MarmoPose: A Deep Learning-Based System for Real-time Multi-Marmoset 3D Pose Tracking

Chaoqun Cheng[1], Zijian Huang[1], Ruiming Zhang[1], Guozheng Huang[1], Han Wang[1], Likai Tang[1], Xiaoqin Wang[1,2*]

[1]Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China.

[2*]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, US.

***Correspondence**: xiaoqinwang@tsinghua.edu.cn;

Abstract

The common marmoset has become an important experimental animal model in scientific research. The ability to capture and quantify behaviors of marmosets in natural environment and social scenarios is highly desired by marmoset research community. Although existing pose tracking methods have enabled multi-marmoset two-dimensional (2D) tracking and single-marmoset three-dimensional (3D) pose estimation, they have not fully addressed the practical demands of marmoset research. Here, we introduce MarmoPose, a real-time 3D pose tracking system based on deep learning for automatically estimating 3D poses of multiple marmosets freely roaming in their homecage. MarmoPose employs a marmoset skeleton model to optimize 3D pose computation and estimate invisible body locations. Furthermore, MarmoPose provides a real-time processing module to enable short-latency closed-loop experimental control based on the 3D poses of marmosets. The objective of our effort is to design an efficient, cost-effective and user-friendly system that can be easily adapted by marmoset researchers to quantify natural behaviors in typical housing environment.

Keywords

Marmoset, 3D Pose Tracking, Skeleton Model, Closed-loop Control, Behavioral Analysis

Introduction

The common marmoset (Callithrix jacchus) has emerged in recent years as a promising non-human primate model in neuroscience research, offering unique advantages over other animal models. Compared to rodents, marmosets have more complex brain architecture and exhibit closer cognitive ability to humans. Unlike larger primates like macaques, marmosets are easier to breed in captivity and have a shorter developmental stage and faster reproductive cycle [1, 2]. Marmosets have been widely used in various fields of scientific research, including vocal and auditory studies [3–8], visual neuro- science [9–11], and transgenic studies for disease modelling [12–16].

Marmosets are particularly suitable for a wide range of behavioral experiments due to their small body size and social behaviors [17]. However, most behavioral experiments on marmosets have been conducted based on manual recordings or with movement constraints [18–20]. Therefore, the ability to automatically capture and quantify behaviors of marmosets in natural environment and social scenarios is highly desired by marmoset research community. Such a system could be integrated with other experimental methodologies to advance marmoset research. For instance, using quantified behaviors to identify the differences in behavioral phenotypes between normal and genetically modified marmosets could shed light on the relationships between genes and behaviors. In addition, synchronizing behavioral quantification with neural activity recording technologies could potentially reveal the neural mechanisms underlying behaviors.

In recent years, several effective pose tracking systems have been developed to automatically estimate the 2D or 3D poses of animals. For example, DeepLabCut offers multi-animal 2D pose estimation capabilities [21, 22], and can be extended to provide low-latency 3D pose

estimation for a single animal [23–25]. SLEAP is another versatile system supporting multi-animal 2D pose tracking across various network architectures [26, 27]. DANNCE enables direct 3D pose estimation for single rodent and can be transferred to other species [28]. Besides, there are some 3D pose tracking systems designed for one particular species, such as OpenMonkeyStudio for macaques [29], DeepFly3D for drosophila [30] and MAMMAL for pigs[31]. However, although DeepLabCut provides 2D pose tracking for two marmosets and DANNCE could be transferred to estimate 3D poses for single marmoset, their practical usage on marmoset behavioral experiments is fairly limited. Actually, considering the highly social nature of marmosets and their rapid movements in three-dimensional space, a system capable of tracking 3D poses of multiple marmosets is greatly needed. In addition, researchers require the ability to control stimuli of an experiment in real-time based on the marmoset's positions or actions [32], necessitating a tool capable of low-latency closed-loop experimental control. Existing systems are unable to fulfill these specific requirements. In this study, we aim to develop a user-friendly and efficient 3D pose tracking system for multiple marmosets, which can be used by a wide range of researchers to study the marmoset's natural behaviors.

The MarmoPose described in this report is a deep learning-based 3D pose tracking system, with minimum hardware requirement, specifically designed for reconstructing 3D poses of single or multiple marmosets freely moving in their homecage environment. In MarmoPose, multi-view images captured by four cameras are first processed by deep neural networks to predict 2D locations of 16 body locations and identify each marmoset. Subsequently, visible 3D body locations are reconstructed using triangulation while the invisible 3D body locations are estimated

through a variational autoencoder incorporating a marmoset skeleton model. MarmoPose offers several advantages over existing systems: (1) This is the first system enables 3D pose tracking for multiple marmosets, with a comprehensive processing pipeline. (2) This system supports short-latency closed-loop experimental control based on the 3D poses and positions of marmosets, which could be integrated with other experimental functions including stimulus playback and neural recording. (3) This system is designed for user-friendly deployment in a typical marmoset family cage (0.7m x 1m x 0.8m) without additional modifications and can therefore be easily transferred to other housing or experimental environments. (4) This system employs a marmoset skeleton model for 3D coordinates optimization, thereby improving the precision of the reconstructed 3D poses and rendering it possible to estimate invisible body locations in the cameras' blind spots.

Results

Overview of the MarmoPose system

MarmoPose is a 3D pose tracking system specifically designed for both single and multiple marmosets. It receives video streams from multiple cameras and outputs estimated positions of body points of marmosets in a 3D space. We developed the system for a typical marmoset family cage (1m x 0.7m x 0.8m) with four video cameras mounted on the upper corners as shown in Fig. 1a. Marmosets can freely move around in the cage. Two wooden logs and two mental shelves are placed inside the cage to encourage the marmosets exhibit natural behaviors. Note that the cameras are fixed on four top corners inside the homecage so that such a setup doesn't require any modifications of the cage and thus can be easily deployed in other housing cages.

We selected 16 locations of the body to comprehensively represent the posture of a marmoset (head, left/right ear, neck, spinemid, left/right elbow/hand/knee/foot, tailbase, tailmid, tailend). As annotated in Fig. 1b, each dot represents one of the body locations and the lines represent the marmoset skeleton. Invisible body locations from this view are indicated by transparent dots. In scenarios involving multiple marmosets, maintaining consistent identification across different camera views is crucial for accurate 3D triangulation. To ensure reliable identification of individuals, we applied a harmless dye to their ears to clearly distinguish them (Among n marmosets, $n-1$ should be marked by different colors while one remains unmarked). For instance, the marmoset marked with blue dye is correspondingly annotated with blue in Fig. 1b.

