

Reviewed Preprint

v1 • April 1, 2026

Not revised

✉ For correspondence:

justine.clery@mcgill.ca

Competing interests: No

competing interests declared

Funding: See [page 22](#)

Reviewing editor: Arun SP, Indian Institute of Science Bangalore, India

© 2026, Yang et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

A real-time, multi-animal model for automatic face detection and identification of freely moving common marmosets based on YOLOv8 algorithms

Jiayue Yang^{1,2}, James Wang^{1,2}, Justine Cléry^{1,3,4} ✉

¹The Neuro, Department of Neurology and Neurosurgery, McGill University, Montreal, Canada • ²Integrated Program in Neuroscience, McGill University, Montreal, Canada • ³McConnell Brain Imaging Centre, The Neuro, Montreal Neurological Institute and Hospital, McGill University, Montreal, Canada • ⁴Azrieli Centre for Autism Research, The Neuro, Montreal, Canada

eLife Assessment

This **valuable** study presents a real-time system for identifying multiple unrestrained marmosets in a home cage setting using a combination of face detection and color-coded beads. However, there is **incomplete** evidence regarding the generalizability and robustness of the system to unconstrained multi-animal environments.

<https://doi.org/10.7554/eLife.110932.1.sa3>

Abstract

Precise and up-to-date information about animal location and identity allows us to better quantify individual behaviors in studies of neural activity, cognition, and animal health. In socially housed laboratory animals, identification is usually defined by observation or invasive markers, making the data collection time-consuming, variable across experimenters, and disruptive to animals. We established an automatic pipeline for real-time identification of common marmosets in captivity using a close-view camera. It uses the supervised deep-learning YOLOv8 model to localize individuals, detect faces, and classify identities. Moreover, we use recognition of uniquely color-coded collar beads to improve detection accuracy among visually similar individuals. Across adult and juvenile marmosets, our system automatically identifies marmosets with > 82.9% precision and > 91.5% recall, achieving human level performance. This pipeline is designed to be easy-to-use and generalizable across non-human primate species, ages, and recording hardware, providing rapid and automatic identity recognition from real-time video.

Introduction

In behavioral neuroscience, an accurate identification of laboratory animals is crucial for welfare assessment and ethical care (National Research Council (U.S.), 2011 [↗](#); Vidal et al., 2021 [↗](#)). More importantly, the animal identity is also required when conducting individual-specific behavioral experiments. However, identifying specific animals within a family group can be challenging, especially if they are not marked and share similar facial features. Invasive recognition methods, including branding, tattooing, and ear tagging, may disrupt animals' behaviors and cause potential stress (Carstens and Moberg, 2000 [↗](#); Lim et al., 2019 [↗](#); Roughan and Sevenoaks, 2019 [↗](#)). To avoid such harms, video analysis is becoming more popular as a non-invasive method to identify the animals by their appearance, with no direct contact or handling being involved (Norouzzadeh et al., 2018 [↗](#); Schindler and Steinhage, 2021 [↗](#)). However, collecting video data of animals that is

subsequently labelled by human observation can be time-consuming and hard to reproduce across experimenters. (Buchan et al., 2003 [↗](#); Marion et al., 2020 [↗](#)). Therefore, considering replacement of human work, machine learning tools have been introduced and increasingly used, providing an accurate and automatic estimation of animal identities.

Computer vision and machine learning have made many advancements and offered the foundation to build these automatic systems (Sturman et al., 2020 [↗](#)). Among these approaches, deep learning, a subset of machine learning, has been widely used in animal identification based on video and image data. Specifically, by utilizing deep learning with video recordings, object detection allows an accurate localization and identification of animals, even with those who live in complicated environments (Yu et al., 2018 [↗](#); Zhuang et al., 2025 [↗](#)). Algorithms like ResNet (He et al., 2015 [↗](#)) and You Only Look Once (YOLO) models (Jocher et al., 2023 [↗](#); Redmon et al., 2015 [↗](#)) have demonstrated promising accuracy and computational speed in identifying animals in naturalistic environments (Bakana et al., 2024 [↗](#); Petso et al., 2021 [↗](#)). In addition, unique facial features can also be used as inputs to classify and recognize animals. This approach has the advantage that it provides information regarding whether a specific animal is present or absent in the camera recording view at each timepoint. Although facial recognition has been demonstrated to be successful in wild and domestic animals (Bergman et al., 2024 [↗](#); Norouzzadeh et al., 2021 [↗](#); Schofield et al., 2023 [↗](#)), its application in controlled laboratory settings is not yet extensively evaluated.

The common marmoset (*Callithrix jacchus*), a small non-human primate, is becoming more popular recently in neuroscience research for many reasons (Kishi et al., 2014 [↗](#); Okano, 2021 [↗](#); Sasaki et al., 2009 [↗](#)). Marmosets are family-bonded, cooperative, and notably pro-social. Moreover, they can perform multiple series of behavioral and cognitive tasks, such as observational and reversal learning (Koski and Burkart, 2015 [↗](#); Miller et al., 2016 [↗](#)). As marmosets are typically maintained in social family groups (Yoshimoto et al., 2018 [↗](#)), accurate identification of specific individuals is therefore essential for assessing which animal is performing a given task, which also enables a tracking of their behavioral performance over time. However, existing identification techniques of marmosets usually involve wearable micro-sensing devices, requiring stable device positioning to the sensor and constant adjustments to ensure the devices fit the animal. This automatic system can also become unreliable during rapid movement or if multiple marmosets are present at the sensor, which limits the consistency and stability of identity tracking in freely moving marmosets during their tasks. Thus, facial recognition offers a non-invasive alternative, without a strict need of a physical identity marker, enabling a continuous, accurate, and automated identification when the marmoset enters the designated task performance space.

Here, we developed a pipeline for automatically detecting and identifying marmosets simultaneously from real-time videos, based on their faces. Previous studies have introduced the feasibility of the methods we used in this study (Dave et al., 2023 [↗](#); XiaoAn et al., 2024 [↗](#)), of which we adapted them for facial-based marmoset recognition. We apply the YOLOv8 model (Jocher et al., 2023 [↗](#)), trained on images of specific individuals (three adult and two young marmosets), to generate a real-time marmoset recognition on a frame-by-frame manner. In addition to marmoset faces, we use the color-coded bead on individual collars to facilitate the identity prediction accuracy. Finally, we further evaluated our model performance on a dataset from both adult marmosets and young marmosets at different developmental stages. We show that our facial-based pipeline achieves high detection performance comparable to human expert level, with a strong generalizability across recording settings and marmosets at various ages.

Methods and Materials

Animals

Three adult common marmosets (*Callithrix jacchus*, one female, aged 2 – 6 years, weighted 450g – 566g) and two young common marmosets (two females, aged 7 – 11 months, weighted 274g – 386g) were involved in this study. The three adult marmosets belonged to the same family unit (parents:

Adult1, Adult2, and their adult offspring: Adult3), while the young marmosets were twin siblings (Young1 and Young2) from a separate family. The marmosets are housed at The Neuro's animal facility in family units in indoor enclosures. Housing cages included two sizes. The first size housed three adult marmosets with dimensions of 1.372m length x 0.760m width x 2.092m height, and the second cage that housed the two young marmosets with their family has the dimensions of 1.065m length x 1.067m width x 2.092m height). We placed a collar with a uniquely colored bead on each marmoset to allow consistent identification of individual marmoset during the experiment. All experimental procedures were conducted under the Canadian Council on Animal Care guidelines, Standard Operating Procedures of marmosets, and Animal Use Protocol (AUP# 10000 and 10001) approved by The Neuro and McGill's Animal Care Committee.

Video collection

Marmoset face detection and identification data were collected by mounting a primate testing chair with an added camera component to the door of marmoset housing cages (Figure 1A [↗](#) and 1C [↗](#)). High-resolution videos were acquired using an industrial color/RGB camera (JIERUIWEITONG DF200-1080P). The industrial camera was selected specifically for its small size (camera box: 36mm width x 36mm height) and capability of close-distance recordings. High-resolution recordings (1920×1080 pixels, 30 frames per second) obtained with the small focal length lens (2.8mm) enabled a wide field of camera view, which was sufficient to capture full facial features and posture variability at close recording distances. The camera was adjusted for focus and placed approximately 10.50 cm from the housing cage door (Figure 1A [↗](#) and 1B [↗](#)). The camera was fixed in position using a transparent protective case, which was also designed to prevent damage to the camera by the marmosets (Figure 1B [↗](#)).

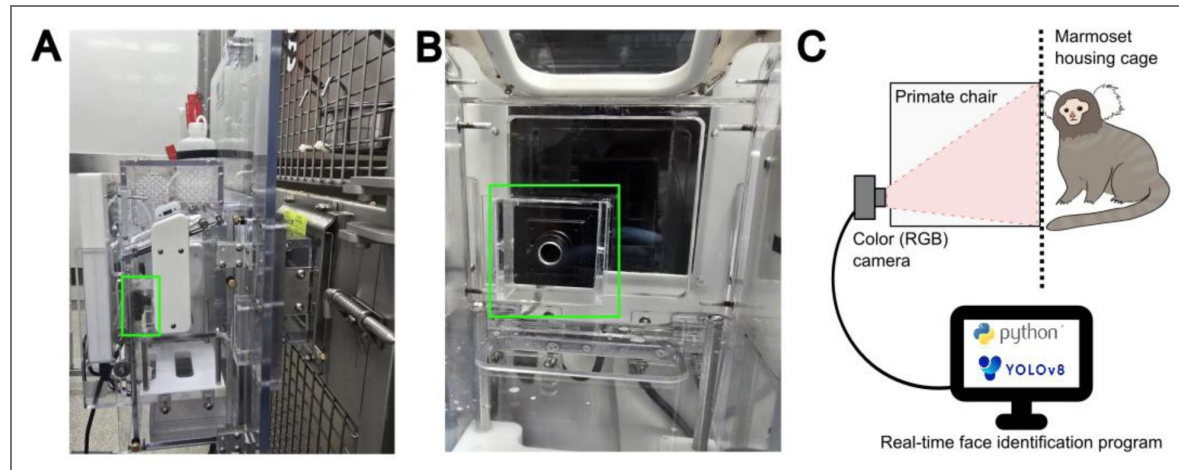
Prior to the experiments, marmosets were acclimated to entering the primate testing chair space. The transparent primate chair space minimized the visual reflection of the animal on the chair surfaces, reducing interference between true facial features and reflections during face detection and identification. Adult marmosets (n=3) from one family were recorded for one hour on a given recording day, with unrestricted voluntary access to the primate chair space throughout the video collection period. Two young marmosets were briefly isolated and recorded individually for testing and improving the automatic face extraction model. We recorded them at two developmental timepoints, 7 months old and 11 months old. During video collection, a sliding panel and an in-cage box were positioned near the housing cage door to temporarily isolate individual marmosets from other family members. Individual isolation was kept brief (approximately 10 minutes) to prevent disturbance and potential stress due to family separation. Marmosets accessed the primate chair space through the housing cage door (11cm width x 10.16cm height). These dimensions of the cage door allowed entry of individual marmosets at one time, ensuring face video collection of one individual per entry. For the adult marmosets, the housing cage door was opened at the beginning of the recording session allowing them to enter and exit freely into the primate chair space for food rewards and observation (Figure 1C [↗](#)).

Video preprocessing and bounding box annotation

The collected videos were preprocessed to identify which video clips included valid face and collar features of the marmoset. Marmoset identities and corresponding collar colors were recorded during clip selection. We defined valid face and collar features by whether the full-face details and colored bead appeared visible from the clips (Figure 2A [↗](#), Step 1). To capture sufficient variability in postures, individuals, and lighting conditions, selected clips were distributed across the entire adult marmoset videos (approximately 5 hours) (Figure 2B [↗](#)). We used OpenCV to extract frames from the selected clips, and the extracted frames were reviewed to exclude those that were blurry (Bradski, 2000 [↗](#)). Computer Vision Annotation Tool (CVAT) was applied to perform bounding box annotation of marmoset faces and collar colors from the extracted frames (Sekachev et al., 2020 [↗](#)).

