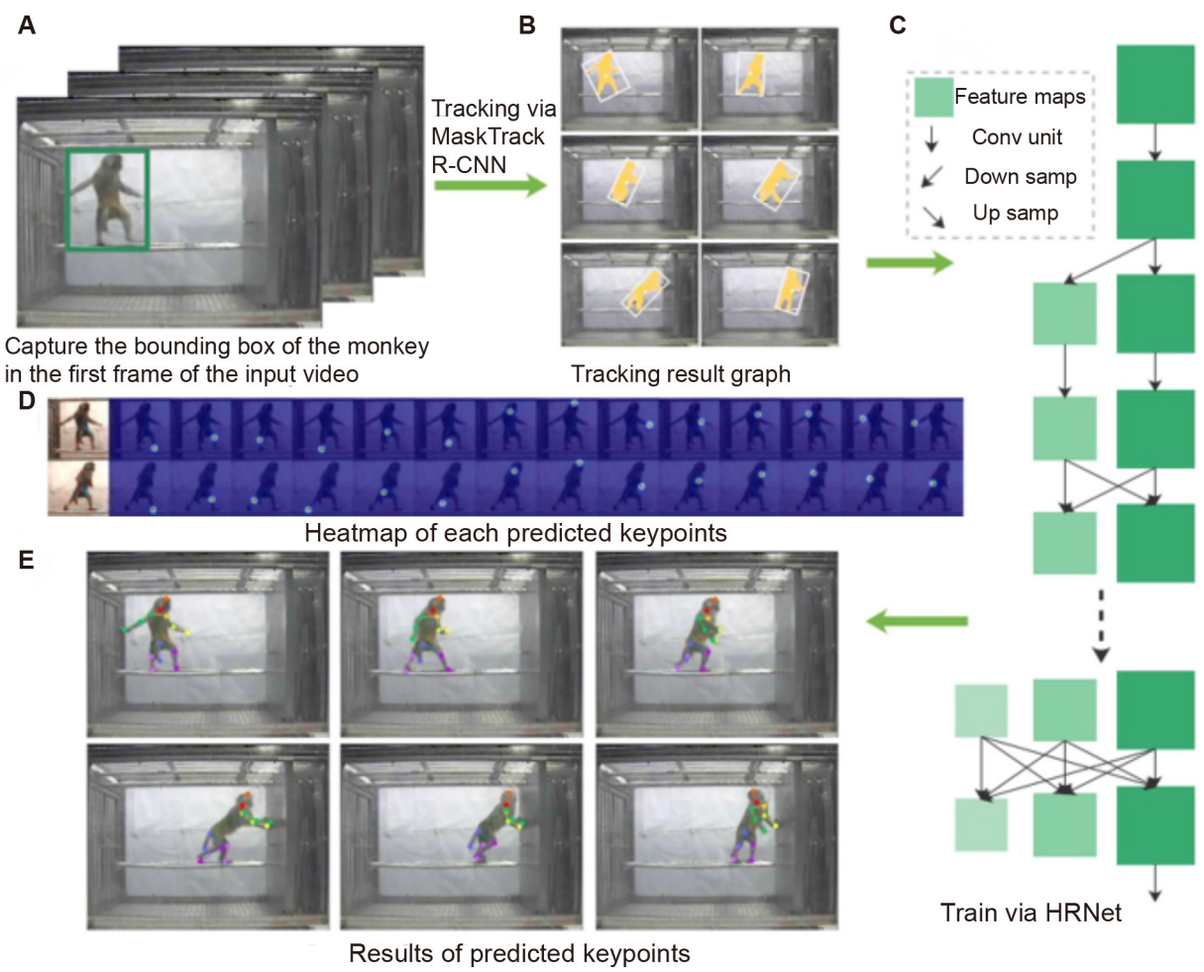

Figure 5 MonKit for action tracking and posture estimation based on MaskTrack R-CNN and HRNet

A: Input videos. B: Tracking results obtained by MaskTrack R-CNN, showing frames 1, 4, 7, 10, 13, 16, and 19, respectively. C: Inputs of tracked monkeys in HRNet network for training or testing. D: Heatmap of 15 keypoints (neck position is obtained by taking the center of the left and right shoulders). E: Heatmap conversion to obtain x and y positions of bone points.

specific formula used was as follows:

$$\left(V_{x_i} + V_{x_{i+1}} + \cdots + V_{x_{i+n}}\right) + \left(V_{y_i} + V_{y_{i+1}} + \cdots + V_{y_{i+n}}\right) < T \qquad (2)$$

where vector (V) of the x and y dimensions in a certain period is less than a certain threshold (T). Three of the five *MECP2* mutant monkeys spent more than 20% of their time in