Fig. 1c illustrates the workflow of MarmoPose, consisting of three main stages: In the first stage, we employed deep neural networks (modified U-net [33]) building upon the architecture proposed by SLEAP [26, 27] to generate 2D predictions from each camera. For single marmoset analysis, one network is used to directly predict the 2D coordinates of all body locations; for multiple marmosets analysis, we used two networks, the first one to locate the central body locations (spinemid) of each instance within each image frame, and the second one to process cropped images centered around the locations detected by the first network to predict both the body locations and identity of each marmoset. This top- down method was selected due to its flexibility to be easily extended to varying number of marmosets with different identity features. To train the deep neural networks, we labeled Marmoset3K dataset (see Methods), containing 1527 images containing one marmoset and 1646 images containing two marmosets across different cam-

era views. Each image is annotated with the 16 body locations and identity information for each instance. Additionally, a module called identity tracker was applied to further improve the accuracy of identity recognition by correcting misidentified instance using trace information (see Methods). In the second stage, instances with the same identity from all camera views are integrated and triangulated into 3D poses using the intrinsic and extrinsic camera parameters, as illustrated by Fig. 1d. In the third stage, we used a variational autoencoder (VAE) [34] integrated with prior knowledge of marmoset skeleton model which is measured manually to reconstruct the invisible body locations, and then an iterative optimization method to refine the 3D poses further (see Methods). As a result, MarmoPose can accurately estimate 3D poses of a marmoset with invisible body locations. To help visualize the estimated 3D poses, MarmoPose also generates videos combining images from camera views with predicted body locations and identity information, as illustrated in Fig. 1e.

In addition, MarmoPose also provides a real-time process module to enable low-latency experimental control based on the 3D poses of marmosets. This module will be discussed in detail in the fourth section below.

MarmoPose can accurately track 3D poses of multiple marmosets

In order to evaluate the accuracy of MarmoPose, we constructed a dataset referred to as Marmoset3D (see Methods) by triangulating hand-labeled 2D coordinates from multiple camera views at the same timepoint. The Marmoset3D dataset contained 522 3D ground-truth instances with 8352 body locations (16 body locations per each instance), consisting of 140 instances

collected from 140 timepoints with single marmoset (from four marmosets) and 382 instances collected from 191 timepoints with paired marmosets (from three pairs of marmosets). The hand-labeled 2D coordinates were first annotated by one person and then proofread by a second person to ensure accuracy. However, because marmosets are relatively small in each camera view and there were no obvious landmarks on their bodies for precise localization, there was inevitably variability in the 2D coordinates annotated by the two persons. Theoretically, this human variability is the lower bound of the error of MarmoPose. We computed the human variability by measuring the error between hand-labeled ground-truth on the same data from two different persons. Fig. 2a shows the 3D Euclidean errors of MarmoPose and human variability, in which 16 body locations are grouped into 4 categories based on their anatomical locations for clarity (Head: head, left/right ear; Spine: neck, spinemid; Limbs: left/right elbow/hand/knee/foot; Tail: tailbase, tailmid, tailend). MarmoPose can achieve comparable 3D errors to the human variability, with median error of 4 mm for Head, 10mm for Spine, 17 mm for Limbs and 12 mm for Tail, respectively.

Fig. 2b shows the 3D reconstruction accuracy as a function of error threshold for both training data and test data for paired marmosets, where the accuracy is defined as the percentage of 3D body locations whose 3D Euclidean errors are below the error threshold, with 16 body locations grouped into four categories. The typical body length of an adult marmoset is about 20cm (excluding the tail, which is also approximately 20cm long). If we choose 20 mm (about 10% of body length) as a threshold, the accuracy of MarmoPose is 86% for Head, 80% for Spine, 58% for Limbs and 65% for Tail respectively on the test data (Fig. 2b). The accuracy increases as the

threshold criterion increases. Note that the accuracy still can't reach 100% even with the largest threshold analyzed. This is because the identities of marmosets might be incorrect in some cases, resulting in large errors after 3D triangulation. To intuitionally display the 3D reconstruction errors, we selected two representative images and drew the 75th percentile of the 3D Euclidean errors for each body location on them (Fig. 2c, e) with full error distribution (Fig. 2d, f). Generally, body locations on the head have smaller errors due to the existence of boundary between brown and white hair, while body locations on the limbs have larger errors due to high degrees of freedom and frequent occlusion. This result was evaluated in the current setup using four cameras on a typical marmoset homecage, and the accuracy might vary with changes in the environment and settings.

Use trace information to improve identification accuracy

Maintaining consistent identification of the same instance across different camera views is crucial for accurate 3D reconstruction. If the unique identity information is occluded by other body locations or items inside the cage, the identities of marmosets recognized by deep neural networks might be incorrect. To address this issue, we employed a tracking method using the trace information to correct the identities of marmosets recognized by neural networks. As illustrated by Fig. 3a, misidentification occurred in the last frame due to occlusion of the dyed ears, while the instances in the previous frames are recognized correctly. Therefore, we can use the trace information computed through positional similarity to correct the identities. We evaluated the identification accuracy with and without the identity tracker on paired marmosets dataset across

11 video clips (n=1666 frames). This tracking method improved the accuracy by 1% percent (ID model, 94.17%; ID model+tracker: 95.06%, Fig. 3b).

Estimate invisible body locations using variational autoencoder

In the scenarios involving multiple marmosets freely moving in the homecage, some of their body locations will inevitably be self-occluded or occluded by other animals and objects like logs or shelves in the cage, leading to missing data in the reconstructed 3D poses. Inspired by previous study on human [35], we developed a variational autoencoder (VAE) [34] to reconstruct the missing 3D body locations, which receives 3D coordinates of 16 body locations with missing data as input and outputs the estimated complete coordinates. We trained the VAE by randomly masking some body locations to simulate inputs with missing data based on the Marmoset3D dataset (see Methods).

We added a loss term constraining the limb length of reconstructed 3D poses to guide the VAE to estimate missing coordinates better. The limb length is a prior knowledge of marmosets. We measured the median length of distances between two body locations on two normal adult marmosets to form a skeleton model (Fig. 3c). Incorporating prior knowledge of marmosets into the VAE enhances its ability to capture the underlying structure of marmosets. As a result, the VAE can more accurately reconstruct the missing body locations, even with a limited amount of training data. The estimated missing data have a median error of 20mm for Head, 35mm for Limbs and 50mm for Tail. Note that, Spine ground has no invisible body locations because they are the center parts used to normalized the poses. Although missing data predicted by VAE might not

match the accuracy of visible body locations, it still offers a reasonable estimation of the missing body locations under fixed experimental environment.

Real-time experimental control based on 3D poses

One objective of the MarmoPose is to provide a real-time process module to enable real-time experimental control by analyzing images from multiple live video streams and output 3D poses frame by frame with a small latency. To achieve this goal, we adopted multi-processing and multi-threading to deal with different tasks in parallel. As the workflow chart shown in Fig. 4a, in one process, live video streams are first cached by multiple threads, then 2D detection and 3D reconstruction are performed on the latest cached frame. In the other process, poses and images are combined together for display and an interface returning the latest poses is provided for customized detection and corresponding experimental control.