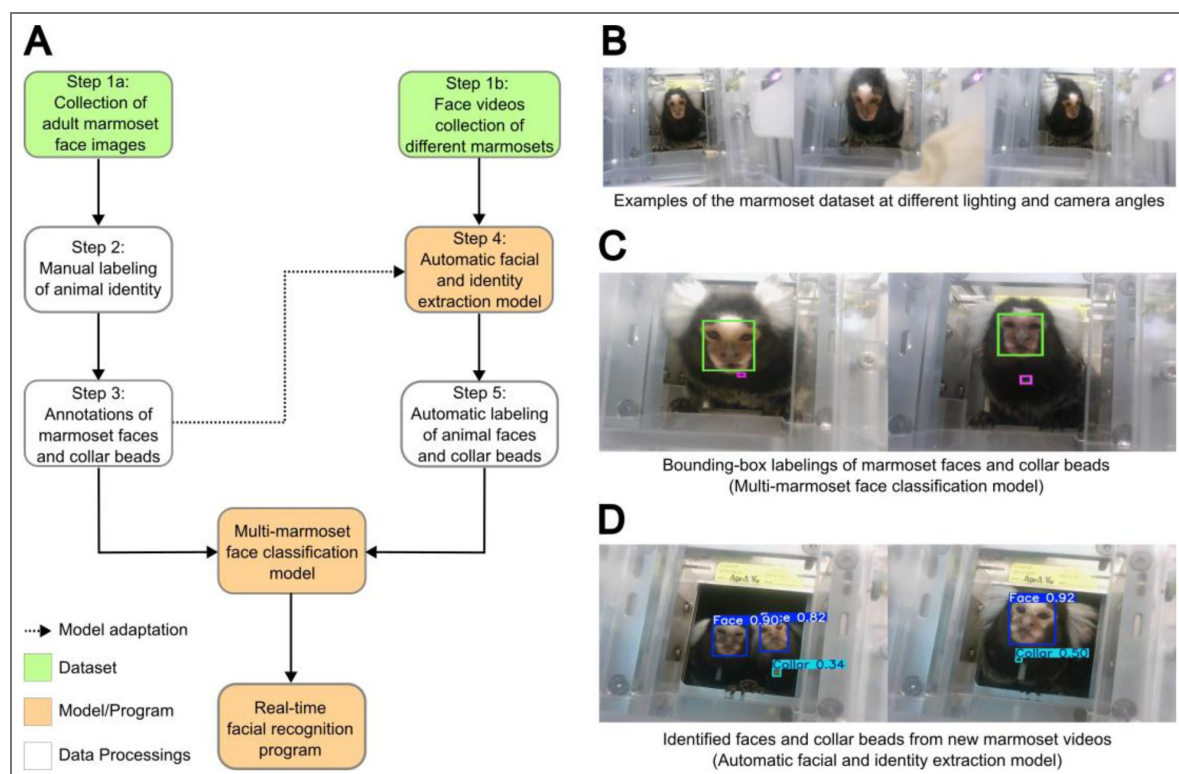
Figures 1. Experimental setup of the camera system and primate chair for face detection and identification.

(A) Side view of the camera (green box) and primate chair, attached to the housing cage. (B) Frontal view of the camera fixed on the primate chair, enclosed within a protective cover. (C) A schematic illustration of the real-time facial image recording and automatic identification of the marmoset entering the primate chair.



Figures 2. Workflow and design of the marmoset facial detection and identification model.

(A) The architecture of the real-time marmoset facial recognition program. (B) Marmoset face images from three camera angles. (C) Bounding boxes of marmoset faces (green box) and collars (pink box) were manually labeled in the training and validation datasets to train the multi-marmoset face classification model. (D) The schematic of the automatic face (blue box) and collar bead (cyan box) extraction model.



Marmoset face and identity dataset

To minimize image computations and data storage, we created a dataset of 2498 annotated images from the three adult marmosets. All images were manually annotated to label marmoset faces, individual identities, and the collar bead colors (Figure 2C [↗](#)). The annotated images were used for training models of multi-marmoset face classification and the automatic identity extraction, which can automatically detect, localize, and identify marmoset faces (Figure 2A [↗](#), Step 1 – 4). We created another dataset of two young marmosets at 7 months old (total images = 502) for testing the automatic facial and identity extraction model (Figure 2A [↗](#), Step 4 – 5). For both adult and young marmoset datasets, images were randomly divided into a training set and a validation set at a ratio of 8:2. Moreover, new videos from the adult marmosets and from the same young marmosets at 11 months, which were not involved during initial model training, were used to further evaluate model performance and assess its generalization across developmental stages.

Marmoset face recognition pipeline

The marmoset facial detection and identification model was primarily based on the YOLOv8 algorithm (Jocher et al., 2023 [↗](#); Varghese and Sambath, 2024 [↗](#)). The architecture of the program consists of two main models: a multi-marmoset face classification model and an automatic facial and identity extraction model (Figure 2A [↗](#)). In the initial step, we deployed multiple pre-trained YOLOv8 models (YOLOv8 nano, YOLOv8 small, and YOLOv8 medium) on adult marmosets' dataset (total images = 2498) to train the multi-marmoset facial classification model (Lin et al., 2015 [↗](#)). Model performance and detection accuracy of these pre-trained models, generated by YOLO metrics as a CSV file, were evaluated to select the optimal model for subsequent training (see Results section). The YOLOv8 nano model was selected and used for the training of all models in the program (Figure 2C [↗](#)). The adult marmosets' dataset was also applied in the training of automatic facial and identity extraction model. This trained automatic face extraction model was utilized on the young marmoset dataset (total images = 502). Since young marmosets were recorded individually, we assigned the automatic annotation of marmoset faces and collar beads based on their recorded identity (Figure 2D [↗](#)). All automatic labeling was manually reviewed to remove mislabeled annotations. We used this verified young marmosets' dataset (total images = 449) to train the multi-marmoset face classification model for the young marmosets.

The marmoset facial recognition program was trained in Anaconda virtual environment. The final models were selected if they reached the early stopping criteria, which indicated that no improvement was observed in the following 100 epochs (i.e. training iterations). This program involved training of three final models using the pre-trained YOLOv8 nano model: (a) multi-marmoset facial classification model for three adult marmosets was trained for 183 epochs; (b) automatic facial and identity extraction model was trained for 124 epochs; (c) multi-marmoset facial classification model for two young marmosets was trained for 245 epochs. The final models generated 2D bounding boxes, assigning customized ID of individual marmosets and collar beads. The detection results of input video frames were exported in a text file, which represented the corresponding marmoset label identity and the bounding box size and location (Jocher et al., 2023 [↗](#); Redmon et al., 2015 [↗](#)).

Model evaluation and analysis

The marmoset facial recognition program was evaluated using a range of performance evaluation metrics. The evaluation metrics included: precision, recall, Mean Average Precision (mAP), Validation Distribution Focal Loss (DFL), and training time. We also computed the F1 score to select the best setup to be applied in the final real-time detection pipeline. In addition, we calculated the inter-individual face similarity between marmosets to test whether the face similarity between family members can explain the mislabeling of our marmoset facial recognition model.

Intersection over Union (IoU)

Intersection over Union (IoU) calculates the amount of spatial overlapping between a predicted bounding box and the ground truth (i.e. manually labeled bounding box) (Jocher et al., 2023 [↗](#); Rezatofighi et al., 2019 [↗](#)). It measures the localization accuracy and error amount between the predicted annotation and the manually labeled ground truth. A detection was considered correct if the IoU reached specific thresholds. In this study, we reported the mean averaged precision (mAP) at the IoU thresholds from 0.5 to 0.95 (increments = 0.05). The IoU is calculated by the area of overlap ($A \cap B$) and area of union ($A \cup B$) of the predicted bounding box (A) and ground truth bounding box (B). If $\text{IoU} = 1$, the predicted bounding box is perfectly aligned with the true annotation. If $\text{IoU} = 0$, there is no overlap between the two boxes.

$$\text{IoU} = \frac{A \cap B}{A \cup B}$$

Precision, recall, and F1 score

Precision, recall, and F1 score are commonly used in the evaluation of classification models (Dehmer and Basak, 2012 [↗](#); Powers, 2008 [↗](#)). Precision measures the overall proportion of correct face identifications over all positive results, indicating the accuracy of the detection model. Recall quantifies the proportion of the face identification that are correct compared with all actual positives, which reflects the ability of the model to predict correct marmoset face and collar beads. The F1 score is the harmonic mean of precision and recall, considering the false positives and false negatives in the model assessment.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Mean Average Precision (mAP)

Mean Average Precision (mAP) assesses how accurate the model detects objects and how well the model localizes the object in the image. In object detection algorithms, precision is the number of correctly detected objects, with recall being the number of true objects that are successfully detected. As the precision and recall are affected by different detection confidence thresholds, the performance is evaluated by the precision-recall curve, which plots the model precision (y-axis) against recall (x-axis) for a particular threshold. The average precision (AP) is the area under the precision-recall curve (Padilla et al., 2021 [↗](#)). Before calculating the AP, the precision-recall pairs are interpolated to create a monotonically non-increasing precision-recall function, meaning that the interpolated precision ($P_{\text{interpolation}}$) is the maximum precision ($\max_{j \geq i} \text{Precision}_j$) at the recall level larger than or equal to Recall_i . This interpolation assigns a highest precision value at each recall level, which smooths the precision-recall curve and reduces the impact of measurement fluctuations and noises to the evaluation of model performance.

$$P_{\text{interpolation}}(\text{Recall}_i) = \max_{j \geq i} \text{Precision}_j$$

For each label class in the object detection model, the AP is the area under the curve (AUC) of the interpolated precision-recall curve, calculated by summing the interpolated precision weighted by the incremental increase in the recall (Maxwell et al., 2021 [↗](#); Padilla et al., 2021 [↗](#)).

$$\text{AP} = \sum_{i=1}^N (\text{Recall}_i - \text{Recall}_{i-1}) P_{\text{interpolation}}(\text{Recall}_i)$$

The AP is extended further to calculate the Mean Average Precision (mAP), which is the averaged AP values of different label classes in the object detection model (total class number = n). The AP_i is the mean AP calculated per label class at index i at multiple IoU levels (IoU = {0.50, 0.55, ..., 0.95}). This calculation generates a more comprehensive evaluation with multiple label classes and their predicted localization accuracy of our marmoset facial recognition program.

$$mAP@50-95 = \frac{1}{n} \sum_{i=1}^n AP_i$$

Validation Distribution Focal Loss (DFL) and training time

The distribution focal loss (DFL) and training time were also used to support the selection of the final marmoset facial recognition model. The DFL function measures the model's performance to refine the bounding box predictions, based on the localization uncertainty of the bounding box annotations (Jocher et al., 2023 [↗](#)). The predicted bounding box position is compared with the ground truth annotations, with a lower DFL value indicating a more precise bounding box prediction.

