Zhang et al. *Zool. Res.* 2025, 46(2): 339−354
https://doi.org/10.24272/j.issn.2095-8137.2024.296
https://cstr.cn/32248.14.zr.2024.296

ZR | Zoological Research

Article Open Access

A deep learning lightweight model for real-time captive macaque facial recognition based on an improved YOLOX model

Jia-Jin Zhang[1,2,#,*], Yu Gao[2,#], Bao-Lin Zhang[1,3,4], Dong-Dong Wu[1,3,5,*]

[1] *Key Laboratory of Genetic Evolution & Animal Models*, *Kunming Natural History Museum of Zoology*, *Kunming Institute of Zoology*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650201, *China*

[2] *College of Big Data*, *Yunnan Agricultural University*, *Kunming*, *Yunnan* 650201, *China*

[3] *National Resource Center for Non-Human Primates*, *Kunming Primate Research Center*, *and National Research Facility for Phenotypic & Genetic Analysis of Model Animals (Primate Facility)*, *Kunming Institute of Zoology*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650107, *China*

[4] *Yunnan Key Laboratory of Biodiversity Information*, *Kunming*, *Yunnan* 650223, *China*

[5] *Center for Excellence in Animal Evolution and Genetics*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650201, *China*

ABSTRACT

Automated behavior monitoring of macaques offers transformative potential for advancing biomedical research and animal welfare. However, reliably identifying individual macaques in group environments remains a significant challenge. This study introduces ACE-YOLOX, a lightweight facial recognition model tailored for captive macaques. ACE-YOLOX incorporates Efficient Channel Attention (ECA), Complete Intersection over Union loss (CIoU), and Adaptive Spatial Feature Fusion (ASFF) into the YOLOX framework, enhancing prediction accuracy while reducing computational complexity. These integrated approaches enable effective multiscale feature extraction. Using a dataset comprising 179 400 labeled facial images from 1 196 macaques, ACE-YOLOX surpassed the performance of classical object detection models, demonstrating superior accuracy and real-time processing capabilities. An Android application was also developed to deploy ACE-YOLOX on smartphones, enabling on-device, real-time macaque recognition. Our experimental results highlight the potential of ACE-YOLOX as a non-invasive identification tool, offering an important foundation for future studies in macaque facial expression recognition, cognitive psychology, and social behavior.

Keywords: YOLOX; Macaque; Facial recognition; Identity recognition; Animal welfare

INTRODUCTION

Non-human primates (NHPs) are phylogenetically similar to humans, sharing numerous physiological, anatomical, immunological, and neurological traits, making them excellent models for biomedical research. Effective management of NHP populations in modern breeding facilities requires the systematic collection of individual data, including birth dates, pregnancy status, health records, parentage, and vaccination history. Such data collection enhances operational efficiency, promotes animal welfare, and reduces labor demands (Cai & Li, 2013).

The growing scale of research facilities necessitates precise and reliable methods for individual animal identification, particularly for NHPs. Accurate identification is critical for maintaining the integrity of research data while enabling personalized care and monitoring, which are fundamental to advancing both research quality and animal welfare. This requirement extends globally, with laboratories worldwide facing similar challenges. However, identification systems for macaques remain limited, highlighting an urgent need for cost-effective, accurate, and practical solutions.

Recognizing individual macaques in group-housed settings is vital for studies focused on social behavior (Sheehan et al., 2014). Research has demonstrated that individual recognition influences key aspects of social dynamics, including territoriality (Tibbetts & Dale, 2007) and mate choice (Gokcekus et al., 2021). Additionally, welfare standards in many countries are shifting toward pair and group housing for NHPs, increasing the demand for effective identification

Received: 10 September 2024; Accepted: 03 December 2024; Online: 04 December 2024

Foundation items: This work was supported by the grants from Yunnan Province (202305AH340006, 202305AH340007) and CAS Light of West China Program (xbzg-zdsys-202213)

[#]Authors contributed equally to this work

*Corresponding authors, E-mail: zjjclc@ynau.edu.cn; wudongdong@mail.kiz.ac.cn

methods that can operate in these complex environments (Harding, 2017).