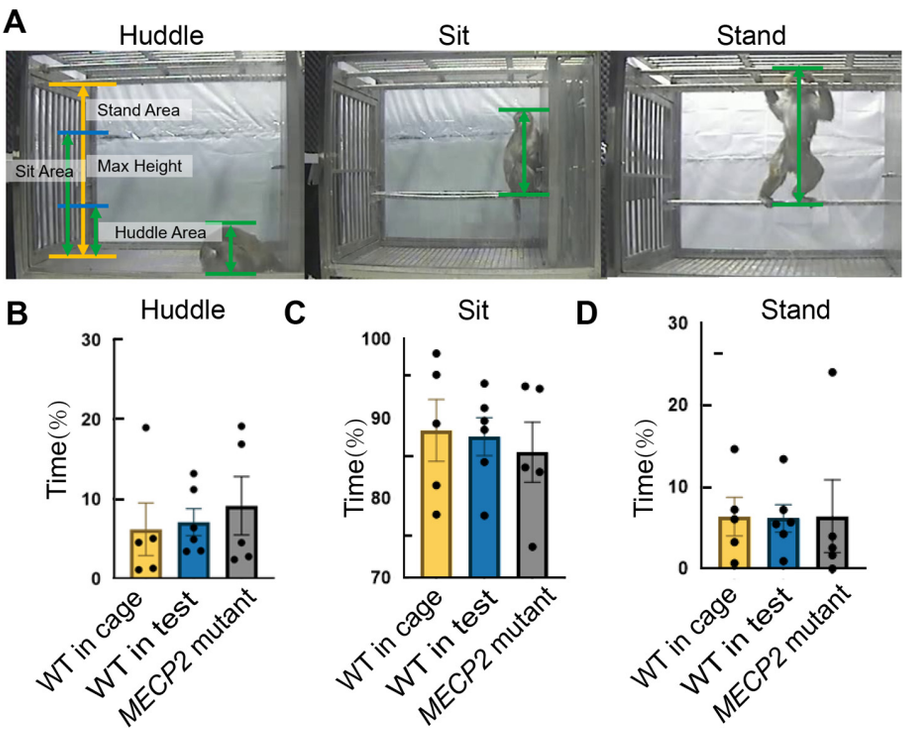

Figure 6 Posture recognition detected by MonKit in WT monkeys in home cage, WT monkeys in test cage, and *MECP2* mutant monkeys
A: Diagram of height calculation for huddle, sit, and stand postures. B–D: Time spent in B: Huddle, C: Sit, and D: Stand. Each dot represents an individual monkey with average time spent in each video clip.

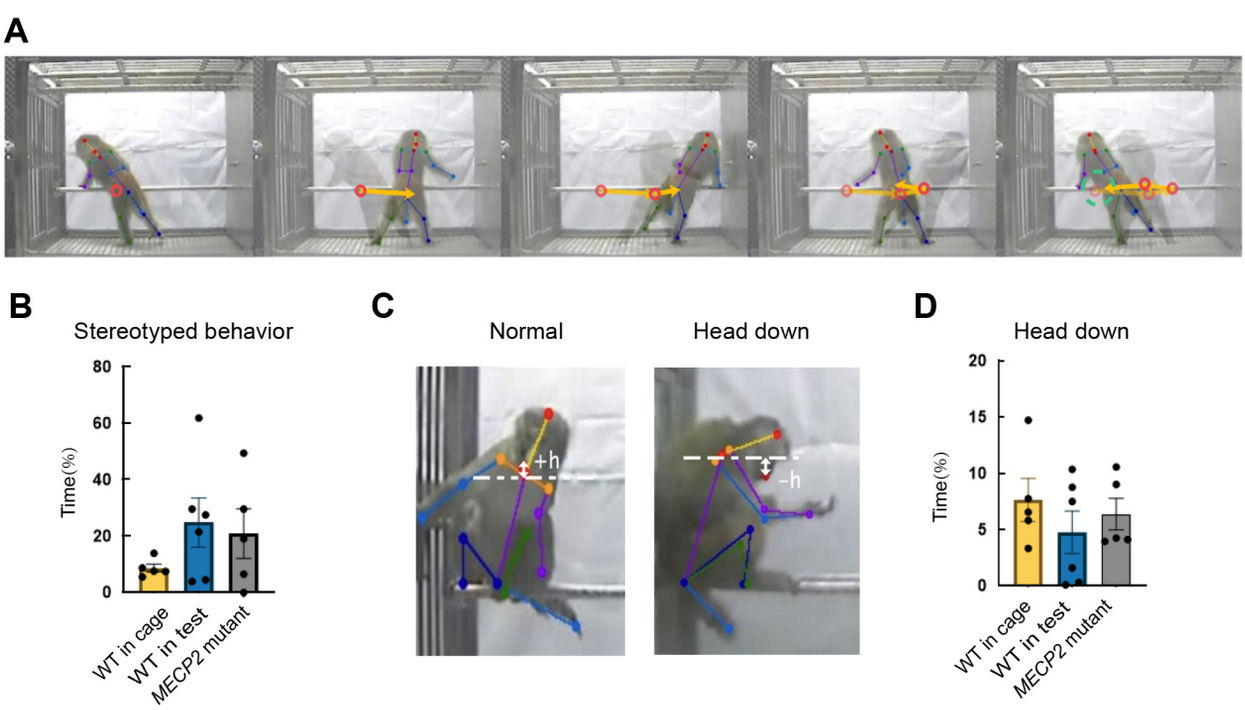

Figure 7 Stereotyped behavior patterns and head-down posture detected by MonKit
A: Series of representative images of stereotyped pacing behavior. B: Stereotyped behavior in WT monkeys in home cage, WT monkeys in test cage, and *MECP2* mutant monkeys. C: Representative images of normal and head-down behavior and diagram of chin and neck bone points. D: Head-down posture in WT monkeys in home cage, WT monkeys in test cage, and *MECP2* mutant monkeys. Each dot represents an individual monkey with average time spent in each video clip.