Training time is not involved in the model evaluation metrics, but we used the training time to assess the efficiency and the feasibility of the real-time marmoset face recognition program. The training time estimated the computational cost of the program, which provided additional information to select models with similar accuracy. The training time is calculated by the total time duration to train the model. For the number of training epochs, N_{epoch} , the total training time is calculated through multiplying N_{epoch} by the training time per epoch (tepoche):

$$Training\ time = N_{epoch} \times t_{epoch}$$

Inter-individual face similarity analysis

We calculated the inter-individual face similarity using customized Python scripts (See Data Availability). The image feature embeddings, based on YOLOv8 algorithms, calculate the numerical representations (i.e. vectors) using a computer vision model, which encodes the semantic and visual maps of the image (Long Chai et al., 2023 [↗](#); Varghese and Sambath, 2024 [↗](#)). In our marmoset facial recognition program, the intermediate layers of the YOLOv8 model included extensive feature maps of the visual details of the marmoset face images. By extracting and analyzing the image feature embeddings of our model, we obtained various numerical representations of the marmoset faces, which allowed us to compare face shapes, distances between face features (eyes, mouth, nose, etc.), and the fur patterns of different individuals. We assessed the inter-individual face similarities for 2 family groups: (a) within the adult marmoset family (parents and adult offspring) and (b) between the two young marmosets (twin siblings). For each individual marmoset, 10 images, with similar head orientation, lighting conditions, and camera angles, were selected and cropped to only include marmoset face regions. As the marmoset face images passed through the YOLO convolutional layers, the images were outputted into a feature map (Aghdam and Heravi, 2017 [↗](#); Dumoulin and Visin, 2016 [↗](#); Long Chai et al., 2023 [↗](#); Redmon et al., 2015 [↗](#)). For one marmoset face image, the feature map (F_i) is encoded in the number of inputs (n_{input}) and the feature map shapes (Aghdam and Heravi, 2017 [↗](#); Dumoulin and Visin, 2016 [↗](#)).

$$F_i(h, w, c) = n_{input} \times height \times width \times channel$$

We extracted face embeddings from our models trained separately on a family of adult marmosets (parents and son) and young twin marmosets from another family. To generate a single vector (i.e. embedding, e_i) representing the marmoset face feature per channel, we averaged across the spatial dimension (i.e. height and width of the image) per channel for one marmoset face image (Lee et al., 2025 [↗](#)). The extracted embedding (e_i) was averaged across the 10 images for each individual marmoset, which allowed us to have an averaged embedding vector ($e_{marmoset}$)

across frames and feature maps to calculate the inter-individual face similarities. We performed L2 normalization on this embedding (emarmoset) to create a unit vector (enorm) for inter-individual comparison, using scikit-learn preprocessing function (Pedregosa et al., 2011 [↗](#)).

$$e_i = \frac{1}{height} \frac{1}{width} \sum_{h=1}^{height} \sum_{w=1}^{width} F_i(h, w, c)$$

$$e_{marmoset} = \frac{1}{10} \sum_{i=1}^{10} e_i$$

$$e_{norm} = \frac{e_{marmoset}}{\|e_{marmoset}\|}, \text{ where } \|e_{marmoset}\| = \sqrt{\sum e_{marmoset}^2}$$

As such, with the normalized embedding vectors from the marmosets in this study, we calculated the inter-individual face similarities between marmoset pairs using cosine similarity (Li and Han, 2013 [↗](#); Nguyen and Bai, 2011 [↗](#); Thongtan and Phientrakul, 2019 [↗](#)). Moreover, we also calculated the Euclidean distances between extracted embeddings from two marmosets (Jozwik et al., 2022 [↗](#); Sugase-Miyamoto et al., 2014 [↗](#)). The calculation equations were described using the normalized embedding vectors of a marmoset pair (enorm - 1 and enorm - 2) in scikit learn library (Merchant et al., 2023 [↗](#); Pedregosa et al., 2011 [↗](#)):

$$\text{Cosine similarity} = \frac{e_{norm-1} \cdot e_{norm-2}}{\|e_{norm-1}\| \|e_{norm-2}\|}$$

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^n (e_{norm-1} - e_{norm-2})^2}$$

We plotted a violin plot of cosine similarity and Euclidean distance to interpret the inter-individual facial similarities calculated from the marmoset pairs' face embeddings. In addition, the similarity values were standardized using within-model z-scores. This allowed us to plot and visualize the distribution of face similarity scores across family relationships and models.

Statistical tests were performed only within each marmoset face classification model, as embedding spaces may vary in scaling, learned features, and baseline metrics making cross-model comparison of inter-individual face similarity unreliable (Bollegala, 2017 [↗](#)). To test whether the learned facial embeddings between different marmosets is related to the biological facial feature similarities, we performed the pairwise Welch's t-test and the Cohen's d to examine the significance and effect size. Within each family unit, facial similarity was compared between two types of family relationships (e.g. mother-son vs. father-son).

Real-time face recognition program

We performed the real-time marmoset face detection and identification using the customized Python scripts, based on the trained weights from the final multi-marmoset facial classification model in YOLOv8. Live videos were collected using the color (RGB) camera and simultaneously processed frame-by-frame to detect the marmoset identities. For each detected bounding box, the scripts returned a corresponding label of marmoset face and collar bead color. We assigned the detected collar beads as the corresponding marmoset identity, with higher weights than the detected face, which improved the final identity stability against potential marmoset facial mislabeling.

To ensure the program accuracy and efficiency, we temporally stored the detection results of most recent 30 frames (approximately one second). The most frequently detected marmoset identity within this time window was exported as the current output. We wrote and updated the detection results continuously into an output JSON file, enabling real-time identity reading of the marmoset who entered the primate chair most recently.

Results

Comparison of pre-trained YOLOv8 models

The results of the training and evaluation of the pre-trained YOLOv8 models (YOLOv8 nano, YOLOv8 small, YOLOv8 medium) were shown in Table 1. The models reached their best performance at 180 epochs (YOLOv8 nano), 132 epochs (YOLOv8 small), and 124 epochs (YOLOv8 medium). Each of the models was trained until reaching the early stopping criteria (i.e. no improvement within the last 100 training epochs). By comparing the performance differences between the YOLOv8 pre-trained networks, we aimed to find the most suitable model for the real-time marmoset identification program.

As presented in Table 1, YOLOv8 nano model reached the best overall detection performance of the adult marmosets' dataset. YOLOv8 medium model achieved a higher precision among all three pre-trained networks. However, the recall and F1 score were higher for the YOLOv8 nano model, indicating robust model sensitivity and a balanced precision-recall value. Mean average precision (mAP@50–95) showed values ranging from 0.710 – 0.713 for the three models. YOLOv8 medium model reached the highest detection accuracy for adult marmoset faces, though the difference was minor while comparing to the YOLOv8 nano and small models. In addition, we considered the validation DFL and training time to support the selection of the optimal pre-trained model. Validation DFL increased with model sizes, suggesting that the detection localization became less accurate from YOLOv8 nano to YOLOv8 medium. Moreover, the training time increased from YOLOv8 nano to medium model, while YOLOv8 medium took about twice the training time compared to YOLOv8 nano model. Thus, we selected the YOLOv8 nano model for training the real-time program due to its high computational speed and efficiency, while maintaining reliable marmoset face prediction compared to larger YOLOv8 models (Jocher et al., 2023).

Model prediction of adult marmoset faces approaches high accuracy comparable to human experimenters

We applied the YOLOv8 nano model for the adult marmoset face classification model. The training curves of the overall precision, recall, and mAP@50-95 (IoU = 0.5:0.95) were visualized in Figure 3A, B, C, and Supplementary Figure 1. The final training model was selected at the training epoch 183 (denoted by the red dotted line), reaching the precision at 0.932, recall at 0.964, and the mAP@50-95 (IoU = 0.5:0.95) at 0.710 (see Table 1). We evaluated the model performance per label class using the adult marmoset' validation dataset (Figure 3D, Supplementary Table 1). Analyzing the label classes, the final training model showed stable and high precision and recall across both marmoset individual identity (i.e. Adult1, Adult2, Adult3) and bead-color (i.e. collar_Adult1, collar_Adult2, collar_Adult3) label classes. In addition, the marmoset face classes showed high mean average precision across varied IoU thresholds, suggesting a robust prediction localization for marmoset faces (mAP@50-95 = 0.841, 0.787, 0.868 for three adults respectively). The lower mAP value observed for the bead-color label category (mAP@50-95 = 0.603, 0.665, 0.515 for collars of three adults) suggested a greater variability in bounding box localization for the marmoset collar beads. This effect was consistent with the small bounding box sizes of the collar-bead label class, which increased the model sensitivity to localization variability in the prediction (example detection shown in Figure 2C). The normalized confusion matrix comparison was shown in Figure 4. The detection accuracy showing similar performance using the adult marmosets' training dataset (Figure 4A) and validation dataset (Figure 4B), with very low occurrence of false detection between the marmoset individuals and the background images.

Automatic facial and identity extraction model accurately localizes unseen marmoset faces and their collar beads

The labelled adult marmosets' dataset was used for training the automatic facial and identity extraction model, using the YOLOv8 nano pre-trained model. This model aimed to extract and localize the marmoset faces and collar beads from the collected video, including unseen

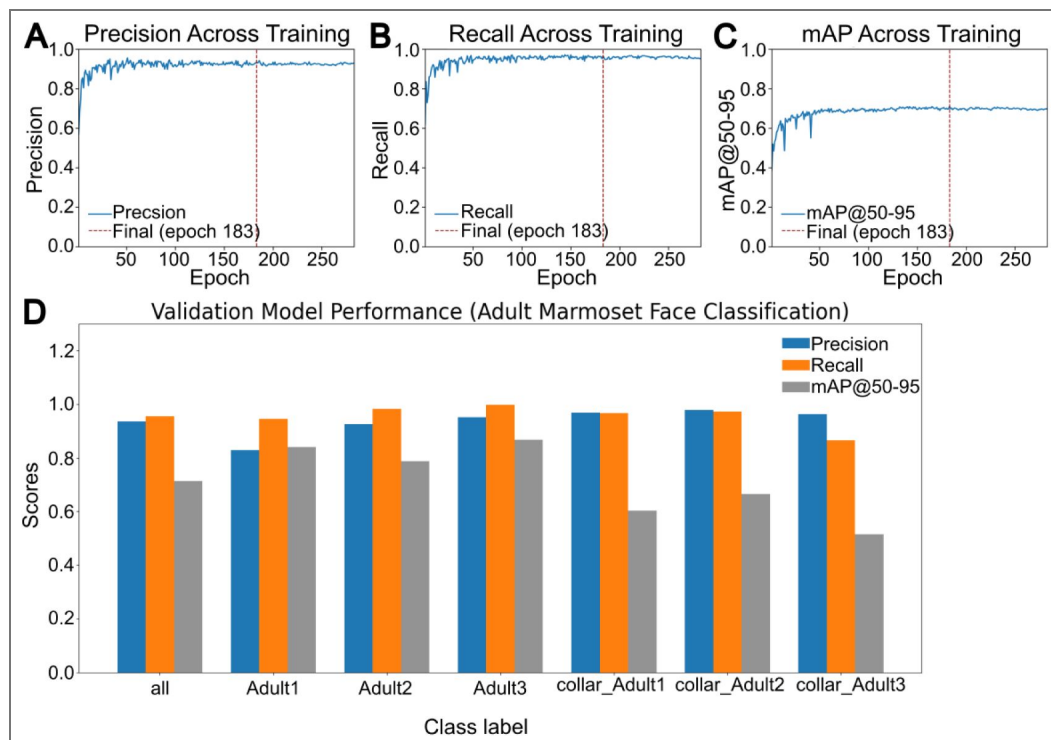
Table 1. Performance comparison of multi-marmoset face classification models.