Traditional identification approaches for group-housed macaques, such as manual observation, depend heavily on the experience and memory of the observer, resulting in low accuracy and high labor intensity (Ait-Saidi et al., 2014). Other methods include the use of tracking devices, such as colored jackets (Rose et al., 2012), collars (Ballesta et al., 2014), or RFID tags (Floyd, 2015). While these methods reduce reliance on human memory, they still present challenges regarding animal welfare due to their invasive nature and associated costs. Conventional marking techniques for animals are not only resource-intensive and risky but are increasingly considered unacceptable from an ethical standpoint (Fernandez-Duque et al., 2018).

Biological image recognition is an efficient, accurate, and non-invasive technique for identifying biological features, which relies primarily on computer vision technology (Corrêa et al., 2019). Facial features, particularly the eyes and nose, serve as a critical source of information about individual animals and serve as a reliable basis for identification (Gao et al., 2021). Recent advances in image-matching methods and machine learning have enabled the successful identification of various species through facial characteristics, including sheep (Billah et al., 2022), pandas (Chen et al., 2020), cattle (Xu et al., 2021), pigs (Marsot et al., 2020), great apes (Ernst & Küblbeck, 2011), lemurs (Crouse et al., 2017), and rhesus macaques (Witham, 2018).

Deep learning, a transformative branch of machine learning, has surpassed traditional shallow learning approaches by automatically extracting features from large datasets, becoming a key area in pattern recognition (Sarker, 2021). Its application in primate identification has yielded remarkable success. Freytag et al. (2016) achieved over 90% accuracy in identifying captive chimpanzees using deep learning techniques. Similarly, Schofield et al. (2019) introduced a Convolutional Neural Network (CNN)-based approach for face detection and recognition in wild chimpanzees, achieving accuracies of 92.5% for individual identification and 96.2% for sex recognition. Subsequent work by the same group utilized deep-learning models to generate association networks among wild chimpanzees (Schofield et al., 2023). Guo et al. (2020) developed an automated system for face detection and identification, achieving 94.1% accuracy. Despite these advancements, most existing studies have focused on primates in wild environments, with limited research on social macaques in captive settings.

You Only Look Once (YOLO) is a single-stage target detection framework designed for real-time applications (Redmon et al., 2016). This architecture integrates object localization and classification into a single operation, enabling the simultaneous identification of object locations and recognition of their categories in one pass (Redmon & Farhadi, 2018). Recent advancements in YOLO variants have demonstrated considerable adaptability across diverse fields. For instance, RS-YOLOX, which combines ECA and CIoU blocks, has been effectively applied to object detection in satellite imagery (Yang et al., 2022). Additionally, Liu et al. (2019) demonstrated that incorporating the ASFF block enhances the performance of the YOLOv3 model, underscoring the potential of these components to improve detection accuracy. These developments informed the design and feasibility of the ACE-YOLOX model proposed in this study.

YOLO-based architecture has gained significant attention in deep learning applications for NHP identification. Studies have utilized YOLO frameworks to detect and identify Japanese macaques (Paulet et al., 2024; Ueno et al., 2022), providing important insights into their utility for NHP research. Building on the improved YOLOX framework (Ge et al., 2021), our study introduces ACE-YOLOX, designed to enable real-time individual recognition of macaques in complex captive environments, which often involve occlusions and minor facial differences. ACE-YOLOX incorporates advanced attention mechanisms and diverse feature extraction structures, designed to achieve superior accuracy, reduced model size, and faster detection speeds. The model was trained on a dataset of 179 400 labeled facial images collected from 1 196 captive macaques at the Kunming Institute of Zoology, Chinese Academy of Sciences. This extensive dataset ensured the robustness and reliability of the identification model. ACE-YOLOX facilitates the precise tracking of known individuals and behavioral monitoring, such as grooming and foraging activities, while effectively reducing the need for repeated capture and anesthesia. These features not only enhance animal welfare but also streamline research processes.