stereotyped behavior, while four of the WT monkeys also performed stereotyped behaviors in the test cage, suggesting that stereotyped patterns may also represent anxiety or hyperactivity states in monkeys (Figure 7B).

Depressive behavior has been observed in a considerable minority of female RTT patients (Hryniewiecka-Jaworska et al., 2016). In addition to the huddling posture, a fetal-like, self-enclosed posture, with the head positioned at or below the shoulders during the awake state, is also considered as a measure of depression-like behavior in monkeys (Hryniewiecka-Jaworska et al., 2016). Here, we detected head-down behavior using the y-axis coordinates of the bone points corresponding to the neck and chin in monkeys during performance of low activity. The formula used was:

$$h = y_{chin} y_{neck} (3) \quad (3)$$

with h<0 indicating a head-down posture (Figure 7C). Based on MonKit detection, the duration of time spent in the head-down posture showed no significant increase in either the *MECP2* mutant monkeys or WT monkeys in the test group compared to the WT monkeys in the cage group (Figure 7D).

Quality control and manual comparison

To confirm the accuracy of MonKit, we performed manual analysis of the videos and compare the results to those obtained from MonKit detection. As manual detection is very labor-intensive and time-consuming, we randomly selected one video from each monkey and analyzed the first 20 min of each video (Figure 8A, B). Discrepancies among the 11 action categories were compared based on MonKit and manual detection, revealing that the detection error of MonKit was consistently below 5% for all actions (Figure 8C). The precision and recall results are shown in Supplementary Table S1 and Supplementary Videos S1 and S2. The low activity category showed the largest detection error (4.9% on average) (Figure 8C). Similarly, three monkeys were randomly selected to manually analyze posture, stereotyped behavior, and head-down behavior. MonKit detection yielded similar time counts for the sit (Figure 8D), stand (Figure 8E), and huddle (Figure 8F) postures as manual detection. Additionally, the detection accuracy for stereotyped behavior and head-down behavior showed similar results between MonKit and manual detection (Figure 8G, 8H).

DISCUSSION

In the current study, the daily behaviors of cynomolgus monkeys in both home and test cages were video recorded and automatically analyzed using MonKit. To the best of our knowledge, MonKit is the first deep learning-based toolkit designed for identifying fine motor activities in NHPs. *MECP2* mutant monkeys, as a disease model of RTT, and age-