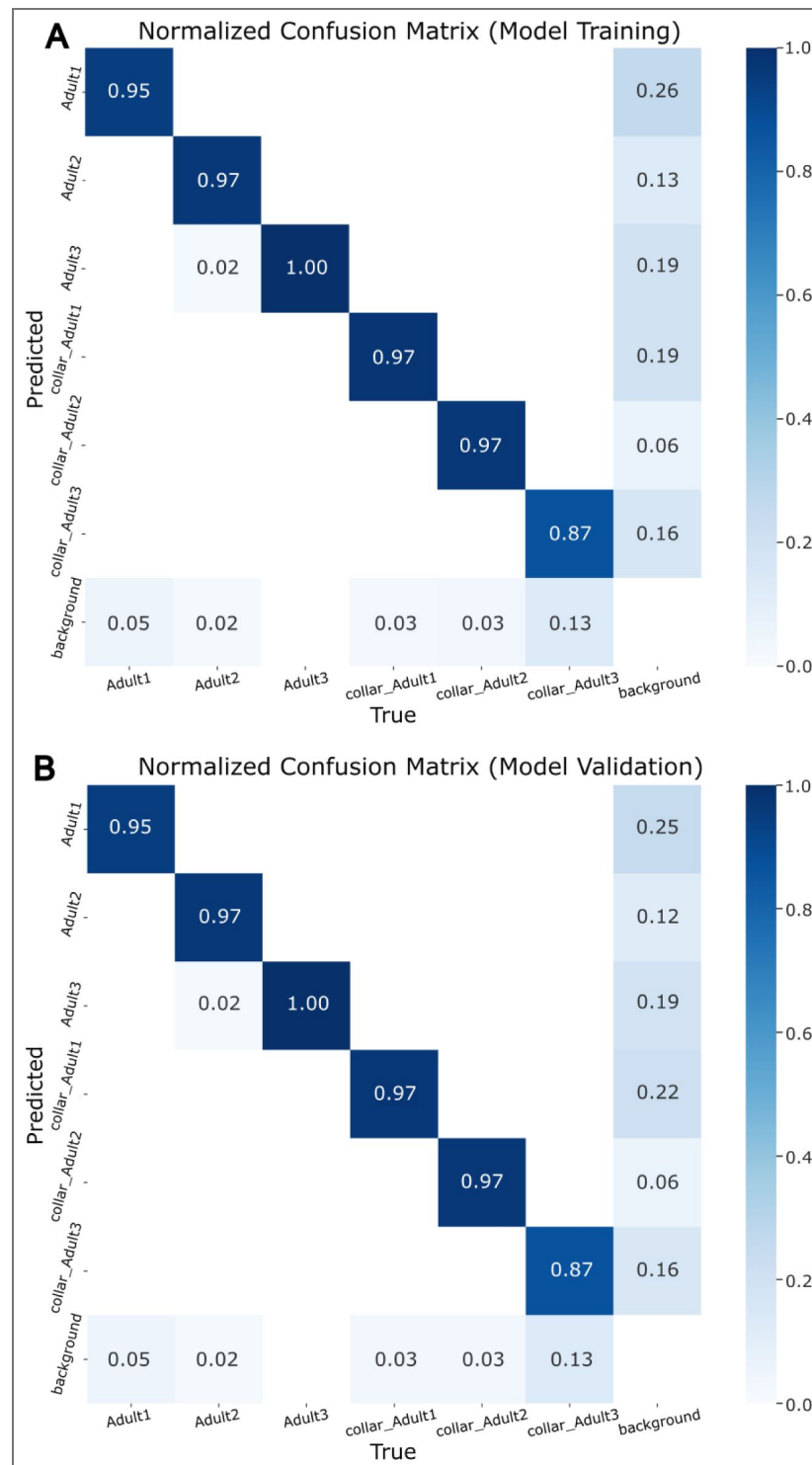
Comparison of Recall, Precision, F1 score, mAP at IoU = 0.5:0.95, Validation DFL, and training time of the three pre-trained model, based on the performance of marmoset face detection and identification on the adult marmosets' dataset. The highest values of each parameter were highlighted in bold.

Metrics	YOLOv8 nano	YOLOv8 small	YOLOv8 medium
Recall	0.964	0.957	0.956
Precision	0.932	0.927	0.937
F1 score	0.948	0.942	0.946
mAP@50-95	0.710	0.710	0.713
Val DFL	0.917	0.953	1.010
Training time (s)	5544	5908	10705

Figures 3. Training performance of the multi-marmoset face classification model for three adult marmosets, using the pre-trained YOLOv8 nano model.

(A) Precision of all detection classes across training epochs. The red dotted line denoted the final model with the best performance at training epoch 183. (B) Similar to (A), except for overall recall. (C) Similar to (A), except for the mAP at the IoU at 0.5:0.95. (D) Overall Model precision, recall, and the mAP@50-95 (IoU = 0.5:0.95) for each label class.





Figures 4. Normalized confusion matrix per-class classification across the 6 label classes in the adult marmoset recognition model.

The y-axis represents the predicted class, and the x-axis represents the manually labelled class. Proportion was generated by the (A) training dataset and the (B) validation dataset, showing whether certain classes were frequently mislabelled as a different class.

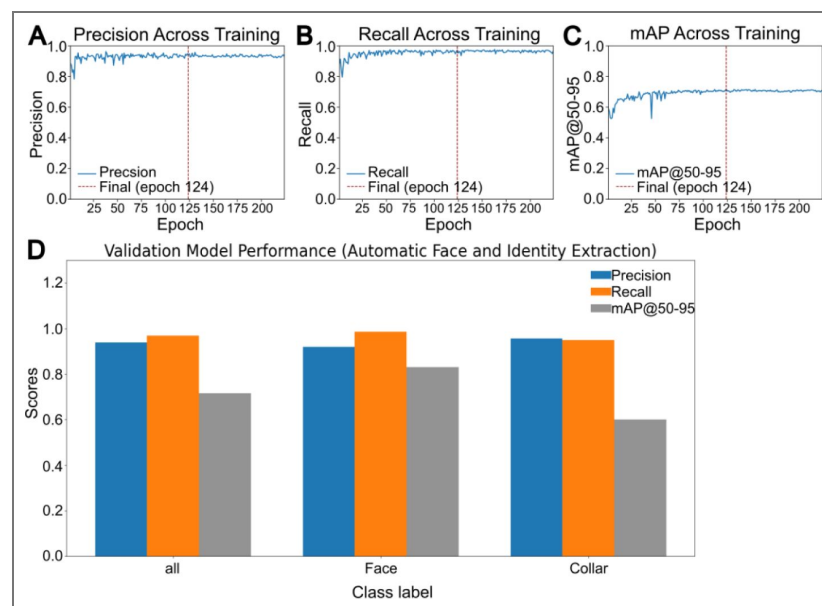
marmosets. The trend of the training was shown in [Figure 5A – C](#) and [Supplementary Figure 2](#), where the detection accuracy improved rapidly in the beginning, with the increasing rate then stabilizing and reaching a plateau. We selected the final model at the training epoch of 124, with a prediction score at 0.940, a recall score at 0.970, and the mAP@50-95 (IoU = 0.5:0.95) at 0.716. In addition, precision, recall, and mAP@50-95 (IoU = 0.5:0.95) were calculated per label class for marmoset faces and collar beads ([Figure 5D](#), [Supplementary Table 2](#)). Based on the model performance of the adult marmosets' validation dataset, the final model accurately predicted the marmoset faces (precision = 0.919, recall = 0.968) and collar beads (precision = 0.957, recall = 0.95). The bead-color class showed lower mAP@50-95 (IoU = 0.5:0.95) values at 0.6, compared to the mAP@50-95 (IoU = 0.5:0.95) for marmoset face class at 0.95, suggesting a similar effect for the face classification model of adult marmosets in [Figure 3D](#). The normalized confusion matrices were compared between the training and validation dataset ([Figure 6](#)). It was evident that the automatic face and identity model exhibited a high detection accuracy. Nonetheless, the background was frequently mislabeled as both marmoset faces and bead-color label classes for the training ([Figure 6A](#)) and validation set ([Figure 6B](#)).

To test the performance of the automatic face and identity extraction model, we used different short clips of new marmosets. We showed that our automatic face and identity extraction model could detect and localize unseen marmoset faces and collar beads at different ages. [Supplementary Video 1](#) showed an example of the detection of faces and bead-color labels for two unseen younger marmosets at 7 months and 17 months, with the confidence threshold of the detection set at 0.6.

Comparison of identification accuracy in young marmosets across developmental stages

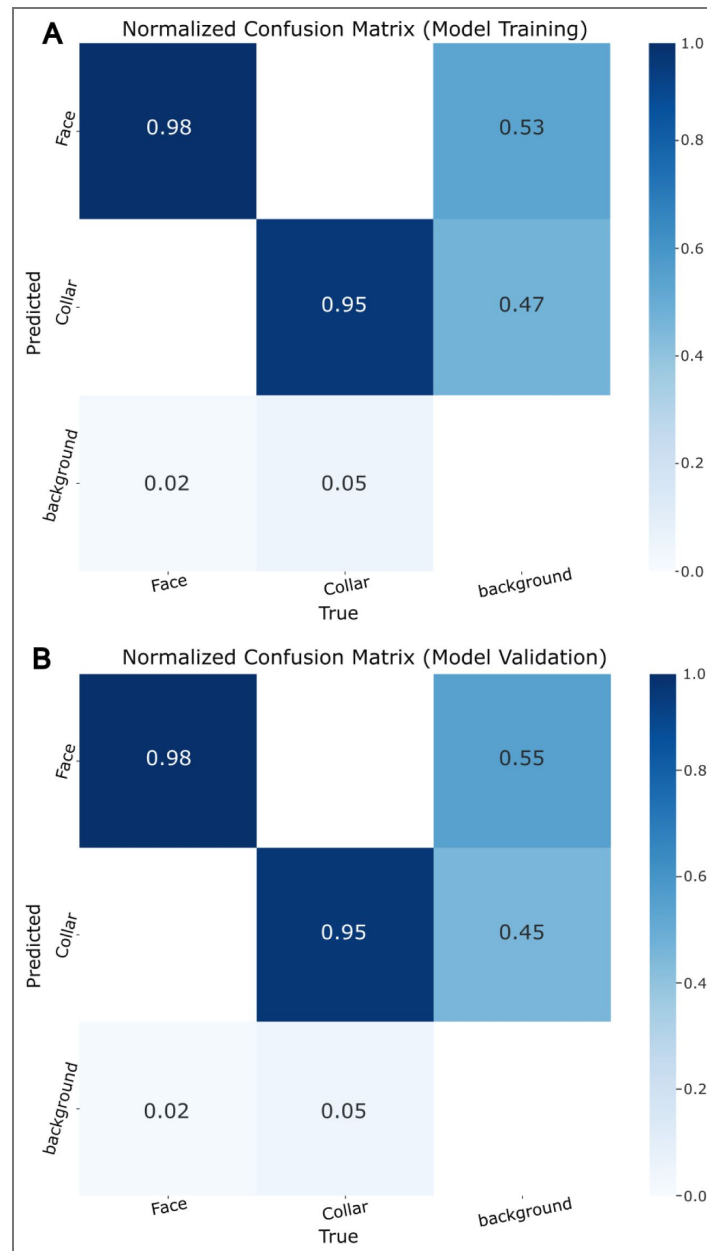
The training and validation datasets of the young marmosets were annotated automatically using the automatic facial and identity extraction model. All automatically generated labels were reviewed by the experimenter, and mislabeled images were removed prior to the model training. The training performance curves were presented in [Figure 7A – C](#) and [Supplementary Figure 3](#). The final model at the training epoch of 245, where the training performance metrics achieved a plateau. This final training epoch achieved the best performance matrices, with precision score at 0.979, recall at 0.975, and mAP@50-95 (IoU = 0.5:0.95) at 0.892. In the young marmoset faces and their collar bead label classes, our final model reached high precision and recall performances using the validation dataset ([Figure 7D](#), [Supplementary Table 3](#)). The precision and recall scores for the label classes were: 1) face of Young1: precision = 1, recall = 0.935; 2) face of Young2: precision = 0.958, recall = 1; 3) collar of Young1: precision = 0.998, recall = 1; 4) collar of Young2: precision = 0.963, recall = 0.964. In terms of localization accuracy, we found a similar pattern of reduced mAP@50-95 (IoU = 0.5:0.95) in bead-color class (mAP@50-95 = 0.808, 0.855), compared to the marmoset faces (mAP@50-95 = 0.95, 0.947). We further computed the normalized confusion matrices of this model using the training and validation dataset. We found that our face classification model was relatively accurate across detection of most class labels ([Figure 8](#)). The normalized confusion matrices showed high accuracy and consistency of most marmoset faces and collars detection in training ([Figure 8A](#)) and validation ([Figure 8B](#)) tests, with some exceptions. Particularly, the background images were frequently identified as the collar of Young2 marmoset, which was likely due to the difficulty in distinguishing multi-color combination from one single collar around the neck region.

To test the detection performance of the model trained on 7-month-old young marmosets across different developmental stages, we used short clips of the two young marmosets at the age of 7 months ([Supplementary Video 2](#), [Supplementary Video 3](#)) and 11 months ([Supplementary Video 4](#), [Supplementary Video 5](#)). [Supplementary Video 2](#), [3](#), [4](#), [5](#) presented the example detections of the young twins at different ages, showing the difficulty of identifying and distinguishing between twin marmosets.