The system latency mainly consists of three parts: 2D detection, 3D reconstruction and inherent processes such as data transfer and display, where 2D detection is the most time-consuming part. To speed up the 2D detection, we adopted image downsampling and dynamic cropping to reduce the computational cost. The original image size is 1920 × 1080 × 3, where marmosets only occupy a small part of the image, thus it is unnecessary to perform detection on the whole image. Cropped images based on the marmoset's positions in the previous frame could be used as the input of the current frame with an offset, because the marmoset's position in consecutive frames will not differ significantly due to spatial continuity. We tested the system latency under different cropping size and downsampling scale (Fig. 4b) and found that when the

input image size was 384 × 384, the system could achieve real-time closed-loop feedback with a median latency of 40ms. In the scenarios of multiple marmosets, the time consumed by the 2D detection is longer than single marmoset because the original image needs be first passed to a network to locate the central body locations of each instance. The time consumed by each part for both scenarios are shown in Fig. 4c. For single marmoset, the median system latency is 40ms which is less than the frame interval (40ms); for multiple marmosets, the median system latency is 110ms (<3 frames), which can still be used for some real-time experimental control as long as it meets the minimum latency requirement.

We tested this system in a practical experiment in which we want to detect the head orientation of a marmoset and provide real-time feedback to control the stimulus. Poses of the marmoset at different timepoints are shown in Fig. 4d. The marmoset looks forward at 40ms and then turns to left at 80ms, MarmoPose can reliably track such quick movements in real-time, thereby providing possibilities for real-time experimental control.

Discussion

We described here MarmoPose that enables real-time 3D pose tracking for multiple freely moving marmosets in a housing cage situation. The primary considerations in designing the MarmoPose is to make it an efficient, cost effective (using minimum hardware), real-time and easily deployable in most housing environment. This system is specifically designed for marmosets in order to optimize its performance. It has several distinct features comparing to other published pose detection systems for freely moving animals. First, MarmoPose leverages the prior

knowledge of marmoset skeleton model to estimate invisible body locations and refine 3D poses. Second, MarmoPose is user-friendly and can be deployed in a typical marmoset homecage environment. Third, MarmoPose provides a real-time process module for users to perform short-latency closed-loop experimental control based on the 3D poses of marmosets, which can be integrated with other experimental methodologies such as stimulus playback and neural recording.

In scenarios involving multiple marmosets, particularly when they are in close proximity, distinguishing individual body locations accurately is a challenging task. This is a common problem in the field of multi-object pose tracking, which has not been addressed very well. Since each marmoset occupies only a small part of the whole image, and their rapid movements often lead to motion blur, it is challenging even for human annotators to accurately locate specific body locations. Consequently, MarmoPose might not perform optimally in such scenarios. A practical solution is to keep a balance between specific experimental demands and system accuracy. In many experimental situations, the focus might be only on a subset of body locations. Therefore, we can choose to ignore the misidentified data and concentrate on the relevant body locations. Additionally, in social scenarios, the overall interaction of the marmosets might be more significant than the precise poses of each body locations. Therefore, it might be more effective to treat marmosets that are close to each other as a single entity without attempting to distinguish every body part clearly.

In the current setup, four cameras are fixed on the upper corners inside the homecage to maximize coverage. This is a simple and effective arrangement, which can be easily deployed in other housings without extra modifications, and keep the minimum computational cost at the same

time. While adding more cameras to cover more blind spots is a straightforward solution to reduce invisible body locations and improve the system accuracy further, it is not always necessary. In fact, the number of cameras used should be determined by the specific experimental demands. The current arrangement with four cameras is sufficient for a range of research purposes. For researchers aiming for higher precision, more cameras can be added to cover more blind spots, which is supported by MarmoPose. However, it is important to consider that increasing the number of cameras would lead to longer processing time and output latency.

Currently, the overall system latency of real-time process module is less than 40ms for single marmoset and around 110ms for multiple marmosets. These latencies should be adequate for conducting real-time behavioral analyses. If an experiment demands shorter latencies, one can consider using a subset of 16 body locations speed up the processing. Additionally, skipping some of frames can be another effective strategy, as long as it meets the experimental requirements.

The ability to track poses of freely moving marmosets in 3D space automatically in real-time using MarmoPose could significantly improve studies of natural behaviors in this field. Traditionally, studies on vocal communications [4, 37] and visual directional preferences [9–11] have been relying on manually processed audio and video recordings. MarmoPose could be used to boost these studies by providing more accurate and comprehensive behavioral quantifications. Give their social nature and ease of handling, marmosets are ideal for comparative behavioral experiments with humans. For example, MarmoPose could be used to describe and compare the behavioral evolution and natural preferences between marmosets and humans [1, 17, 38, 39]. Moreover, due to their relatively high reproductive cycles compared to other primates, marmosets are well-

suited for transgenic modifications [14, 40] and disease modeling [41–44]. Integrating MarmoPose with other methodologies such as neural recording and magnetic resonance imaging (MRI), offers possibilities to explore neural mechanisms underlying natural behaviors in both normal and genetically modified marmosets.

Methods

Datasets

We collected videos for single marmoset and paired marmosets to develop and evaluate MarmoPose. Videos were all recorded by four synchronized cameras (HIKVISION DS-2CD252CZY-ZFR) mounted on the upper corners of a typical marmoset homecage (1m x 0.7m x 0.7m) with 1920 × 1080 × 3 frame size at 25 fps. Two wooden logs and two mental shelves were placed in the homecage to encourage the marmosets to engage in naturalistic behaviors. We used tools provided by SLEAP [26, 27] to label the ground truth, including 2D locations of 16 body locations for each marmoset and their identities when multiple marmosets are in the same frame.

All of the marmosets are healthy adults with normal body size and weight, and they have been lived in the same homecage for more than half a year. All of the procedures in these experiments were approved by the Science and Technology Ethics Committee at Tsinghua University.

Marmoset3K - Single marmoset. Two human annotators labeled 1527 frames (1527 instances) from 4 different camera views across 16 videos capturing a single marmoset freely

moving in the homecage. Four adult marmosets with- out additional modifications were used in this dataset. For each marmoset instance, 16 body locations were labeled (head, leftear, rightear, neck, spinemid, leftelbow, lefthand, rightelbow, righthand, leftknee, leftfoot, rightknee, rightfoot, tail- base, tailmid, tailend) carefully. If a body part is occluded from a camera view, it is labeled as invisible and will not be included in the training and testing process. To train the 2D pose predictor, we randomly selected 1270 frames as training data and the other 257 frames as test data.

Marmoset3K - Paired marmosets. Three human annotators labeled 1646 frames (3292 instances) from 4 different camera views across 12 videos capturing a pair of marmosets freely moving in the homecage. Six (three pairs) adult marmosets were used in this dataset. Since the consistence of marmosets' identities across different camera views is essential for accurate 3D triangulation, we dyed the ears of one marmoset in each pair with harmless blue color to significantly distinguish them. For each pair of marmosets, the one dyed blue was annotated ID '1' and the other one with normal white ears was annotated ID '2', and each instance was also labeled 16 body locations. To train the 2D pose and identity predictor, we randomly selected 1354 frames as training data and the other 292 frames as test data.