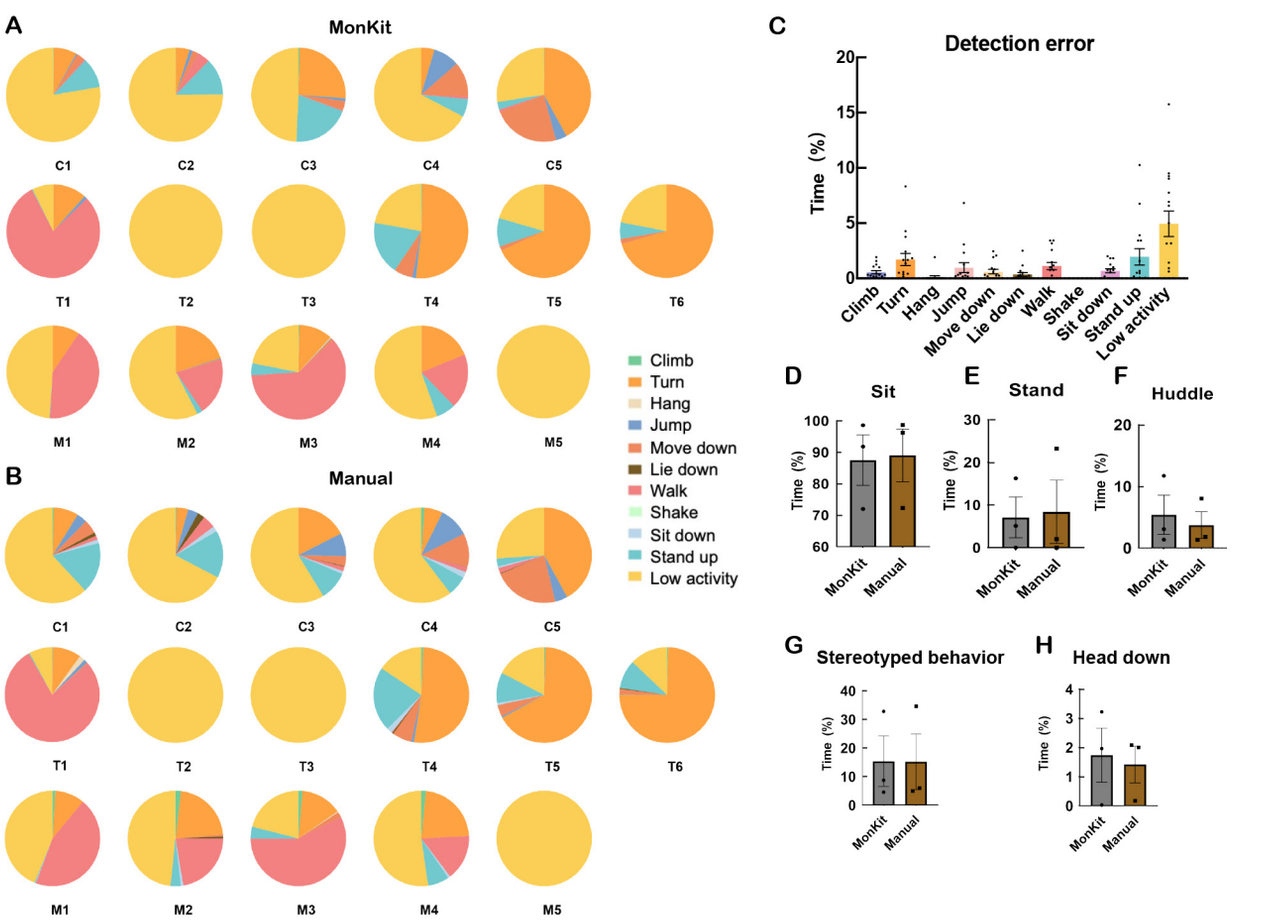

Figure 8 Comparison of MonKit and manual detection accuracy

A: MonKit detection of 11 action categories in all individual monkeys. B: Manual detection of 11 action categories in all individual monkeys. C: Detection error (time in MonKit–time in Manual) of 11 action categories. D–H: Time spent in sit (D, monkeys C1, C5, and M5), stand (E, monkeys C1, C5, and M5), huddle (F, monkeys C1, C5, and M5), stereotyped (G, monkeys of M5, T1, and T4), and head-down behaviors (H, monkeys M1, T1, and C1).

matched control monkeys were placed in test cages to evaluate their motor functions, stereotyped behaviors, and depressive phenotypes.

We first created two benchmark datasets (MiL and MiL2D) to facilitate action recognition in free-moving monkeys in their daily life settings, serving as experimental and preclinical models in the laboratory (Figure 9A). Using longitudinal videos recorded using one camera at a single angle, which is comparatively low cost and easy to produce, we proposed MonKit as an effective deep learning and 2D skeleton-based model for the capture of discriminative spatiotemporal features of daily actions and postures in monkeys, enabling the identification of variances between different environments and genotypes (Figure 9B). Our findings revealed that monkeys exhibited increased physical activities and higher levels of stereotyped behaviors in the test cages compared to their