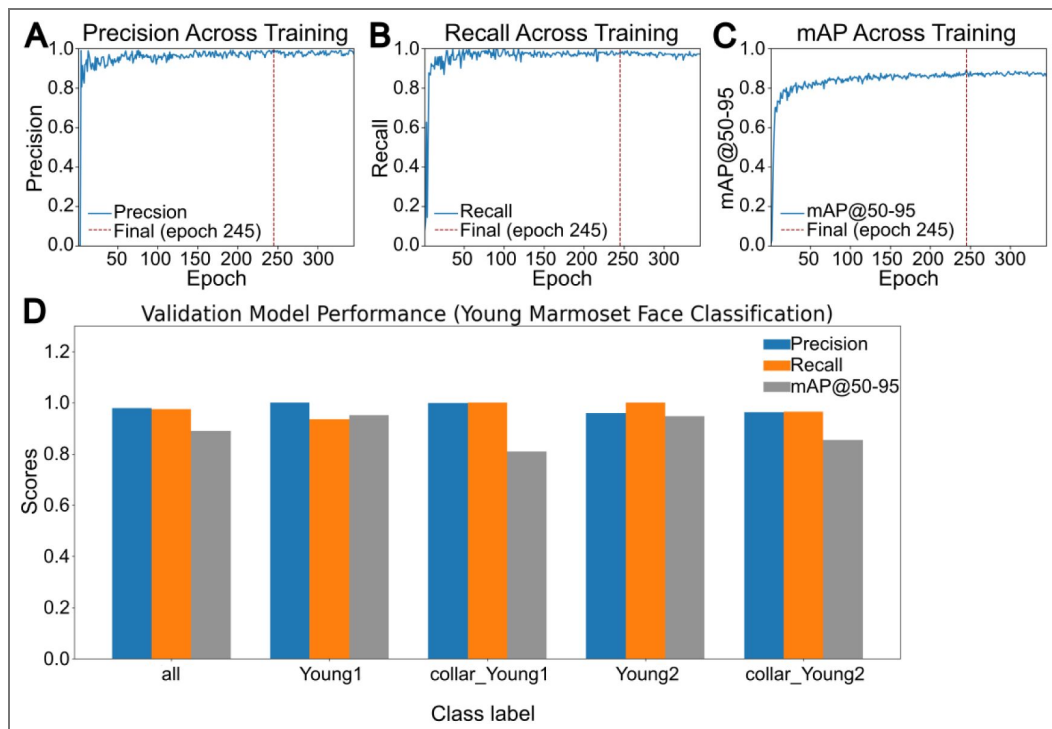
Figures 5. Training performance of the automatic facial and identity extraction model.

(A) Precision of all detection classes across training epochs. The red dotted line denoted the final model with the best performance at training epoch 124. (B) Similar to (A), except for overall recall. (C) Similar to (A), except for the mAP at the IoU at 0.5:0.95. (D) Overall Model precision, recall, and the mAP@50-95 (IoU = 0.5:0.95) for each label class.



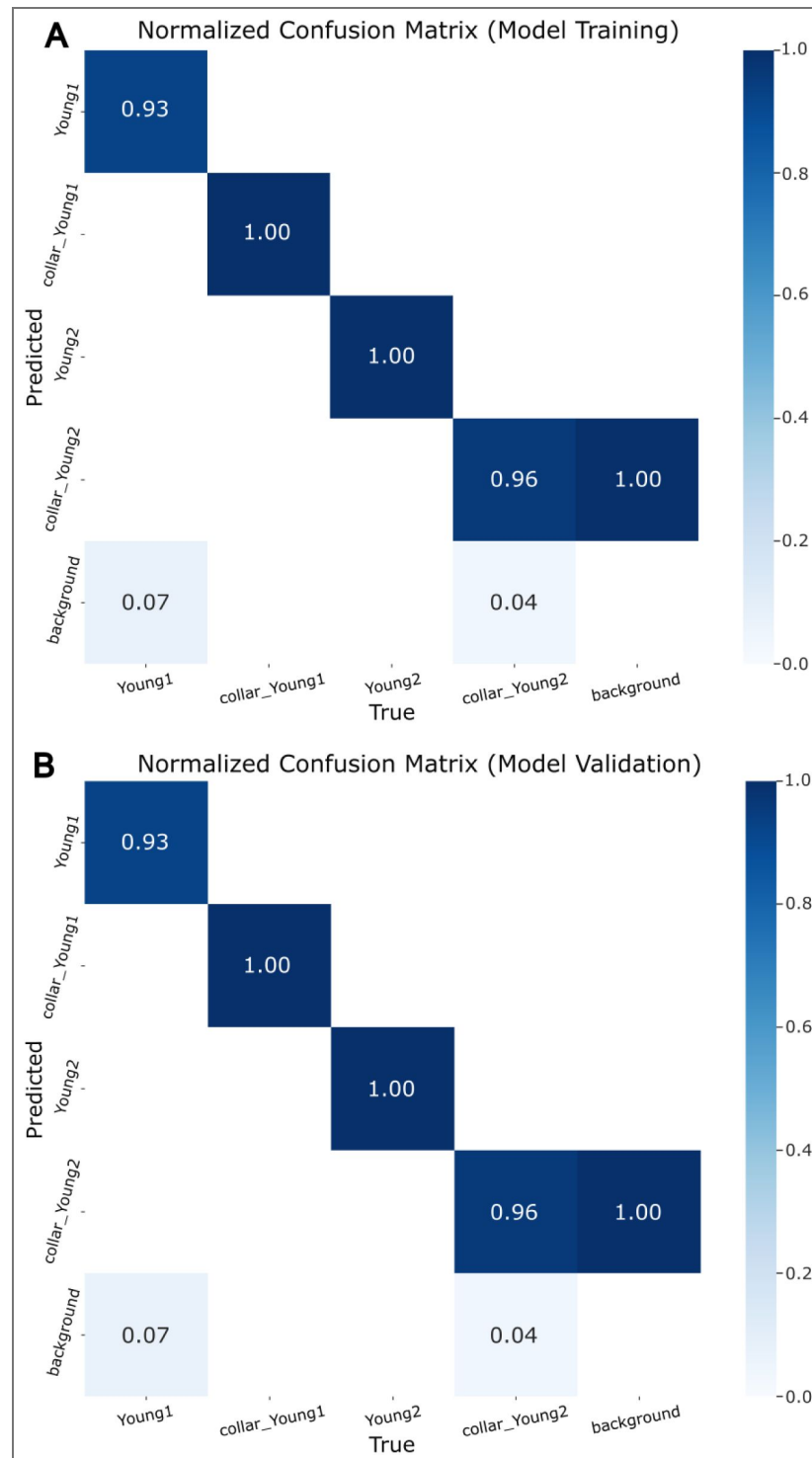
Figures 6. Normalized confusion matrix per-class classification across the 2 label classes in the automatic identity extraction model.

Proportion was generated by the (A) training dataset and the (B) validation dataset, showing whether certain classes were frequently mislabelled as a different class.



Figures 7. Training performance of the multi-marmoset face recognition model for young marmosets at 7 months.

(A) Precision of all detection classes across training epochs. The red dotted line denoted the final model with the best performance at training epoch 245. (B) Similar to (A), except for overall recall. (C) Similar to (A), except for the mAP at the IoU at 0.5:0.95. (D) Overall Model precision, recall, and the mAP (IoU = 0.5:0.95) for each label class.



Figures 8. Normalized confusion matrix per-class classification across the 4 label classes in the young marmoset recognition model.

Proportion was generated by the (A) training dataset and the (B) validation dataset, showing whether certain classes were frequently mislabelled as a different class.

By combining the identification of marmoset faces and collar beads, the program showed valid detection of the correct identity of individual marmosets. However, the detection of the twin marmosets also posed various challenges for computer vision and resulted in mislabeling, including highly similar faces of twin marmosets, face blur or mislabeling due to complex backgrounds and dark lighting conditions (Supplementary Video 2 [↗](#)), fur occluding collar beads, marmosets only showing faces during a short time intervals (Supplementary Video 5 [↗](#)), similar or same color of collar beads between two marmosets (Supplementary Video 4 [↗](#), Supplementary Video 5 [↗](#)), and fewer training data (young marmosets' model only included 449 training and validation images for two marmosets at 7 months old).

Facial similarities of marmosets affect the performance of the facial detection and identification model

We quantified the inter-individual face similarities using cosine similarity and Euclidean distance between mean normalized embeddings of marmoset pairs. We included both adult and young marmosets' dataset and computed the face similarity based on the 4 relationship pairs: 1) mother-father; 2) father-son; 3) mother-son; and 4) twin-twin. To visualize across the two independently trained model, the inter-individual face similarity scores were normalized using z-score within each model (Figure 9 [↗](#)).

Within the model trained using the adult marmoset family, both face similarity measures showed a consistent pattern of across the family relationships (Supplementary Figure 4 [↗](#), Supplementary Figure 5 [↗](#), Supplementary Figure 6 [↗](#), Supplementary Figure 7 [↗](#)). In the adult family, we found that father-son pair showed the highest face similarity values, with the cosine similarity at a positive mean z score ($z = 0.43$) and the Euclidean distance at a negative mean z score ($z = -0.47$) (Figure 9A [↗](#) and B [↗](#)). This pattern suggested that father-son pair exhibited a higher-than-average similarity relative to other family relationship. On the other hand, mother-father relationship showed the lowest face similarity in the family (negative cosine similarity $z = -0.38$, positive Euclidean distance $z = 0.43$) compared to the adult marmosets' model mean. The similarity of the mother-son pair was calculated at cosine similarity $z = -0.05$ and Euclidean distance $z = 0.04$, indicating an intermediate face similarity in the adult family.

The young marmosets' model was independently trained specifically on the two twin marmosets. Thus, the normalized face similarity z scores were centered near zero and not informative of the twin similarity in family relationship comparison (Figure 9 [↗](#)). However, we found a narrow distribution of the image embeddings between the twin marmosets, indicating a nearly invariant facial embeddings and structure between the two twins, with an extremely low raw variance for cosine similarity (std = 0.0005) and Euclidean distance (std = 0.0034).

We performed the statistical tests only on the adult marmoset family model, as the twin marmoset model only involved two individuals and thus not valid for within-model statistical analysis (Table 2 [↗](#), 3 [↗](#)). Cosine similarity showed a trend of lower similarity in mother-father pair compared to the father-son pair ($t = -1.941$, $p = 0.083 > 0.05$, Cohen's $d = -0.868$ (large effect)), though not significant. Differences between other relationship pairs were not significant (Table 2 [↗](#)). Our analysis on the Euclidean distance indicated a significant difference between mother-father and father-son pairs ($t = 2.28$, $p = 0.046 < 0.05$, $d = 1.02$ (large effect)), while the remaining relationship pairs appeared non-significant (Table 3 [↗](#)).

A real-time marmoset identification program, based on the trained networks

We developed a real-time interface for detecting and identifying the marmosets, based on the trained models of multi-marmoset classification. The real-time footages were acquired using the same camera and experimental setup as the training videos (see examples in Supplementary Video 1 [↗](#) – 5 [↗](#)). Our real-time program processes the real-time footage like the detection of the offline collected videos, using the best model selected from each multi-marmoset face classification training model. Prior to running the real-time program, the experimenter first trains the multi-

Figures 9. Across-model visualization of the face similarity between marmoset pairs.

Four types of family relationships (mother-father, father-son, mother-son, and twin1-twin2) were compared, based on the training results of adult and young marmosets. The similarity was calculated using (A) cosine similarity and the (B) Euclidean distance.