Marmoset3D. Ground truth of 3D poses are required in training VAE and evaluating the accuracy of MarmoPose, which could be obtained by triangulating precisely hand-labeled 2D coordinates from multiple camera views at the same timepoint. To establish the Marmoset3D dataset, three human annotators labeled 522 3D groundtruth, consisting of 140 instances triangulating from 560 images containing single marmoset and 382 instances triangulating from

191 images containing paired marmosets. Two annotators labeled the images first and then one annotator proofread the labels to ensure the accuracy.

Pose and identity predictor

Different neural networks are used to detect the 2D pose and identity for single marmoset and multiple marmosets, both are modified on the basis of models provided by SLEAP[27].

For single marmoset, we constructed a U-net like neural network consisting of 6 downsampling convolutional blocks and 4 upsampling convolutional blocks. Each downsampling block contains two convolutional layers with 3x3 kernel size and two relu activation layers, followed by a maxpooling layer with 2x2 kernel size, where the first block has 64 channels and the number of channels 1.5x increases in the following blocks. Correspondingly, each upsampling block has the same structure as the downsampling block expect the channels are 1.5x decreases and the maxpooling layer is replaced by a transposed layer with 2x2 kernel size. The original image size is 1080 × 1920 × 3 but rescaled to 0.8x smaller (896 × 1536 × 3) as the input of the network, and the output size is 224 × 384 × 16 with 16 channels corresponding to 15 body locations. Then the output is upsampled to the original size as the final prediction. Skip connections are used to connect the downsampling and upsampling blocks to provide spatial context. The network is trained with a batch size of 4 for 200 epochs using Adam optimizer with initial learning rate of 0.0001 and early stop for 10 epochs. Data augmentation is applied to the training data, including rotation between -180 to 180 degrees, note that flipping is note used because it will change the left/right body locations.

For multiple marmosets with identity, we used top-down approach containing two networks, the first network detected the 'spinemid' body part of each marmoset within the full-resolution image (1080 × 1920 × 3), and then cropped the image around the 'spinemid' body part with a fixed size (640 × 640 × 3) as the input of the second net- work. Note that the cropped image is centered on one instance but might contain pixels belonging to other marmosets, this centering is crucial as it provides spatial context to the second stage. The second network is the almost same as the single marmoset network except that an extra head consisting of 3 full-connected layers with 64 channels for predicting the identities of marmosets are added after the deepest encoder layer, and the weights of loss for 2D pose and identity are 999 and 0.01 respectively.

Additional identity tracker

Although the identities of marmosets have been predicted by the 2D predictor, sometimes it might be incorrect especially when the dyed ears are occluded by other stuffs. Thus we employed an addition identity tracker [27] to improve identification accuracy further using trace information. Specifically, we extracted candidates with high confidence from a number of previous frames (default 5 frames), then computed the positional similarity between each instance in the current frame and each candidate, in which the similarity is computed by the sum of kernel density estimation (KDE) of distances between corresponding body locations. With the similarity matrix, we performed optimal matching and re-assign identities based on the matching results.

Camera calibration

For camera calibration, we used the model and methods provided by Anipose [24, 45] to estimate the intrinsic and extrinsic parameters of multiple cameras. The calibration process involved placing a standard checkerboard pattern (11 × 9 squares with 45mm square size in this study) in the homecage and rotating it to various angles, then the recorded videos from different camera views are used for calibration.

In order to align the default coordinate system with actual spatial dimensions, we developed a labeling tool for users to define new coordinate system. This tool allows users to define a new coordinate system by marking three specific points in at least two camera views: the origin point, a point on the x-axis, and a point on the y-axis, then these points will be triangulated into 3D coordinates and cameras' extrinsic parameters will be updated to fit the new coordinate system.

Triangulation

We used random sample consensus (RANSAC) triangulation to triangulate the 3D coordinates [24, 46], which can reduce the influence of outlier 2D detections by finding a subset of detected 2D points that minimizes the reprojection error. This is achieved by triangulating all possible combinations of 2D points and then selecting the one with lowest reprojection error.

Pose normalization

The reconstructed 3D poses are in the real spatial space where similar poses might have totally different coordinates, thus it is essential to normalize them into a unified coordinates system for effective VAE model training and subsequent analysis. We used three body locations to define the new coordinates system: the middle of spine ('spinemid') as the original point; the upper end of spine ('neck') defining the x-axis; the lower end of spine ('tailbase') lying on the x-z plane. Then the original 3D poses are rotated into this unified coordinate system with standardize pose lengths by dividing each 3D coordinate with the distance between 'spinemid' and 'neck'.

Variational autoencoder

Inspired by a pose estimation work on human [35], we adopted variational autoencoder to fill in the missing data caused by occlusion, which receives incomplete 3D poses as input and output the predicted complete 3D poses. Specifically, each pose is represented by a matrix $M_{n\times 3}$, where n is the number of body locations (n = 16 by default), and the elements are the 3D coordinated of each body part in the real spatial space. We first normalized them into a unified coordinated system, and then flatten the matrix into a vector $V_{1\times 3n}$ as the input of the VAE. With the Marmoset3D dataset, we generated incomplete 3D poses by randomly masking i ($1 \leq i \leq 4$) body locations to simulate the occlusion in real scenarios. These masked 3D poses and corresponding complete 3D poses are used as input and output of the VAE respectively during training.

In this study, the encoder of VAE is com- posed of 2 fully connected hidden layers, with 128 and 256 hidden units respectively. Symmetrically, the decoder also consists of 2 fully

connected hidden layers with 256 and 128 hidden units. We used 40 latent dimensions to represent the input. The loss function consists of three parts: MSE calculating the difference between ground truth and reconstructed poses; D_{KL} measuring the difference between the latent distribution and the standard normal distribution; and joint loss constraining the length between some body locations. Joint loss incorporates prior knowledge of the marmoset skeleton model into the model for guiding the reconstruction of missing body locations, making the VAE work better with few training data. The skeleton model is constructed by measuring the median length of distances between two body locations which are relatively fixed on adult marmosets, and each body part pair will be assigned with different weights based on their degrees of freedom. The VAE is trained with the Adam optimizer [47] with a learning rate of 10^{-3}.

Dynamic cropping

Dynamic cropping is employed to reduce the latency in real-time detection, involving using cropped images based on the previous positions of the marmosets. Specifically, customized real-time predictor stores the positions of each instance in the last timepoint, then generates cropped images with a user-defined size (640 × 640 × 3 by default) centered around these positions. These images sever as the input of the detection model at the current timepoint, with an offset to get the final coordinates. This method works due to the spatial continuity of the marmosets, i.e., the positions of marmosets in consecutive frames will not differ significantly thus the cropped images can still contain the marmosets.