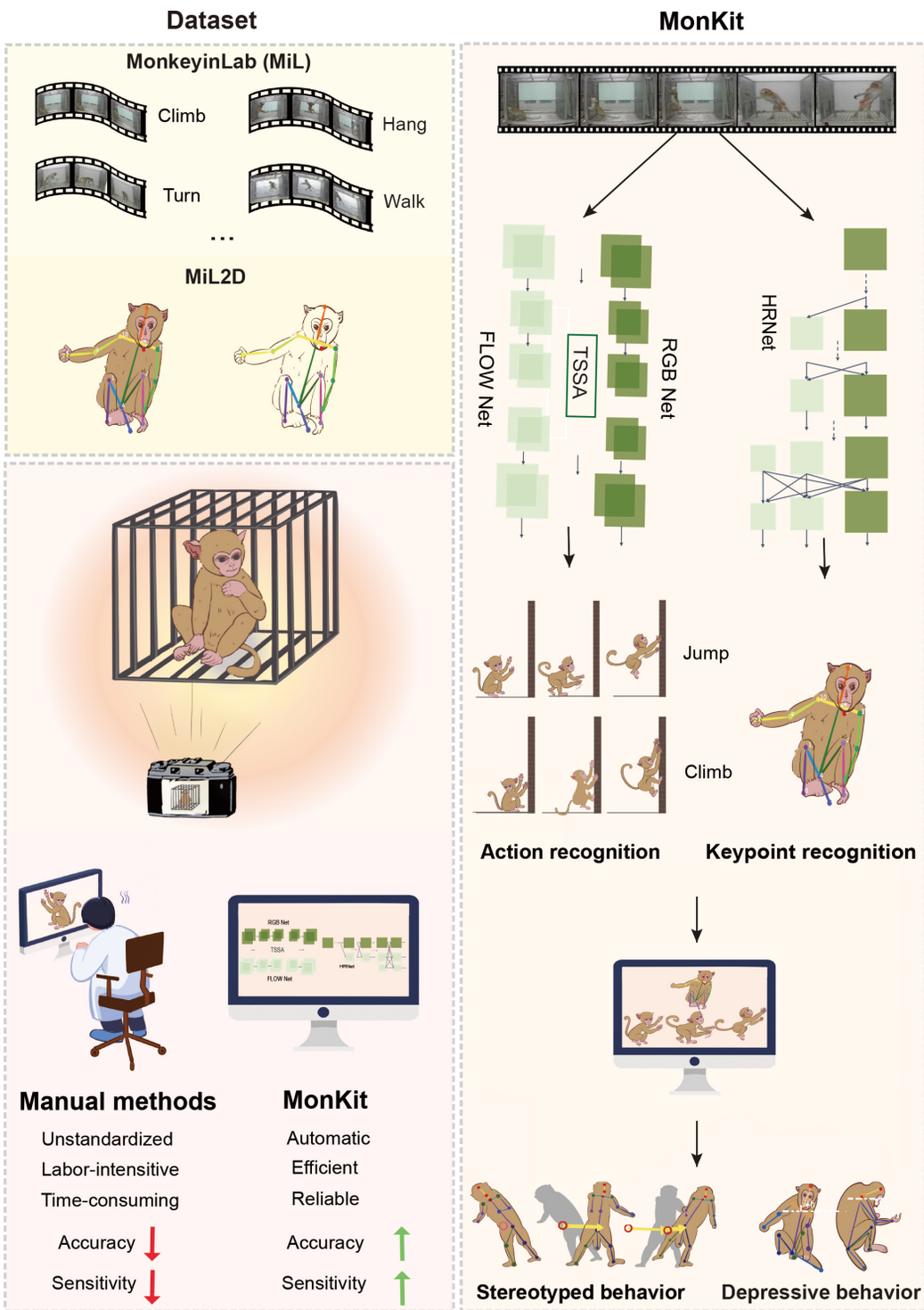

Figure 9 Schematic of datasets and MonKit

A: Action recognition dataset (MonkeyinLab, MiL) and keypoint dataset (MiL2D). B: Original videos are input into TSSA and HRNet networks to obtain action recognition and keypoints, with fine motor identification analysis then conducted on the recognition results. C: Upper part shows angle of the camera shooting the cage. Lower part shows comparison of manual and MonKit methods.

home cages. Notably, the *MECP2* mutant monkeys demonstrated differences in active behaviors, such as climbing, jumping, and moving down, but no significant differences in stereotyped or depressive behavioral phenotypes. MonKit exhibited good performance, with a detection accuracy closely matching that of manual detection. The model achieved a detection error of less than 5% for all actions and showed no difference in the recognition of postures and fine motor activities compared to manual human detection (Figure 9C). These results highlight the efficacy and high accuracy of MonKit in automatic action recognition.

Compared to widely used rodent models, NHP models offer substantial advantages due to their evolutionary homology with humans, similar brain connectivity patterns, advanced cognitive functionality, and other behavioral characteristics (Qin et al., 2019). Thus, action recognition of phenotypic behaviors is a key feature of NHP models (Liu et al., 2016; Zhou et al., 2019). To date, however, action analysis in NHPs has predominantly relied on manual labor and subjective assessments, resulting in labor-intensive and time-consuming processes with limitations in sensitivity and accuracy. To overcome these challenges, the application of deep learning-based methods has achieved high performance via end-to-end optimization and has gained acceptance in many prediction tasks, such as action recognition and feature extraction. Currently, most action recognition and pose estimation models are based on 3D-CNN (Kay et al., 2017; Ng et al., 2015; Tran et al., 2015), long short-term memory (LSTM) (Li et al., 2018; Sharma et al., 2015), two-stream CNN (Feichtenhofer et al., 2016; Karpathy et al., 2014; Simonyan & Zisserman, 2014; Soomro et al., 2012), ResNet (He et al., 2016a, 2016b; Xie et al., 2017), and HRNet models (Wang et al., 2021). The two-stream model, first proposed by Karpathy et al. (2014), has evolved to incorporate CNN-based local spatiotemporal information, thus achieving connectivity in the spatiotemporal domain and improving CNN performance by analyzing additional motion information.

In this study, we adopted a random sampling strategy with sparse temporal grouping (Wang et al., 2016) to ensure effective temporal structure modeling over a long-term range. Additionally, we proposed a novel spatiotemporal two-stream model based on TSSA (Lin et al., 2019; Zhang et al., 2022) modules. The TS module enabled learning of temporal features, while the SA mechanism facilitated focus learning (i.e., with an attention mechanism) to generate further discriminative features for improved recognition. In our previous study, the TSSA network showed 98.99% accuracy on the MiL dataset (Xiao et al., 2022). By employing a random sampling strategy with sparse temporal grouping from input videos, we effectively modeled long-term content with enhanced robustness and generalization.