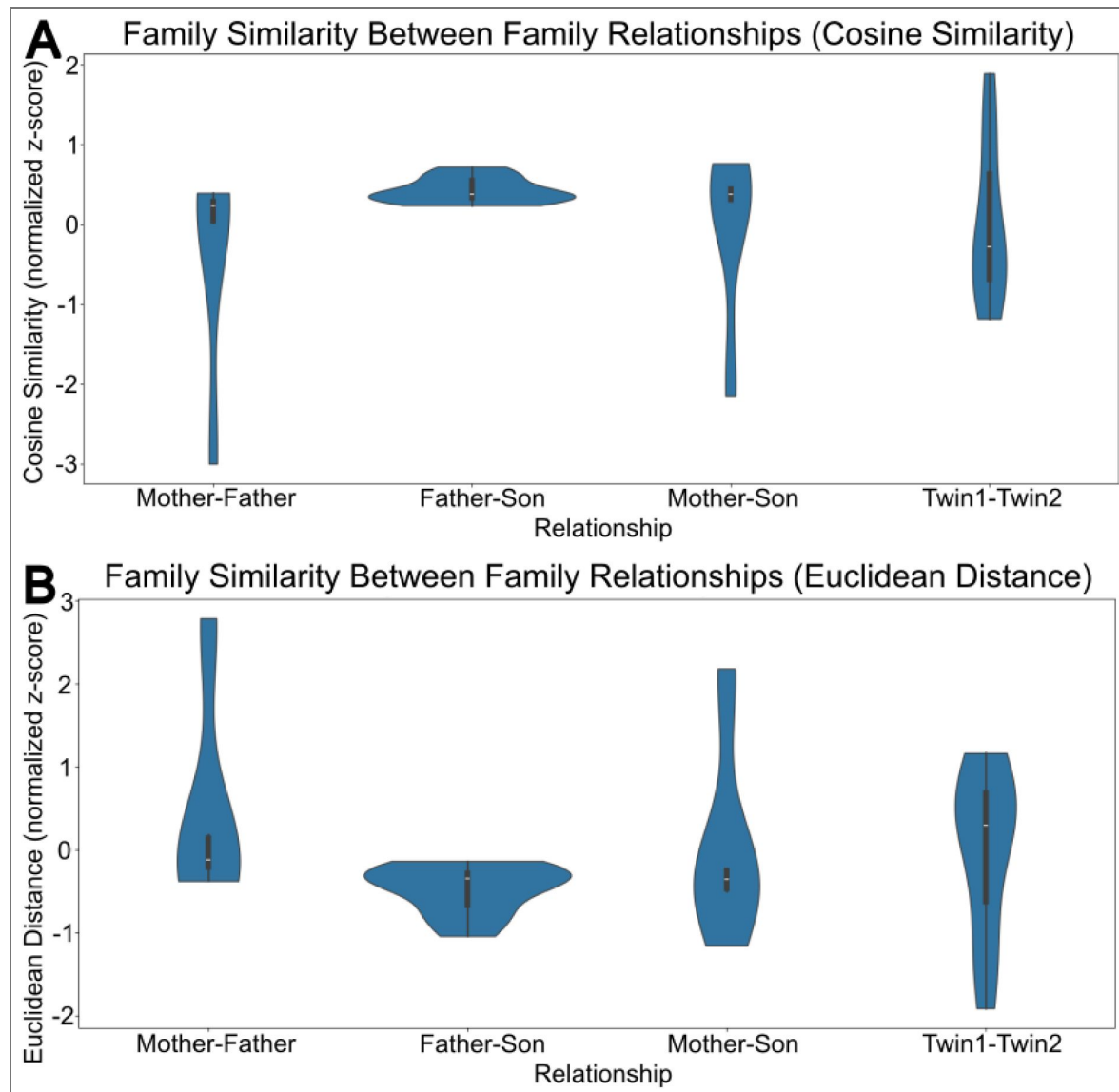


Table 2. Statistical analysis of inter-individual face similarity (cosine similarity) between different family relationships, within the adult marmoset family.

Comparison of t-statistics, p-value, and Cohen's d on cosine similarities were compared between the 3 family relationships. The relationship pairs that showed significant differences were bolded.

Relationships	t-value	p-value	Cohen's d
Mother-father vs. Father-son	-1.941	0.083	-0.868
Mother-father vs. Mother-son	-0.613	0.548	-0.274
Father-son vs. Mother-son	1.460	0.177	0.653

Relationships	t-value	p-value	Cohen's d
Mother-father vs. Father-son	2.276	0.046	1.018
Mother-father vs. Mother-son	0.747	0.465	0.334
Father-son vs. Mother-son	-1.397	0.192	-0.625

Table 3. Statistical analysis of inter-individual face similarity (Euclidean distance) between different family relationships, within the adult marmoset family.

Comparison of t-statistics, p-value, and Cohen's d on Euclidean distances were compared between the 3 family relationships. The relationship pairs that showed significant differences were bolded.

marmoset face classification model based on the specific subjects of interest. Next, the model with the best performance from training are selected and inputted into the real-time program. No programming is required from the experimenter using this real-time marmoset face identification program. The experimenter clicks a run button on the program console to initiate the real-time detection. While this program is running, the experimenter can check the frame-by-frame label detection results (30 frames per second) displayed on the screen. Potential mislabeling are mitigated by combining the marmoset faces and collar beads detection, with greater weights assigned to the collar bead detections than the marmoset face (example mislabelling in [Supplementary Video 4](#) and [Supplementary Video 5](#)). The program automatically records the most frequently detected marmoset identity across 30 frames (approximately one second). At the end of the experiment, the program is terminated by clicking a stop button of the python console. After the completion of the experiment, the full record of marmoset identity detection results is available for review by the experimenter.

Discussion

We developed a real-time computer vision program for automatic marmoset identity recognition, using their facial features and uniquely color-coded bead collars. By combining the automatic annotation of the real-time footage with the individual-specific classification (i.e. faces and collar beads), our program allows continuous identity tracking during behavioral experiments, with minimal human interference. Utilizing the object detection YOLOv8 model based on pre-trained image networks, we adapted it to a non-human primate (i.e. marmosets) face dataset and applied it in real-time tracking. Pose estimation tools have been widely used in characterizing animal behavior; however, these tools require extensive computational power thus limiting the identification between visually similar individuals, especially animals housed in family unit ([Camilleri et al., 2023](#); [Gill et al., 2025](#); [Lauer et al., 2022](#); [Mathis et al., 2018](#); [Nath et al., 2019](#)). Notably, the main goal of the current system is marmoset identity recognition, instead of pose estimation or behavioral analysis, such that identity characterization is not dependent on posture cues or movement tracking. Thus, we selected the YOLOv8 object detection algorithms for the development of our real-time marmoset face recognition pipeline. Among the YOLOv8 pre-trained models, we selected the light-weighted YOLOv8 nano model, which provided the optimal balance between detection accuracy and computational inference speed, supporting its feasibility for the final real-time marmoset identity recognition program.

Our real-time marmoset recognition pipeline includes two face classification models for both adult and young marmoset faces, with an additional automatic face and identity extraction model for all marmosets. The models presented in this paper achieved robust detection accuracy across adult marmosets' and twin marmosets' datasets, with detection accuracy improving with increased amount of training images. While comparing between the adult and young marmoset datasets, we found that the adult marmosets' model, trained with a larger number of varied images, showed more reliability and efficiency in marmoset identity recognition. Moreover, we anticipate that experimenters can integrate this program into behavioral experiments, as an automated marmoset identity extraction and classification tool for real-time video monitoring. Once fully trained on recognizing subject marmosets, this pipeline operates automatically and can be applied across individuals of different ages, with minimal manual work needed. This automated pipeline substantially reduces the time and work required for traditional manual identity labeling, while maintaining an expert-level human performance and reproducibility across experimenters. The pipeline's advantages are particularly efficient for large datasets and longitudinal studies, where manual identity labeling becomes difficult, as variability and errors increase along with dataset size and experimenter number.

The motivation for this real-time marmoset identity recognition program was to develop an easy-to-use, generalizable pipeline that could be applied across different marmosets and lab environments. The pipeline was designed to have no specific hardware requirements and can be implemented for any standard recording device, including any commonly available cameras, primate chair system, and computer-based device. Based on the collected individual marmoset

face data, automatic extraction program allows consistent identity annotation of marmoset faces and collar beads, ensuring accurate and stable identity interpretation during the marmoset face classification model training. The experimenter only needs to run the Python pipeline scripts to perform the real-time marmoset identification during experiments, without the need to manually label large marmoset identity datasets. Moreover, the system operates directly based on the raw video input from the camera, with no pre-processing such as video cropping or resolution modification required. The identity recognition system recognizes each marmoset based on pre-defined labels from collected videos of individuals, without other external tools such as Radio Frequency Identification (RFID) systems (Pereira et al., 2023 [DOI](#)). We focused the marmoset identification on individual-specific facial features and uniquely colored collar beads to ensure the pipeline robustness across various recording conditions, regardless of lighting condition, camera angle, or animal posture. Together, this pipeline design and setup enable fast, reliable, and accurate identity recognition for an efficient real-time monitoring of multiple animals in complex experimental environments. However, with the high visual similarity between closely related marmoset family members, the facial features and the collar beads must be carefully integrated in our pipeline design, as even experienced human observers can misidentify marmoset twins.

Because of the high extent of natural facial similarity between related marmosets, our real-time pipeline can experience challenges in identifying closely related individuals, as the results in Figure 7 [DOI](#) - 9 [DOI](#), and Supplementary Video 2 [DOI](#) - 5 [DOI](#) indicate. Fur textures, coloring patterns, and genetic relatedness contribute to the morphological similarities, making it challenging for human observers when identifying individual primates (Alvergne et al., 2009 [DOI](#); Guan et al., 2023 [DOI](#); Leopold and Rhodes, 2010 [DOI](#)). To overcome this challenge, we implemented the uniquely color-coded bead collars as the complementary cue for marmoset identity prediction. As the collar beads label class were detected with high precision, shown in our results in Figure 3 [DOI](#), 5 [DOI](#), and 7 [DOI](#), it presents as a constant marker as marmosets age, aiming to improve the detection accuracy of marmoset identities. Nonetheless, identity mislabeling could persist when there are collar beads visibility issues. For example, this issue can be caused by identical or similar bead colors between individuals (Supplementary Video 4 [DOI](#), 5 [DOI](#)) and beads that are occluded by fur (Supplementary Video 2 [DOI](#) - 5 [DOI](#)). An alternate experimental solution is to improve collar visibility, including using distinct color code across individuals within a family and increasing collar beads number to reduce occlusion. Moreover, the model performance of the system is strongly dependent on the amount and variability of the training data, with identity classification improving as more marmoset images are involved in the model training. This relation is specifically important when characterizing between young and adult marmosets, as facial features may change across marmosets' developmental stages. Images from both the juvenile and adult periods of the same marmoset can be included in the model training, which could improve model performance and generalizability across developmental stages, and avoid repeated training of the individual recognition model at different ages.

Because our pipeline was designed specifically for marmoset identity recognition, it was optimized for characterizing individuals within a defined housing unit, which usually represents as a family or pair group. With one separated model trained per family unit, our system can assign distinctly colored collar beads to different family members. Even though multiple marmosets with visually similar faces may present close to the camera, the distinctly different color beads can be used for accurate prediction, supporting the reliability of identity recognition. Moreover, while changes in the recording environment setup (e.g. lighting condition, camera angles, etc.) does not affect the model performance, the model needs to be adjusted when family members or composition changes. In particular, the introduction of new members (e.g. newborns) requires assignment of new colored collar beads and collection of the face images, thus it is necessary that the recognition model of this marmoset family has to be retrained.

This system may be particularly useful when experimenters need to know which individual is performing a specific task for cognitive and behavioral experiments (Kangas et al., 2016 [DOI](#); Kangas and Bergman, 2017 [DOI](#); Marshall and Ridley, 2003 [DOI](#)). In these experiments, especially those involving long-term or continuous behavioral responses, experimenters often need to be present

to record animal identity or to manually review video recordings after the experiment. This may distract the animals from their task and affect their behaviors, while also demanding sustained attention and large time commitment for the experimenters. Our pipeline resolves this limitation by automatically detecting and recording marmoset identity throughout the experiment. A typical system requires the experimenter to only collect video clips from each marmoset, verify the automatically annotated marmoset faces and collar beads, and initiating the training of a family-specific face classification model. This process takes approximately 20 - 30 hours of time. Once trained, the system operates automatically to collect real-time identity and can work to present subject-specific tasks based on the identity of the detected animal, with no work or presence needed on the user's end.