Real-time control module

The real-time control module in MarmoPose consists of two processes: (1) Prediction process. It reads the latest images from multiple live video streams cached by separate threads, then perform 2D detection and 3D triangulation and transmit the results to the main process. Customized predictors are used which stores the previous positions of marmosets for dynamic cropping. (2) Main process. It reads poses and images from the prediction process and display the real-time results, and also provides an interface for users to perform customized event detection and corresponding experimental control. An interface, receiving the latest 2D and 3D poses of marmosets, is provided in the real-time module, allowing users to perform customized event detection and corresponding experimental control. In the scenario of head orientation detection, we first got the head orientation of the marmoset by computing the vector from the middle of 'left ear' and 'right ear' to 'head' in the real space, then we defined the head orientation as 'left' if the angle between the head orientation vector and the normal vector of left side of the cage (i.e. x-z plane in the default coordinate system) is less than 45 degrees, and 'forward' if the angle is between 45 and 135 degrees, and 'right' if the angle is larger than 135 degrees.

Cameras produced live video streams at 25fps with 1920×1080×3 frame size, and frames are read by Real-Time Stream Protocol (RTSP). Latencies were evaluated on a computer with Intel Core i7- 13700K CPU, NVIDIA GeForce GTX 4090 GPU and 64GB Ram.

Acknowledgment

This research was supported by Tsinghua University Non-human Primate Research Center (THU-NPRC). We thank Dr. Li Luo and veterinary and husbandry staff at THU-NPRC for help with this research.

Author Contributions

C.C. and X.W. designed the study. C.C. developed the system. Z.H. helped develop the setup for video recording. C.C., R.Z., G.H, H.W. and L.T. labeled the dataset for model training and test. C.C. and X.W. wrote the manuscript.

Declaration of interests

The authors declare no competing interests.

Reference

1. Prins, N. W. *et al.* Common marmoset (callithrix jacchus) as a primate model for behavioral neuroscience studies. *Journal of neuroscience methods* **284**, 35–46 (2017).
2. Schiel, N. & Souto, A. The common marmoset: an overview of its natural history, ecology and behavior. *Developmental Neuro- biology* **77**, 244–262 (2017).
3. Miller, C. T. & Wang, X. Sensory-motor interactions modulate a primate vocal behavior: antiphonal calling in common marmosets. *Journal of comparative physiology A* **192**, 27–38 (2006).
4. Miller, C. T., Beck, K., Meade, B. & Wang, X. Antiphonal call timing in marmosets is behaviorally significant: interactive play- back experiments. *Journal of Comparative Physiology A* **195**, 783–789 (2009).
5. Miller, C. T., Mandel, K. & Wang, X. The communicative content of the common marmoset phee call during antiphonal calling. *American journal of primatology* **72**, 974–980 (2010).
6. Agamaite, J. A., Chang, C.-J., Osmanski, M. S. & Wang, X. A quantitative acoustic analysis of the vocal repertoire of the common marmoset (callithrix jacchus). *The Journal of the Acoustical Society of America* **138**, 2906–2928 (2015).
7. Gultekin, Y. B., Hildebrand, D. G., Hammer schmidt, K. & Hage, S. R. High plasticity in marmoset monkey vocal development from infancy to adulthood. *Science Advances* **7**, eabf2938 (2021).
8. Osmanski, M. S. & Wang, X. Perceptual specializations for processing species-specific vocalizations in the common marmoset (callithrix jacchus). *Proceedings of the National Academy of Sciences **120**, e2221756120 (2023).*
9. de Boer, R. A., Overduinde Vries, A. M., Louwerse, A. L. & Sterck, E. H. The behavioral context of visual displays in common marmosets (callithrix jacchus). American Journal of Primatology 75, 1084–1095 (2013).
10. Mitchell, J. F., Reynolds, J. H. & Miller, C. T. Active vision in marmosets: a model system for visual neuroscience. Journal of Neuroscience 34, 1183–1194 (2014).
11. Mitchell, J. F. & Leopold, D. A. The marmoset monkey as a model for visual neuroscience. Neuroscience research 93, 20–46 (2015).
12. Han, H.-J., Powers, S. J. & Gabrielson, K. L. The common marmoset—biomedical research animal model applications and common spontaneous diseases. Toxicologic pathology 50, 628–637 (2022).
13. Perez-Cruz, C. & de Dios Rodriguez-Callejas, J. The common marmoset as a model of neurodegeneration. Trends in Neurosciences (2023).
14. Bert, A., Abbott, D. H., Nakamura, K. & Fuchs, E. The marmoset monkey: a multi-purpose preclinical and translational model of human biology and disease. Drug

discovery today 17, 1160–1165 (2012).

15. Okano, H., Hikishima, K., Iriki, A. & Sasaki, E. The common marmoset as a novel animal model system for biomedical and neuroscience research applications. Seminars in Fetal and Neonatal Medicine 17, 336–340 (2012).

16. Okano, H. & Mitra, P. Brain-mapping projects using the common marmoset. Neuroscience Research 93, 3–7 (2015).

17. Miller, C. T. et al. Marmosets: a neuroscientific model of human social behavior. Neuron 90, 219–233 (2016).

18. Burkart, J. & Heschl, A. Geometrical gaze following in common marmosets (callithrix jacchus). Journal of Comparative Psychology 120, 120 (2006).

19. Spadacenta, S., Dicke, P. W. & Thier, P. Reflexive gaze following in common marmoset monkeys. Scientific Reports 9, 15292(2019).

20. Pandey, S., Simhadri, S. & Zhou, Y. Rapid head movements in common marmoset monkeys. Iscience 23, 100837 (2020).

21. Mathis, A. et al. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Nat Neuroscience 21, 1281–1289 (2018).

22. Lauer, J. et al. Multi-animal pose estimation, identification and tracking with deeplabcut. Nature Methods 19, 496–504 (2022).

23. Nath, T. et al. Using deeplabcut for 3d markerless pose estimation across species and behaviors. Nat Protocols 14, 2152–2176(2019).

24. Karashchuk, P. et al. Anipose: A toolkit for robust markerless 3d pose estimation. Cell Reports 36, 109730 (2021).

25. Kane, G. A., Lopes, G., Saunders, J. L., Mathis, A. & Mathis, M. W. Real-time, low-latency closed-loop feedback using markerless posture tracking. eLife 9, e61909 (2020).

26. Pereira, T. D. et al. Fast animal pose estimation using deep neural networks. Nature methods 16, 117–125 (2019).

27. Pereira, T. D. et al. Sleap: A deep learning system for multi-animal pose tracking. Nature Methods 19, 486–495 (2022).

28. Dunn, T. W. et al. Geometric deep learning enables 3d kinematic profiling across species and environments. Nat Methods 18, 564–573(2021).

29. Bala, P. C. et al. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. Nature Communications 11, 4560 (2020).

30. Günel, S. et al. Deepfly3d, a deep learning- based approach for 3d limb and appendage tracking in tethered, adult drosophila. Elife 8, e48571 (2019).