Keypoints (body part positions) and bone skeleton-based action recognition have been widely applied in human behavior analysis (Li e t al., 2020; Lo Presti & La Cascia, 2016). While human datasets for bone recognition, such as coco and MPII, are relatively well-established (Andriluka et al., 2014; Chen et al., 2018), animal datasets and keypoint and bone recognition methods are still in their developmental stages. The MacaquePose dataset, comprising 16 393 monkeys captured in 13 083 pictures, provides manually labeled keypoints for macaques in naturalistic scenes, serving as a valuable resource for training and testing networks to analyze monkey movement (Labuguen et al., 2021). Other tools like DeepLabCut (Mathis et al., 2018), LiftPose3D (Gosztolai et al., 2021), DANNCE (Karashchuk et al., 2021), OpenMonkeyStudio (Bala et al., 2020), and MaCaQuE (Berger et al., 2020) have also been proposed for 2D and 3D tracking of animals, such as *Drosophila*, chickadees, rodents, and NHPs, in the laboratory and other environments. In our study, we created the MiL2D dataset containing 2D skeleton annotated images and further generated the MonKit toolkit, not only enabling action recognition and posture estimation but also allowing measurement of fine motor activities from longitudinally observed monkeys in different groups. Fine motor abilities are integral to a diverse range of movement skills and interact continuously with psychological, cognitive, emotional, and social functions (Van Damme et al., 2015). Impairments in motor and fine motor abilities offer insights into pathophysiological disruptions associated with neurological diseases and mental disorders, such as Parkinson's disease, Alzheimer's disease, attention deficit hyperactivity disorder (ADHD), autism spectrum disorders (ASD), schizophrenia, and depression (Downey & Rapport, 2012; Mendes et al., 2018; Sabbe et al., 1996; Viher et al., 2019). As a severe neurodevelopmental disorder, RTT exhibits a phenotype characterized by motor dysfunctions, as well as autistic features and emotional and cognitive deficits. Here, we observed a severe decline in motor ability in *MECP2* mutant monkeys, consistent with that observed in RTT patients. However, we found similar levels of repetitive stereotyped behaviors in the test cage monkeys and *MECP2* mutant monkeys, both higher compared to the home cage monkeys. This observation may be attributed to the test cage monkeys experiencing a heightened state of anxiety, as anxiety is reported to be an intrinsic motivator for repetitive behaviors in children with ASD (Cashin & Yorke, 2018; Joosten et al., 2009).

In conclusion, our MiL and MiL2D datasets, along with the MonKit toolkit, demonstrate the feasibility of an automatic and objective analysis system for quantifying NHP behavioral models (Figure 9A, C). Our experimental setup, consisting of one camera at a fixed angle, is cost-effective, convenient, and simple to install (Figure 9C). Currently, MiL and MiL2D analyses of video recordings have focused on a single monkey in a single cage, but future extensions of the system will encompass social interactions among multiple monkeys. Notably, MonKit can serve as an auxiliary tool for efficient, accurate, and interference-free behavior recognition and symptom identification in NHPs. To further improve the performance of our approach, we will introduce few-shot learning (Cao et al., 2020) and more effective backbones in the future (Qin et al., 2020). Although no significant differences were found in head-down behavior of the *MECP2* mutant monkeys, it does not eliminate the potential existence of emotional or cognitive deficits in monkeys with this genotype. Taking advantage of the capabilities of MiL2D and MonKit in fine activity analysis, we can explore additional phenomena, such as grooming, manual performance in cognitive tasks, eating patterns, circadian rhythm, social isolation, and aggressive behavior, thus contributing to the establishment of a comprehensive behavioral analysis system for NHPs in both basic and clinical studies.

DATA AVAILABILITY

The datasets and MonKit generated and/or analyzed in the current study are available from the corresponding author on reasonable request. Our

model is provided and maintained on our GitHub repository (https://github.com/MonKitFudan/MonKit).

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

C.L., Z.X., and X.X. conceived the research and experiments. C.L., Z.X., and Y.L. designed the network structure and performed data collection and analysis. Y.L., Z.C. and X.J. participated in data collection and analysis. S.F., Z.Z., Y.C., and K.Z. provided video data and suggestions for the design of the dataset. J.F and T.W.R. provided valuable suggestions and advice on the research and experiments. X.X., S.X., and Y.C. provided the funding to support the research. X.X. supervised the study and led the writing of the manuscript. All authors read and approved the final version of the manuscript.

REFERENCES

Ahmad Z, Khan N. 2020. Human action recognition using deep multilevel multimodal (M^2) fusion of depth and inertial sensors. *IEEE Sensors Journal*, **20**(3): 1445−1455.

Amir RE, Van den Veyver IB, Wan MM, et al. 1999. Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nature Genetics*, **23**(2): 185−188.

Andriluka M, Pishchulin L, Gehler P, et al. 2014. 2D human pose estimation: new benchmark and state of the art analysis. *In*: Proceedings of 2014 IEEE Computer Vision and Pattern Recognition. Columbus: IEEE.

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

Ben Mabrouk A, Zagrouba E. 2018. Abnormal behavior recognition for intelligent video surveillance systems: a review. *Expert Systems with Applications*, **91**: 480−491.