Future extensions can further improve the detection accuracy and the pipeline utility. Firstly, the pipeline can be easily combined with additional programs to characterize the detailed facial features of the marmosets (Correia-Caeiro et al., 2022 [↗](#); Kawaguchi et al., 2023 [↗](#)). Considering differences in their eye distance, mouth shape, and fur coloring pattern, in addition to the global facial structure, we can improve the identity prediction, particularly between two closely related marmosets with similar faces. Also, aside of the colored collar beads, experimenters can label marmosets' identities using other visual markers, including color dyes on marmoset ear tufts. Using a more evident visual marker could be helpful for accurate identity prediction and avoid mislabeling. Moreover, it would be possible to integrate pose estimation tools, such as DeepLabCut and MarmoPose (Cheng et al., 2025 [↗](#); Lauer et al., 2022 [↗](#)), with this marmoset recognition pipeline, thus allowing the estimation of the postures and behaviors of specific animals of interest. While we developed this pipeline and demonstrate its utility for common marmosets in laboratory captivity, there is no application restriction of this system. With appropriate training data and experimental design, our pipeline can be applied to other non-human primates in various settings such as lab housing, conservation fields, or even in the wild.

Data availability

All code in this paper is publicly available at Github: <https://github.com/Jy-Yang-bot/real-time-marmoset-recognition> [↗](#). This repository contains the scripts for pre-processing the images, the main facial recognition pipeline, and the analytic tools.

Acknowledgements

We would like to thank our animal health technicians D. Hau-Aquino, V. Comtois, C. Hunt; and veterinarians F. Chaurand, and J. Hutta for animal health care and support. We are thankful to M. Gacoin and T. Cook for helpful discussions and feedback on the visualization and format of the manuscript. We thank J. Smith and C. O'Hare-Freire for the design and construction of transparent protective case for the camera. We acknowledge the support of the Government of Canada's New Frontiers in Research Fund (NFRF), [NFRFT-2022-00051] and by the Fonds de Recherche du Québec-Santé (FRQS), [#347426 and #358082]. Ces travaux ont bénéficié d'un octroi des fonds Nouvelles frontières en recherche du gouvernement du Canada [NFRFT-2022-00051] et du Fonds de recherche du Québec-Santé [FRQS, #347426 et #358082].

Additional information

Author contributions

Jiayue Yang: Conceptualization, pipeline development, data collection and curation, dataset labeling, formal analysis, investigation, visualization, methodology, writing – original draft, and writing – review and editing.

James Wang: Data collection and curation, writing - review and editing

Justine Cléry: Conceptualization, methodology, writing – review and editing, project administration, supervision, resources, funding acquisition.

Funding

Funder	Grant reference number	Author
New Frontiers in Research Fund	NFRFT-2022-00051	Justine Cléry
Fonds de recherche du Québec (FRQ)	https://doi.org/10.69777/347426	Justine Cléry
Fonds de recherche du Québec (FRQ)	358082	Justine Cléry

Author ORCID iDs

Jiayue Yang: <https://orcid.org/0009-0003-3727-9586>

Justine Cléry: <https://orcid.org/0000-0003-1020-1845>

Additional files

Supplementary Figure 1–6 [↗](#)

Supplementary Figure 7, Supplementary Table 1–3 [↗](#)

Supplementary Video 1 [↗](#) Example annotations of the automatic face and identity extraction model from two unseen marmosets at different age. Top panel: annotated output video of the model. Bottom panel: timeline plot of the detected label classes of marmoset faces (blue) and collar (cyan) beads.

Supplementary Video 2 [↗](#) Example annotations of the multi-marmoset face classification model for Young1 marmoset at 7 months. Top panel: annotated output video of the model. Bottom panel: timeline plot of the detected label classes of marmoset Young1 face (blue), Young1 collar beads (cyan), Young2 (pink), and Young2 collar beads (lime green).

Supplementary Video 3 [↗](#) Example annotations of the multi-marmoset face classification model for Young2 marmoset at 7 months. Top panel: annotated output video of the model. Bottom panel: timeline plot of the detected label classes of marmoset Young1 face (blue), Young1 collar beads (cyan), Young2 (pink), and Young2 collar beads (lime green).

Supplementary Video 4 [↗](#) Example annotations of the multi-marmoset face classification model for Young1 marmoset at 11 months. Top panel: annotated output video of the model. Bottom panel: timeline plot of the detected label classes of marmoset Young1 face (blue), Young1 collar beads (cyan), Young2 (pink), and Young2 collar beads (lime green).

Supplementary Video 5 [↗](#) Example annotations of the multi-marmoset face classification model for Young2 marmoset at 11 months. Top panel: annotated output video of the model. Bottom panel: timeline plot of the detected label classes of marmoset Young1 face (blue), Young1 collar beads (cyan), Young2 (pink), and Young2 collar beads (lime green).

References

- Aghdam H.H., Heravi E.J. (2017) *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification* Cham: Springer. <https://doi.org/10.1007/978-3-319-57550-6>
- Alvergne A., Huchard E., Caillaud D., Charpentier M.J.E., Setchell J.M., Ruppli C., Féjan D., Martinez L., Cowlshaw G., Raymond M. (2009) Human Ability to Recognize Kin Visually Within Primates. *Int J Primatol* **30**:199-210 <https://doi.org/10.1007/s10764-009-9339-0>
- Bakana S.R., Zhang Y., Twala B. (2024) WildARe-YOLO: A lightweight and efficient wild animal recognition model. *Ecological Informatics* **80**:102541 <https://doi.org/10.1016/j.ecoinf.2024.102541>
- Bergman N., Yitzhaky Y., Halachmi I. (2024) Biometric identification of dairy cows via real-time facial recognition. *animal* **18**:101079 <https://doi.org/10.1016/j.animal.2024.101079>
- Bollegala D. (2017) Dynamic feature scaling for online learning of binary classifiers. *Knowledge-Based Systems* **129**:97-105 <https://doi.org/10.1016/j.knosys.2017.05.010>
- Bradski G. (2000) The OpenCV Library. *Dr. Dobb's Journal of Software Tools*

- Buchan J.C., Alberts S.C., Silk J.B., Altmann J. (2003) True paternal care in a multi-male primate society. *Nature* **425**:179-181 <https://doi.org/10.1038/nature01866>
- Camilleri M.P.J., Zhang L., Bains R.S., Zisserman A., Williams C.K.I. (2023) Persistent animal identification leveraging non-visual markers. *Machine Vision and Applications* **34**:68 <https://doi.org/10.1007/s00138-023-01414-1>
- Carstens E., Moberg G.P. (2000) Recognizing Pain and Distress in Laboratory Animals. *ILAR Journal* **41**:62-71 <https://doi.org/10.1093/ilar.41.2.62>
- Cheng C., Huang Z., Zhang R., Huang G., Wang H., Tang L., Wang X. (2025) A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates. *Cell Reports Methods* **5**:e100986 <https://doi.org/10.1016/j.crmeth.2025.100986>
- Correia-Caeiro C., Burrows A., Wilson D.A., Abdelrahman A., Miyabe-Nishiwaki T. (2022) CalliFACS: The common marmoset Facial Action Coding System. *PLoS ONE* **17**:e0266442 <https://doi.org/10.1371/journal.pone.0266442>
- Dave B., Mori M., Bathani A., Goel P. (2023) Wild Animal Detection using YOLOv8. *Procedia Computer Science* **230**:100-111 <https://doi.org/10.1016/j.procs.2023.12.065>
- Dehmer M., Basak S.C. (2012) *Statistical and Machine Learning Approaches for Network Analysis* (1) Wiley. <https://doi.org/10.1002/9781118346990>
- Dumoulin V., Visin F. (2016) A guide to convolution arithmetic for deep learning. *arXiv* <https://doi.org/10.48550/ARXIV.1603.07285>
- Gill H., Charles J., Grant R., Gardiner J., Bates K., Brassey C. (2025) Gait tracking in dogs using DeepLabCut: A markerless machine learning approach for controlled settings. *Applied Animal Behaviour Science* **292**:106769 <https://doi.org/10.1016/j.applanim.2025.106769>
- Guan Y., Lei Y., Zhu Y., Li T., Xiang Y., Dong P., Jiang R., Luo J., Huang A., Fan Y., et al. (2023) Face recognition of a Loriscidae species based on computer vision. *Global Ecology and Conservation* **45**:e02511 <https://doi.org/10.1016/j.gecco.2023.e02511>
- He K., Zhang X., Ren S., Sun J. (2015) Deep Residual Learning for Image Recognition. *arXiv* <https://doi.org/10.48550/ARXIV.1512.03385>
- Jocher G., Chaurasia A., Qiu J. (2023) Ultralytics YOLO.
- Jozwik K.M., O'Keeffe J., Storrs K.R., Guo W., Golan T., Kriegeskorte N. (2022) Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proc. Natl. Acad. Sci. U.S.A* **119**:e2115047119 <https://doi.org/10.1073/pnas.2115047119>
- Kangas B.D., Bergman J. (2017) Touchscreen technology in the study of cognition-related behavior. *Behavioural Pharmacology* **28**:623-629 <https://doi.org/10.1097/FBP.0000000000000356>
- Kangas B.D., Bergman J., Coyle J.T. (2016) Touchscreen assays of learning, response inhibition, and motivation in the marmoset (*Callithrix jacchus*). *Anim Cogn* **19**:673-677 <https://doi.org/10.1007/s10071-016-0959-4>
- Kawaguchi Y., Nakamura K., Tajima T., Waller B.M. (2023) Revisiting the baby schema by a geometric morphometric analysis of infant facial characteristics across great apes. *Sci Rep* **13**:5129 <https://doi.org/10.1038/s41598-023-31731-4>
- Kishi N., Sato K., Sasaki E., Okano H. (2014) Common marmoset as a new model animal for neuroscience research and genome editing technology. *Dev Growth Differ* **56**:53-62 <https://doi.org/10.1111/dgd.12109>
- Koski S.E., Burkart J.M. (2015) Common marmosets show social plasticity and group-level similarity in personality. *Sci Rep* **5**:8878 <https://doi.org/10.1038/srep08878>
- Lauer J., Zhou M., Ye S., Menegas W., Schneider S., Nath T., Rahman M.M., Di Santo V., Soberanes D., Feng G., et al. (2022) Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat Methods* **19**:496-504 <https://doi.org/10.1038/s41592-022-01443-0>