Further analysis and visualization of the macaque dataset validated the efficacy of the proposed model. Experimental results demonstrated that ACE-YOLOX achieved high detection accuracy while being lightweight and efficient enough for deployment on smartphones. This capability significantly reduces the labor and time required for long-term data collection in research settings, while also improving animal welfare, aligning with ethical research practices. The methodology described in this study advances the field of NHP identification by providing a scalable, practical solution for individual recognition based on lightweight deep learning techniques.

MATERIALS AND METHODS

Animals and ethical considerations

This study used 1 196 macaques acquired from the National Major Science and Technology Infrastructure for Pattern Animal Phenotyping and Genetics (Primate Facility), Kunming, Yunnan, China. The facility holds international accreditation from the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) (Yao, 2022). All applicable international, national, and institutional guidelines for the care and use of animals were strictly followed. All animal sample collection protocols complied with the current laws of China. All animal procedures were conducted in accordance with the ethical standards of the Institutional Animal Care and Use Committee of the Kunming Institute of Zoology, Chinese Academy of Sciences (Permit No. IACUC-PE-2022-11-003). All animals were cared for in compliance with the standards of local and national animal welfare organizations for proper animal care.

Experimental data collection

Macaques were housed in groups of 6–10 individuals per enclosure at the primate facility, following international animal welfare regulations. Facial data collection was integrated into routine health examinations to minimize disruptions to the daily lives of the macaques. Familiar caretakers assisted in

this process, ensuring compliance with animal welfare laws governing the acquisition of experimental data from NHPs (Hemsworth et al., 2015). Data collection spanned from 10 October to 15 December 2023, during which videos of 1 196 macaques were recorded. Each individual was filmed for approximately 2 min at 30 frames per second (FPS) with an image resolution of 1 920×1 080 pixels.

To facilitate individual identification, the identification of each macaque was determined using a mark located on its inner thigh, as depicted in Figure 1A. Recordings focused on facial images to maximize feature extraction, with facial angles controlled within a 30–60° range from both the left and right sides, as shown in Figure 1C–E. Additionally, an independent external test set was generated by collecting images with partial obstructions from enclosure fences, as shown in Figure 1B. This test set was specifically designed to evaluate the robustness of the model under challenging conditions, ensuring that the dataset accurately reflected the complex living environments of the macaques.

Dataset preprocessing

The inherently active nature of macaques (He et al., 2016) poses significant challenges during video recording, necessitating the use of the zoom function of a smartphone to ensure clear footage. Video-based data collection often results in a high frame-to-frame similarity, leading to redundancy in the dataset. To address this, Python scripts were used to convert each video file into an image sequence, which was then subjected to a rigorous cleaning process. Images that did not contain macaque faces were excluded, and similar images for each individual were systematically removed to ensure diverse representation of facial appearances. Structural similarity index measurement (SSIM) was applied to identify and eliminate similar and low-variance images (Wang et al., 2004).

During preprocessing, an imbalanced dataset was identified, with some macaques represented by significantly fewer images. To address this issue, a suite of data augmentation techniques was implemented, including mirroring (horizontal and vertical flipping), grid overlay (adding grid patterns to increase visual complexity), white block insertion (adding random white squares or rectangles to occlude parts of the image and simulate missing information), brightness adjustments (both enhancement and reduction), multi-angle rotation (rotating images to capture diverse perspectives), and blurring (smoothing edges to reduce image detail). These techniques enhanced the generalization ability of the model and prevented it from learning insufficient features and overfitting individual features, which reduce model performance (Song et al., 2022). Data augmentation also simulated realistic scenarios with macaque enclosures, as shown in Figure 2. Following these steps, a comprehensive facial dataset comprising 179 400 images from 1 196 macaques was constructed, with an average of 150 facial images per macaque.