Berger M, Agha NS, Gail A. 2020. Wireless recording from unrestrained monkeys reveals motor goal encoding beyond immediate reach in frontoparietal cortex. *eLife*, **9**: e51322.

Blake R. 1993. Cats perceive biological motion. *Psychological Science*, **4**(1): 54−57.

Cao KD, Ji JW, Cao ZJ, et al. 2020. Few-shot video classification via temporal alignment. *In*: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 10618–10627.

Cashin A, Yorke J. 2018. The relationship between anxiety, external structure, behavioral history and becoming locked into restricted and repetitive behaviors in autism spectrum disorder. *Issues in Mental Health Nursing*, **39**(6): 533−537.

Chahrour M, Zoghbi HY. 2007. The story of Rett syndrome: from clinic to neurobiology. *Neuron*, **56**(3): 422−437.

Chattopadhay A, Sarkar A, Howlader P, et al. 2018. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. *In*: Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE, 839–847.

Chen YC, Yu JH, Niu YY, et al. 2017. Modeling rett syndrome using TALEN-Edited *MECP2* mutant cynomolgus monkeys. *Cell*, **169**(5): 945−955.e10.

Chen YL, Wang ZC, Peng YX, et al. 2018. Cascaded pyramid network for multi-person pose estimation. *In*: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE.

Delanoeije J, Gerencsér L, Miklósi Á. 2020. Do dogs mind the dots? Investigating domestic dogs' (*Canis familiaris*) preferential looking at human - shaped point - light figures. *Ethology*, **126**(6): 637−650.

Dittrich WH, Lea SEG. 1993. Motion as a natural category for pigeons: generalization and a feature - positive effect. *Journal of the Experimental Analysis of Behavior*, **59**(1): 115−129.

Downey R, Rapport MJK. 2012. Motor activity in children with autism: a review of current literature. *Pediatric Physical Therapy*, **24**(1): 2−20.

Feichtenhofer C, Pinz A, Zisserman A. 2016. Convolutional two-stream network fusion for video action recognition. *In*: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 1933–1941.

Feng XL, Wang LN, Yang SC, et al. 2011. Maternal separation produces lasting changes in cortisol and behavior in rhesus monkeys. *Proceedings of the National Academy of Sciences*, **108**(34): 14312−14317.

Gosztolai A, Günel S, Lobato-Ríos V, et al. 2021. LiftPose3D, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nature Methods*, **18**(8): 975−981.

Harlow HF, Suomi SJ. 1971. Production of depressive behaviors in young monkeys. *Journal of Autism and Childhood Schizophrenia*, **1**(3): 246−255.

He KM, Zhang XY, Ren SQ, et al. 2016a. Deep residual learning for image recognition. *In*: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 770–778.

He KM, Zhang XY, Ren SQ, et al. 2016b. Identity mappings in deep residual networks. *In*: Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 630–645.

Hirasaki E, Kumakura H, Matano S. 2000. Biomechanical analysis of vertical climbing in the spider monkey and the Japanese macaque. *American Journal of Physical Anthropology*, **113**(4): 455−472.

Hossain E, Chetty G, Goecke R. 2013. Multi-view multi-modal gait based human identity recognition from surveillance videos. *In*: Proceedings of the 1st IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction. Tsukuba: Springer, 88–99.

Hryniewiecka-Jaworska A, Foden E, Kerr M, et al. 2016. Prevalence and associated features of depression in women with Rett syndrome. *Journal of Intellectual Disability Research*, **60**(6): 564−570.

Joosten AV, Bundy AC, Einfeld SL. 2009. Intrinsic and extrinsic motivation for stereotypic and repetitive behavior. *Journal of Autism and Developmental Disorders*, **39**(3): 521−531.

Karashchuk P, Tuthill JC, Brunton BW. 2021. The DANNCE of the rats: a new toolkit for 3D tracking of animal behavior. *Nature Methods*, **18**(5): 460−462.

Karpathy A, Toderici G, Shetty S, et al. 2014. Large-scale video classification with convolutional neural networks. *In*: Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 1725–1732.

Kay W, Carreira J, Simonyan K, et al. 2017. The kinetics human action video dataset. arXiv preprint arXiv: 1705.06950.

Labuguen R, Matsumoto J, Negrete SB, et al. 2021. MacaquePose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in Behavioral Neuroscience*, **14**: 581154.

Li CX, Yang C, Li YR, et al. 2021a. MonkeyPosekit: automated markerless 2D pose estimation of monkey. *In*: Proceedings of 2021 China Automation Congress. Beijing: IEEE, 1280–1284.

Li WT, Wang QX, Liu X, et al. 2021b. Simple action for depression detection: using kinect-recorded human kinematic skeletal data. *BMC Psychiatry*, **21**(1): 205.

Li YS, Xia RJ, Liu X. 2020. Learning shape and motion representations for view invariant skeleton-based action recognition. *Pattern Recognition*, **103**: 107293.