- Lee W., Ham B., Kim S. (2025) Maximizing the Position Embedding for Vision Transformers with Global Average Pooling. *arXiv* <https://doi.org/10.48550/ARXIV.2502.02919>
- Leopold D.A., Rhodes G. (2010) A comparative view of face perception. *Journal of Comparative Psychology* **124**:233-251 <https://doi.org/10.1037/a0019460>
- Li Baoli, Han L., Distance Weighted Cosine Similarity Measure for Text Classification, Yin H., Tang K., Gao Y., Klawonn F., Lee M., Weise T., Li Bin, *et al.* (2013) *Intelligent Data Engineering and Automated Learning – IDEAL 2013, Lecture Notes in Computer Science* Berlin, Heidelberg: Springer. pp. 611-618 https://doi.org/10.1007/978-3-642-41278-3_74
- Lim M.A., Defensor E.B., Mechanic J.A., Shah P.P., Jaime E.A., Roberts C.R., Hutto D.L., Schaevitz L.R. (2019) Retrospective Analysis of the Effects of Identification Procedures and Cage Changing by Using Data from Automated, Continuous Monitoring. *J Am Assoc Lab Anim Sci* **58**:126-141 <https://doi.org/10.30802/AALAS-JAALAS-18-000056>
- Lin T.-Y., Maire M., Belongie S., Bourdev L., Girshick R., Hays J., Perona P., Ramanan D., Zitnick C.L., Dollár P. (2015) Microsoft COCO: Common Objects in Context. *arXiv*
- Long Chai J.C., Ng T.-S., Low C.-Y., Park J., Jin Teoh A.B. (2023) Recognizability Embedding Enhancement for Very Low-Resolution Face Recognition and Quality Estimation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada. pp. 9957-9967 <https://doi.org/10.1109/CVPR52729.2023.00960>
- Marion S., Davies A., Demšar U., Irvine R.J., Stephens P.A., Long J. (2020) A systematic review of methods for studying the impacts of outdoor recreation on terrestrial wildlife. *Global Ecology and Conservation* **22**:e00917 <https://doi.org/10.1016/j.gecco.2020.e00917>
- Marshall J.W.B., Ridley R.M. (2003) Assessment of Cognitive and Motor Deficits in a Marmoset Model of Stroke. *ILAR Journal* **44**:153-160 <https://doi.org/10.1093/ilar.44.2.153>
- Mathis A., Mamidanna P., Cury K.M., Abe T., Murthy V.N., Mathis M.W., Bethge M. (2018) DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**:1281-1289 <https://doi.org/10.1038/s41593-018-0209-y>
- Maxwell A.E., Warner T.A., Guillén L.A. (2021) Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 1: Literature Review. *Remote Sensing* **13**:2450 <https://doi.org/10.3390/rs13132450>
- Merchant F.A., Shah S.K., Castleman K.R. (2023) Object Measurement. In: Merchant F.A., Castleman K.R. (Eds). *Microscope Image Processing* Elsevier. pp. 153-175 <https://doi.org/10.1016/B978-0-12-821049-9.00017-4>
- Miller C.T., Freiwald W.A., Leopold D.A., Mitchell J.F., Silva A.C., Wang X. (2016) Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron* **90**:219-233 <https://doi.org/10.1016/j.neuron.2016.03.018>
- Nath T., Mathis A., Chen A.C., Patel A., Bethge M., Mathis M.W. (2019) Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat Protoc* **14**:2152-2176 <https://doi.org/10.1038/s41596-019-0176-0>
- National Research Council (U.S.) (2011) *Guide for the care and use of laboratory animals* (8) Washington, D.C: National Academies Press.
- Nguyen H.V., Bai L., Cosine Similarity Metric Learning for Face Verification, Kimmel R., Klette R., Sugimoto A. (2011) *Computer Vision – ACCV 2010, Lecture Notes in Computer Science* Berlin, Heidelberg: Springer. pp. 709-720 https://doi.org/10.1007/978-3-642-19309-5_55
- Norouzzadeh M.S., Morris D., Beery S., Joshi N., Jojic N., Clune J. (2021) A deep active learning system for species identification and counting in camera trap images. *Methods Ecol Evol* **12**:150-161 <https://doi.org/10.1111/2041-210X.13504>
- Norouzzadeh M.S., Nguyen A., Kosmala M., Swanson A., Palmer M.S., Packer C., Clune J. (2018) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. U.S.A* **115** <https://doi.org/10.1073/pnas.1719367115>

- Okano H. (2021) Current Status of and Perspectives on the Application of Marmosets in Neurobiology. *Annu. Rev. Neurosci* **44**:27-48 <https://doi.org/10.1146/annurev-neuro-030520-101844>
- Padilla R., Passos W.L., Dias T.L.B., Netto S.L., Da Silva E.A.B. (2021) A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit. *Electronics* **10**:279 <https://doi.org/10.3390/electronics10030279>
- Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., et al. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**:2825-2830
- Pereira E., Araújo I'., Silva L.F.V., Batista M., Júnior S., Barboza E., Santos E., Gomes F., Fraga I.T., Davanso R., et al. (2023) RFID Technology for Animal Tracking: A Survey. *IEEE J Radio Freq Identif* **7**:609-620 <https://doi.org/10.1109/JRFID.2023.3334952>
- Petso T., Jamisola R.S., Mpoeleng D., Bennitt E., Mmereki W. (2021) Automatic animal identification from drone camera based on point pattern analysis of herd behaviour. *Ecological Informatics* **66**:101485 <https://doi.org/10.1016/j.ecoinf.2021.101485>
- Powers D. (2008) Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach Learn Technol* **2**
- Redmon J., Divvala S., Girshick R., Farhadi A. (2015) You Only Look Once: Unified, Real-Time Object Detection. *arXiv* <https://doi.org/10.48550/ARXIV.1506.02640>
- Rezatofighi H., Tsoi N., Gwak J., Sadeghian A., Reid I., Savarese S. (2019) Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. *arXiv* <https://doi.org/10.48550/ARXIV.1902.09630>
- Roughan J.V., Sevenoaks T. (2019) Welfare and Scientific Considerations of Tattooing and Ear Tagging for Mouse Identification. *J am assoc lab anim sci* **58**:142-153 <https://doi.org/10.30802/AALAS-JAALAS-18-000057>
- Sasaki E., Suemizu H., Shimada A., Hanazawa K., Oiwa R., Kamioka M., Tomioka I., Sotomaru Y., Hirakawa R., Eto T., et al. (2009) Generation of transgenic non-human primates with germline transmission. *Nature* **459**:523-527 <https://doi.org/10.1038/nature08090>
- Schindler F., Steinhage V. (2021) Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecological Informatics* **61**:101215 <https://doi.org/10.1016/j.ecoinf.2021.101215>
- Schofield D.P., Albery G.F., Firth J.A., Mielke A., Hayashi M., Matsuzawa T., Biro D., Carvalho S. (2023) Automated face recognition using deep neural networks produces robust primate social networks and sociality measures. *Methods Ecol Evol* **14**:1937-1951 <https://doi.org/10.1111/2041-210X.14181>
- Sekachev B., Manovich N., Zhiltsov M., Zhavoronkov A., Kalinin D., Hoff B., TOSmanov, Kruchinin D., Zankevich A., DmitriySidnev, et al. (2020) opencv/cvat. Zenodo. version: v1.1.0 <https://doi.org/10.5281/ZENODO.4009388>
- Sturman O., Von Ziegler L., Schläppi C., Akyol F., Privitera M., Slominski D., Grimm C., Thieren L., Zerbi V., Grewe B., et al. (2020) Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacol* **45**:1942-1952 <https://doi.org/10.1038/s41386-020-0776-y>
- Sugase-Miyamoto Y., Matsumoto N., Ohya K., Kawano K. (2014) Face Inversion Decreased Information about Facial Identity and Expression in Face-Responsive Neurons in Macaque Area TE. *J. Neurosci* **34**:12457-12469 <https://doi.org/10.1523/JNEUROSCI.0485-14.2014>
- Thongtan T., Phienthrakul T. (2019) Sentiment Classification Using Document Embeddings Trained with Cosine Similarity. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy. pp. 407-414 <https://doi.org/10.18653/v1/P19-2057>

- Varghese R., Sambath M. (2024) YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). Chennai, India. pp. 1-6
<https://doi.org/10.1109/ADICS58448.2024.10533619>
- Vidal M., Wolf N., Rosenberg B., Harris B.P., Mathis A. (2021) Perspectives on Individual Animal Identification from Biology and Computer Vision. *Integrative and Comparative Biology* **61**:900-916
<https://doi.org/10.1093/icb/icab107>
- XiaoAn B., LinQing Z., XiaoMei T., Biao W., QingQi Z., YuTing J., Na Z. (2024) Wildlife target detection based on improved YOLOX-s network. *Sci Rep* **14**:23608 <https://doi.org/10.1038/s41598-024-73631-1>
- Yoshimoto T., Takahashi E., Yamashita S., Ohara K., Niimi K. (2018) Larger cages with housing unit environment enrichment improve the welfare of marmosets. *Exp. Anim* **67**:31-39
<https://doi.org/10.1538/expanim.17-0058>
- Yu L., Chen X., Zhou S. (2018) Research of Image Main Objects Detection Algorithm Based on Deep Learning. In: 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). Chongqing, China. pp. 70-75 <https://doi.org/10.1109/ICIVC.2018.8492803>
- Zhuang Y., Liu J., Zhao H., Ma L., Fang Z., Li L., Wu C., Cui W., Liu Z. (2025) A deep learning framework based on structured space model for detecting small objects in complex underwater environments. *Commun Eng* **4**:24 <https://doi.org/10.1038/s44172-025-00367-9>

Peer reviews

Reviewer #1 (Public review):

Summary:

The manuscript by Yang, Wang, and Cléry presents a lightweight pipeline for real-time identification of common marmosets in a laboratory setting. Models were trained and evaluated on data derived from a family of three closely related adults and a set of juvenile twins. Freely moving animals entered an enclosed space fixed to the housing cage door, which permitted the entry of individual animals for data acquisition. Utilizing YOLOv8-nano, identification was improved through the introduction of uniquely colored collar beads. Analyses of facial similarity showed close morphological relatedness amongst individuals and highlighted the need for highly discriminative classification. Overall, the authors offer a framework for identity tracking that prioritizes real-time inference. The authors demonstrate that combining facial detection with visual markers enables adequate identity assignment under controlled laboratory conditions with minimal cross-individual misclassification.

Strengths:

- (1) The proposed pipeline offers a solution for real-time identity tracking in common marmosets. Its lightweight design enables deployment across a wide range of hardware configurations. Furthermore, if similar strategies are employed, this methodology is likely adaptable for other species with minimal modification.
- (2) Evaluation of closely related individuals provides a necessary stress test for the discrimination of facial identity tracking.

Weaknesses:

- (1) The pipeline's reliance on controlled animal isolation and small visual markers raises questions about the approach's generalizability to unconstrained multi-animal environments. The provided confusion matrices (Figures 6-8) indicate that the most common misclassifications are background-related, possibly suggesting that detection specificity is the

primary source of error. All things considered, these findings raise concerns about performance in its use in socially dynamic and visually complex environments.

(2) The manuscript claims performance comparable to that of human experimenters but provides no explicit evidence to support these claims. While it is plausible that human experimenters may be less accurate in facial recognition tasks involving closely related marmosets, the authors don't provide evidence. Moreover, while that might be the case, the color-coded beads provide a salient identity cue for the model, which complicates the interpretation of this comparison grounded in facial recognition.

<https://doi.org/10.7554/eLife.110932.1.sa2>

Reviewer #2 (Public review):

Summary:

In this study, Yang et al. develop a real-time system for automatic face detection and identification of multiple unrestrained common marmosets in a home cage setting.

Strengths:

The study aims to address an unmet need in behavioral neuroscience: the ability to non-invasively identify animals is crucial to the automated and rigorous study of neural behaviors; this is especially true for common marmosets, which are rapidly becoming a model system of choice for the study of complex social cognition. By using a YOLOv8 backbone, the study achieve human level performance, both in terms of precision and recall of the trained models.

Weaknesses:

The robustness of the system is not clear from the limited datasets presented. The use of color-coded beads undercuts the study's premise that the system achieves truly non-invasive tracking. Although the system achieves good performance in face detection, it does not perform as well for classification using faces alone (especially when the faces are similar, as in twin animals). Here, too, the color-coded beads play a key role in identity discrimination. The stated goals of the study and the actual results presented are therefore at odds.

<https://doi.org/10.7554/eLife.110932.1.sa1>

Reviewer #3 (Public review):

Summary:

In this manuscript, Yang et al introduce a new method for automatically identifying marmosets in their home cage using a supervised deep learning method that recognizes the face and colored beads on marmoset collars. The authors show a high precision rate of identifying marmosets to levels comparable to a human experimenter. The method overall seems robust at identifying marmosets at different life stages and different settings; however, given the current form, I'm struggling to see the generalizability and experimental utility of this method.

Strengths:

(1) The authors provide a near-perfect automatic identification of marmosets in their home cage.

(2) This method is robust across lightning, camera angles, etc., making it potentially useful for marmoset (and other NHP) identification outside the housing cage as well

Weaknesses:

(1) Despite the almost perfect precision, in its current form, I'm failing to see how this method can be useful to other labs.

(2) This is a nice methods manuscript, but the authors do not present results to show how their method can be used outside of identifying marmosets inside their home cages in a small field of view.

(3) Reading the manuscript is strenuous, given its repetitive nature. Consolidating and shortening the results, as well as adding some definitions to the results section, would be helpful.

<https://doi.org/10.7554/eLife.110932.1.sa0>