Data annotation and facial landmark detection

High-quality annotated datasets are essential for training macaque facial recognition models to achieve reliable performance. However, the annotation processes for large-scale datasets are typically performed manually and consume considerable time and resources. In this study, 179 400 macaque facial images were annotated. To address the challenges of manual annotation at this scale, a semi-automated annotation approach was adopted, as illustrated in Figure 3A and detailed in Table 1.

The workflow began with an annotated subset (VV) of 55 manually labeled macaque face images and a larger non-annotated subset (U) of 1 141 macaque face images. To minimize the labor involved in manual data annotation, an initial model (IM) was trained using the manually annotated subset V, allowing it to learn essential features required for macaque face identification. The IM achieved a mean average precision (mAP) of 80.14%, as shown in the precision-recall (P-R) curve in Figure 3B, demonstrating adequate performance despite the limited size of the training data. Once trained, the IM was applied to the non-annotated subset U to detect facial regions using the YOLO-landmark detector. This process generated bounding boxes (RectBox) matching the macaque faces in U. These annotations were then manually reviewed and categorized into two quality groups: “Good” (U_1), containing accurate RectBox, and “Bad” ($U_2 = U - U_1$), encompassing images with annotation issues, such as significant positional deviations, missing annotations, duplicates, or incorrect fits. For $U_2 \neq \varnothing$, manual correction was applied to the RectBox annotations in U_2. Following this correction, final annotations ($U + V$) were obtained. Conversely, for $U_2 = \varnothing$, the final annotations were directly obtained without requiring further modification. While this semi-automated approach required some manual checking and correction, the high accuracy of IM significantly reduced the complexity and labor involved in the annotation process.

To annotate the macaque facial image subset V, the annotation tool LabelImg (Tzutalin, 2015) was used to label the ground truth for the macaque faces using RectBox, as shown in Figure 3C. The RectBox coordinates (marked in blue) were transformed into a standard data style consistent with wider face formats (Yang et al., 2016) using formulas (1) and (2), where b_0 and x_{min} are the starting coordinates of the lateral axes, b_1 and y_{min} are the starting coordinates of the vertical axes, x_{max} and y_{max} are the finishing points of the X- and Y-axes, respectively, and b_2 and b_3 are the width and height of the macaque face RectBox.

$$b_0 = x_{min}, b_2 = x_{max} - x_{min} \tag{1}$$

$$b_1 = \mathrm{y}_{min}, b_3 = y_{max} - y_{min} \tag{2}$$

Architecture of ACE-YOLOX

This study developed a lightweight deep learning approach suitable to enable individual identification of macaques from images or videos, specifically tailored for facial recognition. Designing such a network requires a robust capability for detecting subtle defects in facial features while maintaining a compact architecture optimized for deployment on smartphones. The YOLOX detector has been previously verified as a high-performing model that effectively balances accuracy and computational efficiency (Pereira, 2022). Unlike YOLOv4 (Bochkovskiy et al., 2020) and YOLOv5, which face over-optimization challenges, YOLOX uses YOLOv3 (Redmon & Farhadi, 2018) as its foundation, incorporating improvements to achieve better tradeoffs between speed and accuracy in object detection tasks. However, YOLOX exhibits limitations in capturing multiscale information, which is critical for detecting macaque facial features under diverse conditions.