31. An, L. et al. Three-dimensional surface motion capture of multiple freely moving pigs using mammal. Nature Communications 14, 7727 (2023).

32. Kane, G. A., Lopes, G., Saunders, J. L., Mathis, A. & Mathis, M. W. Real-time, low-

latency closed-loop feedback using markerless posture tracking. eLife 9, e61909 (2020).

33. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. arXiv preprint arXiv:1505.04597 (2015).

34. Kingma, D. P. & Welling, M. Auto- encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).

35. Carissimi, N., Rota, P., Beyan, C. & Murino,V. ECCV (ed.) Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. (ed.ECCV) Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 0–0 (2018).

36. Akaike, H. in Information theory and an extension of the maximum likelihood principle (ed.Emanuel Parzen, K. T. . G. K.) Selected papers of hirotugu akaike 199–213 (Springer, 1998).

37. Eliades, S. J. & Miller, C. T. Marmoset vocal communication: behavior and neurobiology. Developmental neurobiology 77, 286–299 (2017).

38. Tardif, S. D., Mansfield, K. G., Ratnam, R., Ross, C. N. & Ziegler, T. E. The marmoset as a model of aging and age-related diseases. ILAR journal 52, 54–65 (2011).

39. Pereira, T. D., Shaevitz, J. W. & Murthy, M. Quantifying behavior to understand the brain. Nature neuroscience 23, 1537–1549 (2020).

40. Sasaki, E. et al. Generation of transgenic non- human primates with germline transmission. Nature 459, 523–527 (2009).

41. Kobayashi, Y. et al. Pre-evaluated safe human iPSC-derived neural stem cells promote functional recovery after spinal cord injury in common marmoset without tumorigenicity. PloS one 7, e52787 (2012).

42. Carrion Jr, R. & Patterson, J. L. An animal model that reflects human disease: the common marmoset (callithrix jacchus). Current opinion in virology 2, 357–362 (2012).

43. Anwar Jagessar, S. et al. The different clinical effects of anti-BLyS, anti-APRIL and anti-CD20 antibodies point at a critical pathogenic role of γ-herpesvirus infected b cells in the marmoset eae model. Journal of neuroimmune pharmacology 8, 727–738 (2013).

44. Smith, D., Trennery, P., Farningham, D. & Klapwijk, J. The selection of marmoset monkeys (callithrix jacchus) in pharmaceutical toxicology. Laboratory Animals 35, 117–130(2001).

45. Zhang, Z. Y. A flexible new technique for camera calibration. Ieee Transactions on Pattern Analysis and Machine Intelligence 22, 1330–1334 (2000).

46. Sheshadri, S., Dann, B., Hueser, T. & Scherberger, H. 3d reconstruction toolbox for behavior tracked with multiple cameras. Journal of Open Source Software 5, 1849 (2020).

47. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

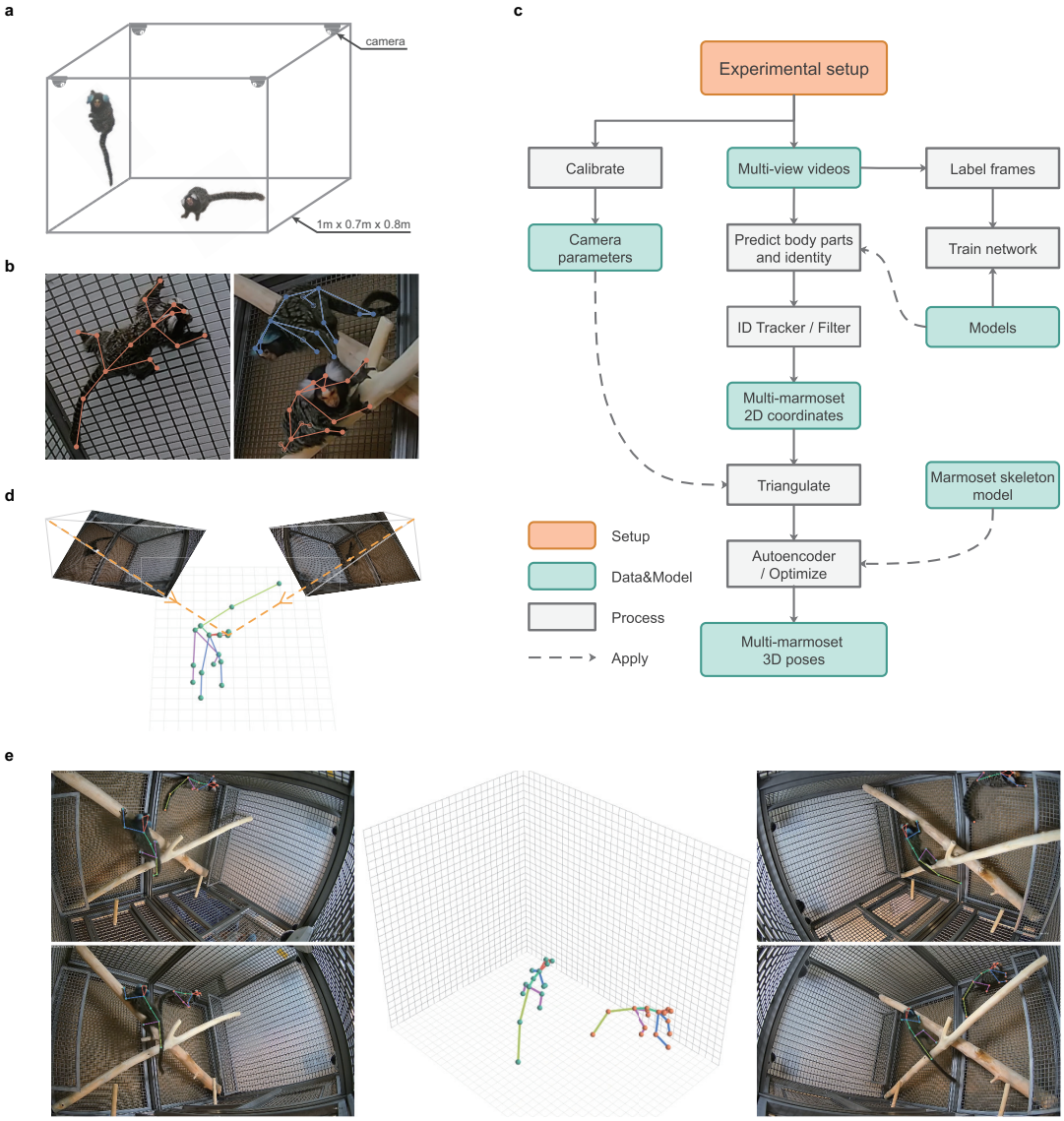

Fig. 1 MarmoPose is a complete system for multi-marmoset 3D pose tracking. a, Experimental setup of MarmoPose. Four cameras are mounted on the upper corners of the marmoset homecage (1m x 0.7m x 0.8m), and two wooden logs and two mental shelves are placed in the cage as their daily environment. Marmosets can freely move without any interference. **b,** Example (cropped) images with manual annotations for the Marmoset3K dataset. Each dot represents one of the 16 body locations (transparent dots denote invisible body locations from this camera view), and the lines indicate the marmoset skeleton. **c,** Diagram of the whole workflow of MarmoPose. **d,** Illustration of 3D triangulation with 2D predictions. By extending the lines from cameras' focal points to the predicted 2D points from more than 2 cameras, a unique spatial position in the 3D space can be pinpointed. **e,** An example demonstrating one frame of the visualized videos generated by MarmoPose. Four corners show images from 4 camera views with predicted 2D body locations and identities, while the central part shows the reconstructed 3D poses.