Li ZY, Gavrilyuk K, Gavves E, et al. 2018. VideoLSTM convolves, attends and flows for action recognition. *Computer Vision and Image*

Understanding, **166**: 41−50.
Lin J, Gan C, Han S. 2019. TSM: temporal shift module for efficient video understanding. *In*: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 7083–7093.
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.
Liu Z, Li X, Zhang JT, et al. 2016. Autism-like behaviours and germline transmission in transgenic monkeys overexpressing MeCP2. *Nature*, **530**(7588): 98−102.
Lo Presti L, La Cascia M. 2016. 3D skeleton-based human action classification: a survey. *Pattern Recognition*, **53**: 130−147.
Ma X, Ma CL, Huang J, et al. 2017. Decoding lower limb muscle activity and kinematics from cortical neural spike trains during monkey performing stand and squat movements. *Frontiers in Neuroscience*, **11**: 44.
Mathis A, Mamidanna P, Cury KM, et al. 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, **21**(9): 1281−1289.
Mendes LST, Manfro GG, Gadelha A, et al. 2018. Fine motor ability and psychiatric disorders in youth. *European Child & Adolescent Psychiatry*, **27**(5): 605−613.
Nath T, Mathis A, Chen AC, et al. 2019. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, **14**(7): 2152−2176.
Ng JYH, Hausknecht M, Vijayanarasimhan S, et al. 2015. Beyond short snippets: deep networks for video classification. *In*: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 4694–4702.
Qin DD, Wu SH, Chen YC, et al. 2019. Behavioral screening tools for identifying autism in macaques: existing and promising tests. *Brain Research Bulletin*, **146**: 87−93.
Qin ZQ, Zhang PY, Wu F, et al. 2020. FcaNet: frequency channel attention networks. arXiv preprint arXiv: 2012.11879.
Ricciardi C, Amboni M, De Santis C, et al. 2019. Using gait analysis' parameters to classify Parkinsonism: a data mining approach. *Computer Methods and Programs in Biomedicine*, **180**: 105033.
Richter CP. 1931. The grasping reflex in the new-born monkey. *Archives of Neurology and Psychiatry*, **26**(4): 784−790.
Sabbe B, Hulstijn W, Van Hoof J, et al. 1996. Fine motor retardation and depression. *Journal of Psychiatric Research*, **30**(4): 295−306.
Shah RR, Bird AP. 2017. MeCP2 mutations: progress towards understanding and treating Rett syndrome. *Genome Medicine*, **9**(1): 17.
Sharma S, Kiros R, Salakhutdinov R. 2015. Action recognition using visual attention. arXiv preprint arXiv: 1511.04119.
Simonyan K, Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos. *In*: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 568–576.
Soomro K, Zamir AR, Shah M. 2012. UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv: 1212.0402.
Sun B, Zhang XY, Liu LZ, et al. 2017. Effects of head-down tilt on nerve conduction in rhesus monkeys. *Chinese Medical Journal*, **130**(3): 323−327.
Tran D, Bourdev L, Fergus R, et al. 2015. Learning spatiotemporal features with 3D convolutional networks. *In*: Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 4489–4497.
Tran TH, Le TL, Hoang VN, et al. 2017. Continuous detection of human fall using multimodal features from Kinect sensors in scalable environment. *Computer Methods and Programs in Biomedicine*, **146**: 151−165.
Van Damme T, Simons J, Sabbe B, et al. 2015. Motor abilities of children and adolescents with a psychiatric condition: a systematic literature review. *World Journal of Psychiatry*, **5**(3): 315−329.
Venkataraman V, Turaga P, Lehrer N, et al. 2013. Attractor-shape for dynamical analysis of human movement: applications in stroke rehabilitation and action recognition. *In*: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland: IEEE, 514–520.
Viher PV, Docx L, Van Hecke W, et al. 2019. Aberrant fronto-striatal connectivity and fine motor function in schizophrenia. *Psychiatry Research:Neuroimaging*, **288**: 44−50.
Vyas S, Rawat YS, Shah M. 2020. Multi-view action recognition using cross-view video prediction. *In*: Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 427–444.
Wang JD, Sun K, Cheng TH, et al. 2021. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(10): 3349−3364.
Wang LM, Xiong YJ, Wang Z, et al. 2016. Temporal segment networks: towards good practices for deep action recognition. *In*: Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 20–36.
Xiao ZF, Liu YQ, Li CX, et al. 2022. Two-stream action recognition network based on temporal shift and split attention. *Computer Systems & Applications*, **31**(1): 204−211. (in Chinese)
Xie SN, Girshick R, Dollár P, et al. 2017. Aggregated residual transformations for deep neural networks. *In*: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 1492–1500.
Yang LJ, Fan YC, Xu N. 2019. Video instance segmentation. *In*: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 5188–5197.
Zhang H, Wu CR, Zhang ZY, et al. 2022. ResNeSt: split-attention networks. *In*: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE.
Zhou Y, Sharma J, Ke Q, et al. 2019. Atypical behaviour and connectivity in *SHANK3*-mutant macaques. *Nature*, **570**(7761): 326−331.