To address these limitations and improve the accuracy,

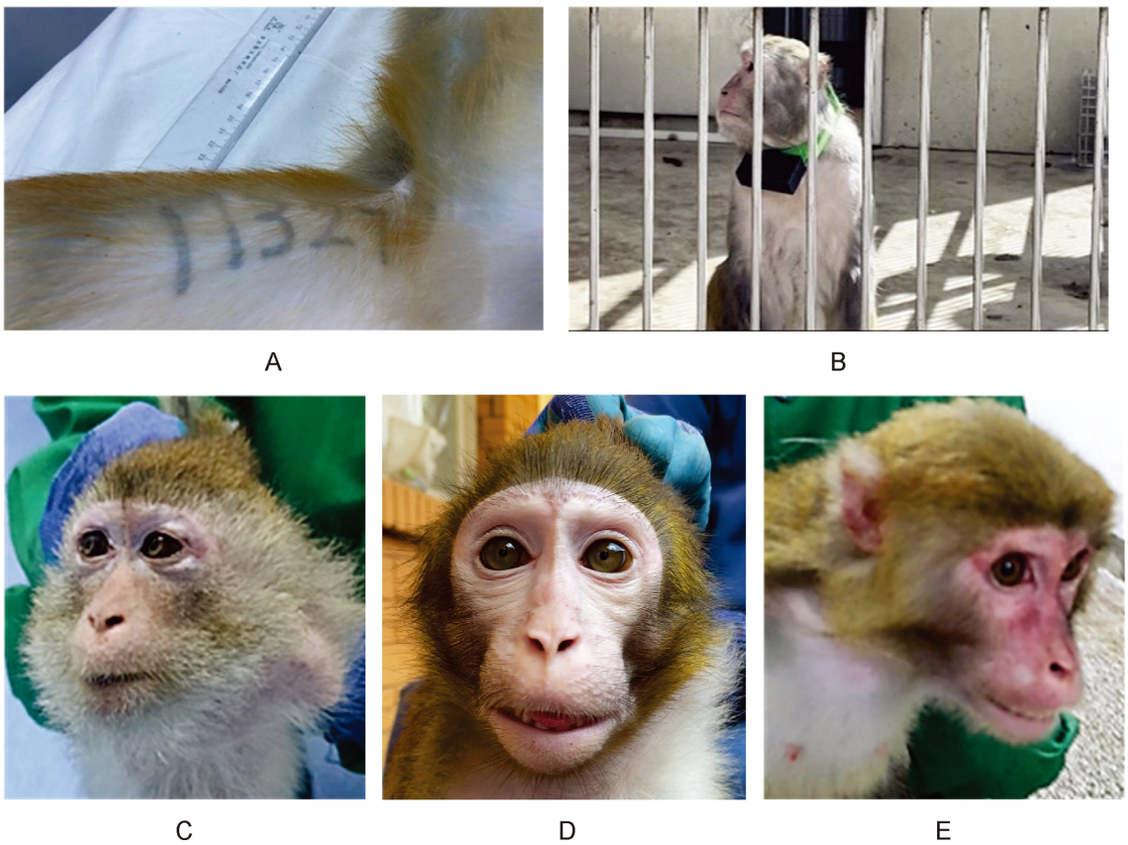

Figure 1 Presentation of experimental data

A: Mark placed on inner thigh of macaque. B: Facial image of a macaque partially obstructed by railings. C–E: Facial images of macaque captured from different angles.

robustness, and real-time performance of macaque facial recognition, an enhanced YOLOX model was developed for automated macaque face detection and identification in challenging environments, including dynamic facial movements, image blurring, and occlusions caused by cage structures. The architecture of ACE-YOLOX integrated several key innovations to enhance performance. At the backbone end of the YOLOX network, an ECA block (Wang et al., 2020) was introduced to enhance its feature-capturing capability. Furthermore, the traditional Intersection over Union (IoU) loss was replaced with CIoU to improve the accuracy of bounding box regression. Additionally, an ASFF block was incorporated after the neck network to strengthen detection capabilities for small facial features. The improved ACE-YOLOX model, named after its three core components —A, C, and E—demonstrates significant advancements in both detection accuracy and robustness as shown in Figure 4.

Efficient channel attention (ECA) block

The ECA block was inserted into the backbone of YOLOX (i.e., dark3, dark4, and dark5) to increase the ability of the network to focus on critical features, enabling the convolutional network to adaptively concentrate on regions of interest (Hu et al., 2018b), suppress unnecessary features (Roy et al., 2023), and improve the representation of meaningful features (Li, 2019). Given the need to identify captive macaques in visually complex environments, it is essential for the model to focus on effective feature channels in macaque facial images.