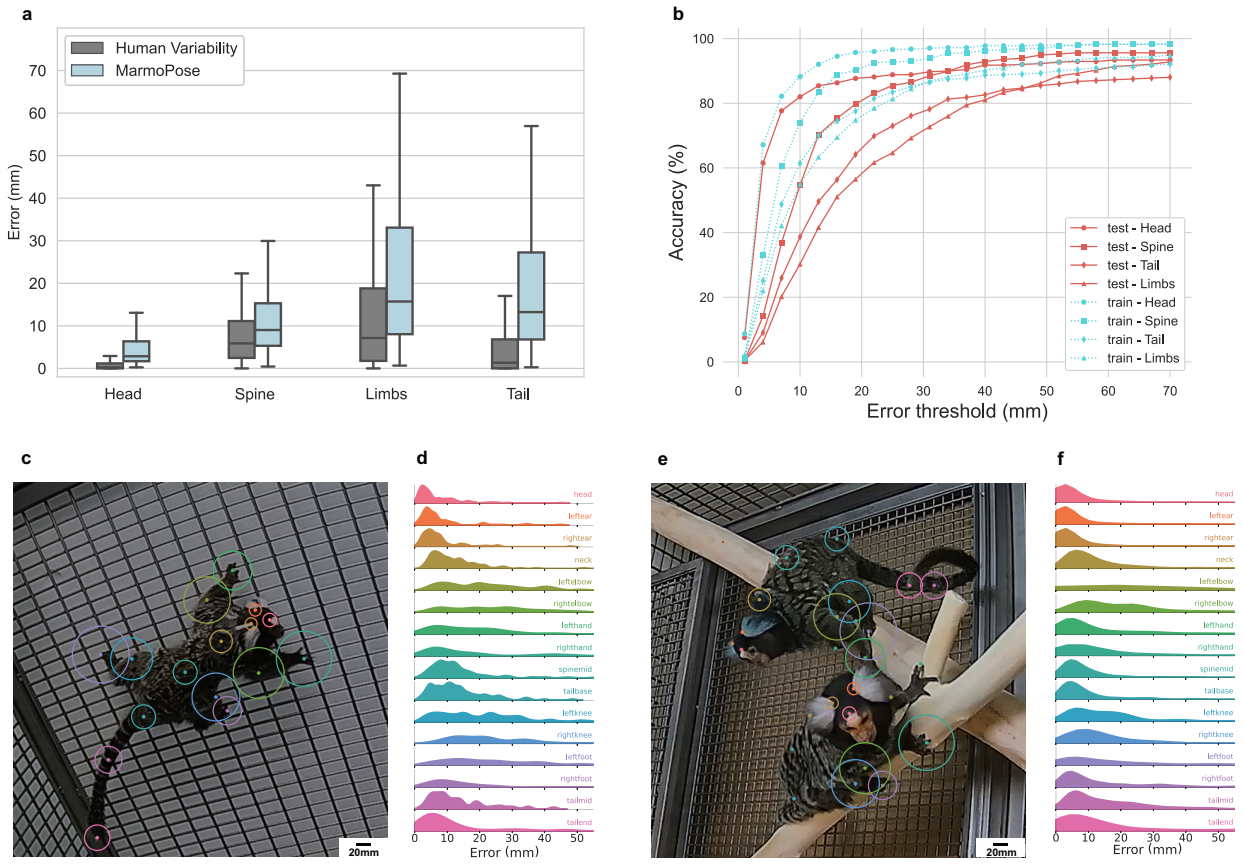

Fig. 2 MarmoPose can accurately track 3D poses of multiple marmosets. a, Box plots of 3D Euclidean errors for MarmoPose and human variability evaluated on the Marmoset3D test dataset, 16 body locations are grouped into 4 categories based on their anatomical locations (Head: head, left/right ear; Spine: neck, spinemid; Limbs: left/right elbow/hand/knee/foot; Tail: tailbase, tailmid, tailend). Human variability indicates the error between hand-labeled ground-truth on the same data from two different people. (n=152 instances). **b**, 3D reconstruction accuracy as a function of error threshold evaluated on part of the training data (n=230 instances) and test data (n=152 instances) for paired marmosets. Body locations are broken down to the same categories as in **a**. **c-f**, 3D reconstruction errors for single marmoset and paired marmosets. Each dot represents one of the body locations, with surrounding circles denoting the 75th percentile of the 3D Euclidean errors, and histograms correspond to full error distribution evaluated on the test data (n=84 instances for single marmoset dataset; n=146 instances for paired marmosets dataset)

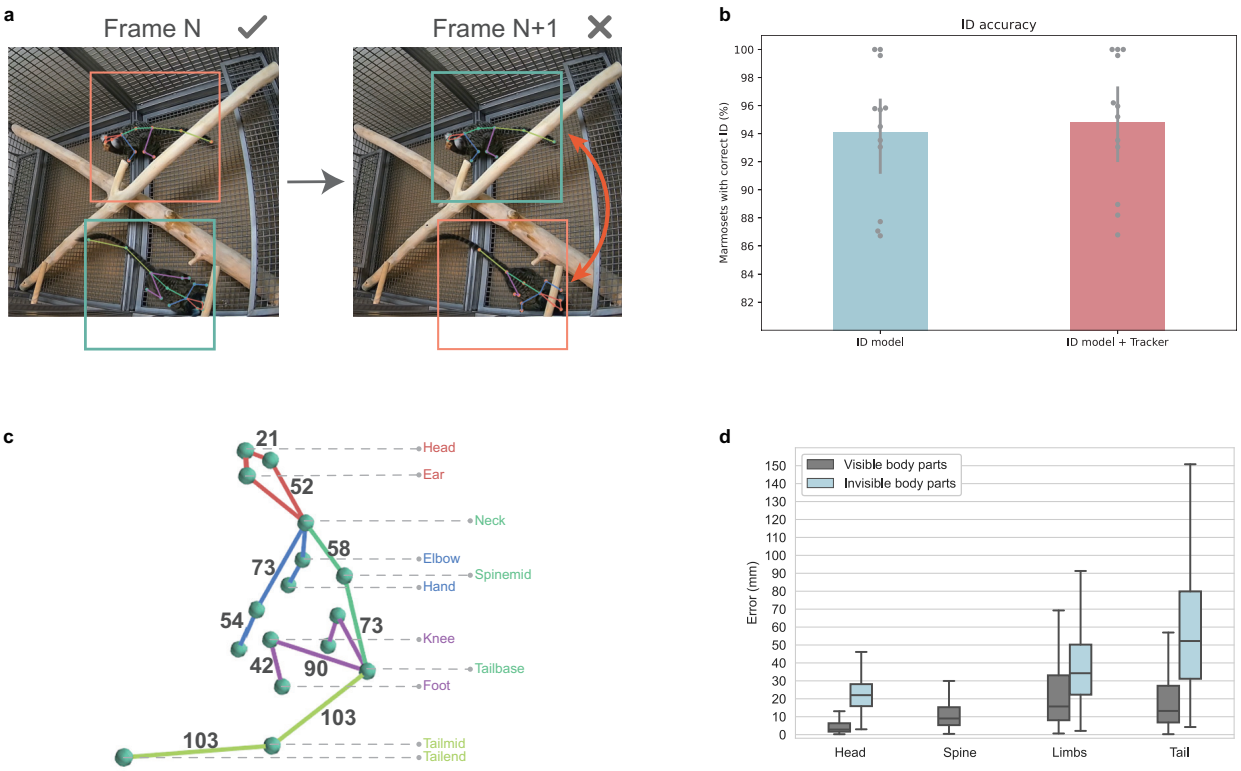

Fig. 3 MarmoPose employs additional methods to improve the accuracy. a, Schematic of using trace information to correct the identities of marmosets. **b**, Identification accuracy with and without identity tracker evaluated on paired marmosets dataset across 11 videos (n=1666 frames). Bars and error whiskers correspond to mean and 95% confidence intervals. **c**, Illustration of the marmoset skeleton model, used as one of the terms in the loss function of VAE to guide the reconstruction of missing body locations. Numbers indicating the median length of the distances between two body locations measured on real marmosets, with different weights assigned based on their degrees of freedom. **d**, Box plots of 3D Euclidean errors in test data for visible and invisible body locations, 16 body locations are grouped into 4 categories based on their anatomical locations. Note that 'Spine' has no invisible body locations because they are used to normalize the poses.

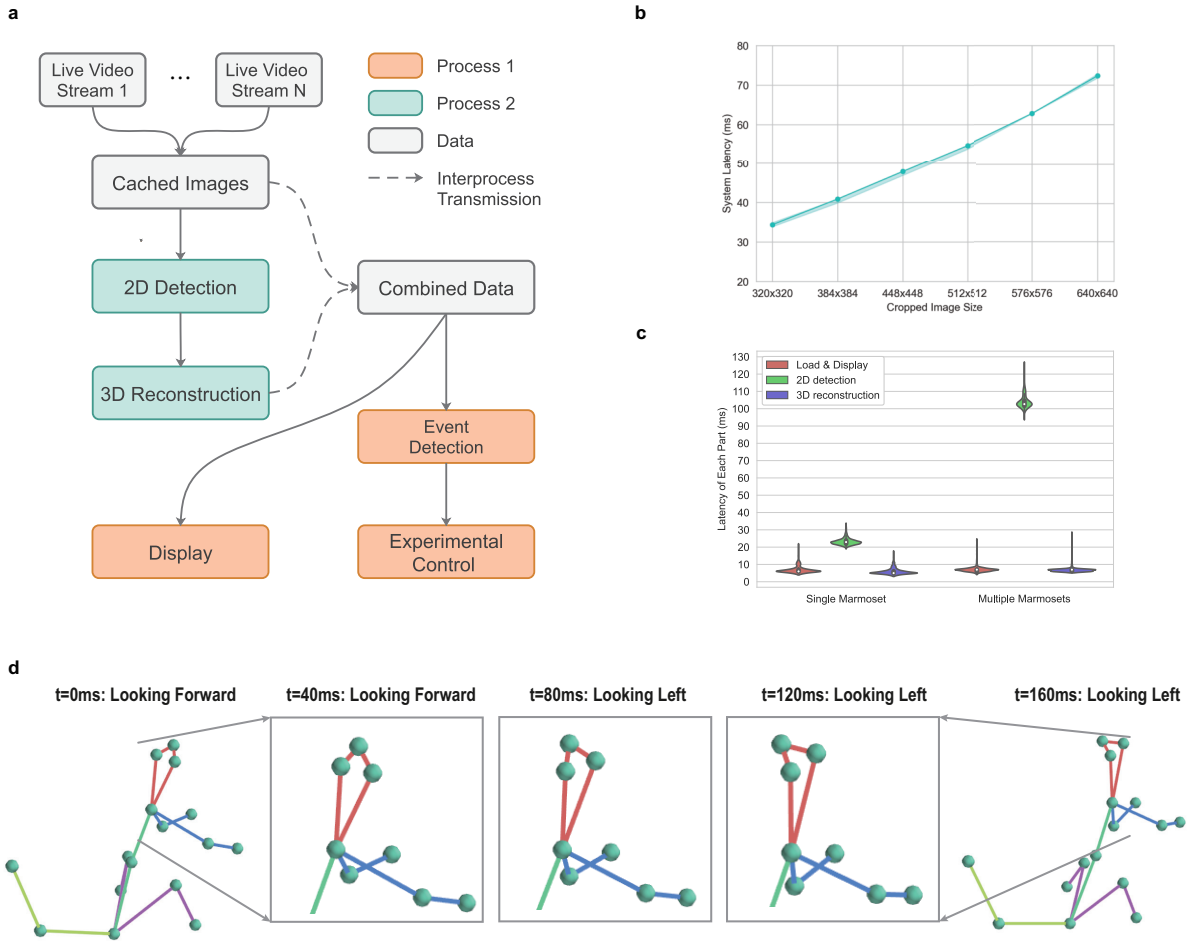

Fig. 4 MarmoPose provides a real-time process module to enable low-latency experimental control. a, Schematic of real-time experimental control. Process 2 reads images from cached live video streams and perform 2D detection and 3D triangulation, while process 1 (main process) reads images and poses to display and perform experimental control. **b**, Overall system latency with different cropped image size. Images from 4 cameras at the same timepoint are cropped around the marmoset's position in the previous frame and stacked together as the input of the current frame. Latencies are measured on a 30s video of single marmoset with size 1920 × 1080 × 3 @ 25fps. **c**, Latency of each part in the real- time control module. Latencies are measured on a 30s video of single marmoset and a 30s video of multiple marmoset with size 1920 × 1080 × 3 @ 25fps. One network is used for single marmoset scenarios, and two top-down networks are used for multiple marmosets scenarios. **d**, An example showing real-time event detection. The marmoset turns its head from forward to left in 160ms (4 frames), while MarmoPose could feedback the head orientation in less than 40ms and perform real-time experimental control based on the detected event.