Therefore, to optimize the model for this task, various channel attention mechanisms were evaluated, including the Squeeze-and-Excitation (SE) block (Hu et al., 2018a), ECA block, and Convolutional Block Attention Module (CBAM) block (Woo et al., 2018). Among these, the ECA block demonstrated superior performance during experiments. It adaptively determined the optimal convolutional kernel size (k) and allocated greater weight to unobstructed macaque faces than to background features or occlusions, such as railings. This approach significantly improved the detection and identification of macaque facial features under challenging conditions.

The ECA block focuses on weighting feature maps and is an extremely lightweight plug-and-play block that can scale up the performance of various CNNs. Dimensionality reduction and the capture of cross-channel interactions can be efficiently avoided. As shown in Figure 5A, after channel-wise Global Average Pooling (GAP) without dimensionality reduction (Gao et al., 2019), the ECA block adaptively selected a kernel size for one-dimensional (1D) convolution, followed by a sigmoid function to capture local cross-channel interactions (Tan et al., 2020). This design incorporated only a few additional parameters and negligible computations, ensuring high efficiency without compromising effectiveness.

Specifically, the ECA block employed a band matrix (W_k) to learn channel attention, involving $k*C$ parameters, where k is the kernel size and C is the number of channels. The interaction between channel y_i and its k nearest neighbors is expressed as:

$$y_i = \sum_{j \in \mathcal{N}(i)} w_{ij} x_j \tag{3}$$

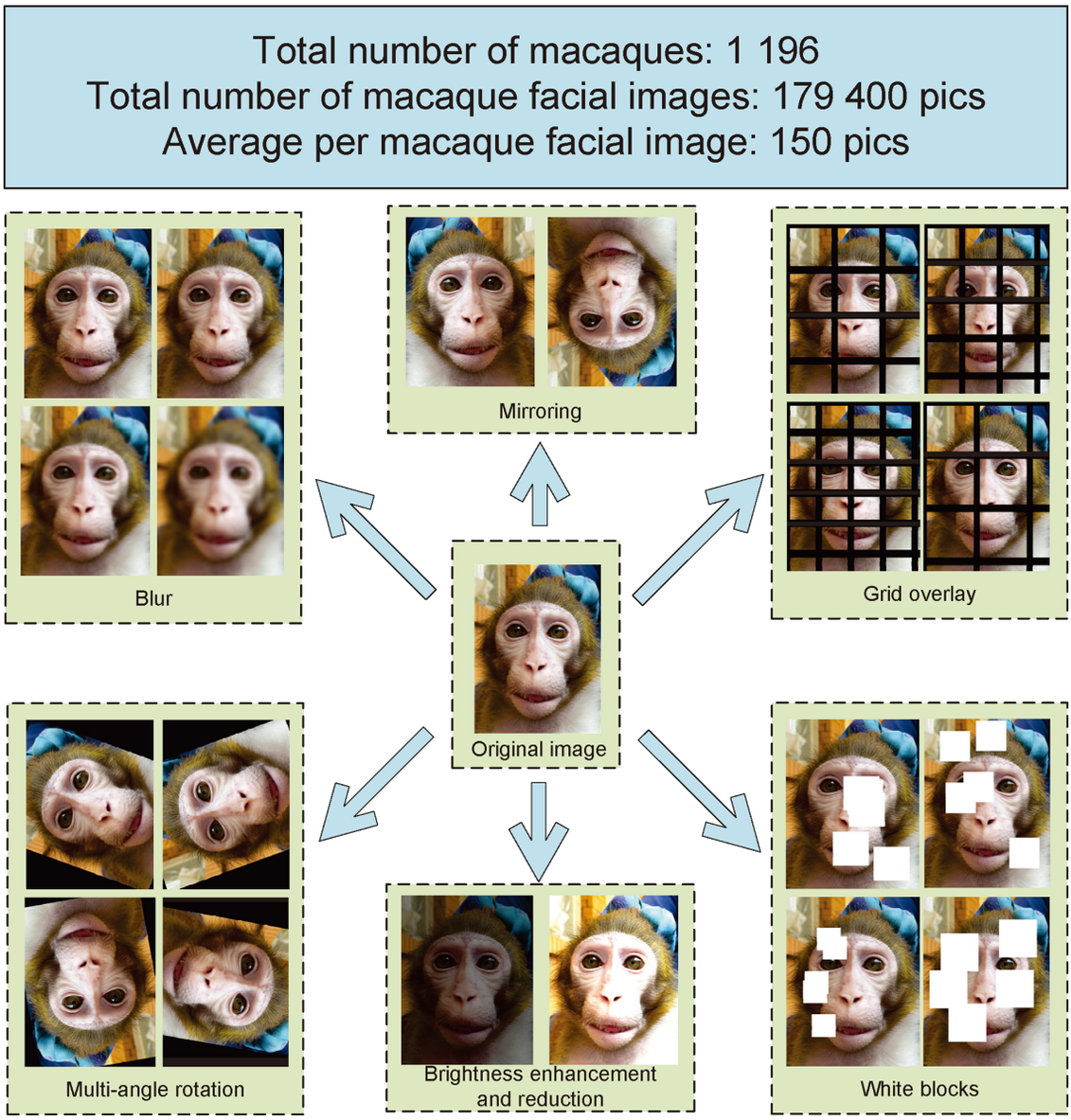

Figure 2 Data augmentation

where $\mathcal{N}(i)$ represents the set of k neighboring channels of y_i and w_{ij} is the learned weight.

The ECA block process was as follows: (1) The input feature map X underwent GAP to produce GAP(X) (Lin et al., 2013). (2) GAP(X) was processed using a 1D convolution to generate weights Conv1D$(\text{GAP}(X))$ (He et al., 2016). (3) The weights Conv1D$(\text{GAP}(X))$ were passed through a sigmoid function to obtain attention weights $\sigma(\text{Conv1D}(\text{GAP}(X)))$. (4) The input feature map $X$was then adjusted by element-wise multiplication with $\sigma(\text{Conv1D}(\text{GAP}(X)))$:

$$Y = X \cdot \sigma\left(\text{Conv}_1^D\left(\text{GAP}(X)\right)\right) \tag{4}$$

Improved path aggregation feature pyramid network (PAFPN) structure

The YOLOX neck network adopted a multilevel feature pyramid (PAFPN) structure, enabling the effective fusion of spatial and semantic information. This structure combined the location details captured by the shallow network with the semantic information of deeper layers, significantly improving both regression and classification performance. By connecting the ASFF module to the PAFPN, feature maps processed by the PAFPN are further refined, allowing the fusion of similar features and enhancing detection accuracy. The structural relationship between PAFPN and ASFF is shown in Figure 5B.

The PAFPN generated three feature maps, labeled from top to bottom as X^1, X^2, and X^3. These color-coded maps represented feature transfer across scales. With the largest receptive field, X^1 is suitable for large object detection, while X^3, with the smallest receptive field, is optimal for small objects, and X^2 is suitable for medium-sized objects.

ASFF fused these three feature maps through two primary processes: Feature resizing and adaptive fusion. Feature resizing scaled each feature map to a uniform size for consistent fusion. For example, $X^{1\to3}$ denotes X^1 resized to the dimensions of X^3. Adaptive fusion involved training weight maps α, β, and γ for each feature map, which were then multiplied element-wise with their corresponding resized maps.

The detailed steps for $ASFF-3$ were as follows: (1) X^1, X^2, and X^3 were resized to the same size, resulting in $X^{1\to3}$, $X^{2\to3}$, and $X^{3\to3}$. (2) The weight maps α, β, and γ were trained to

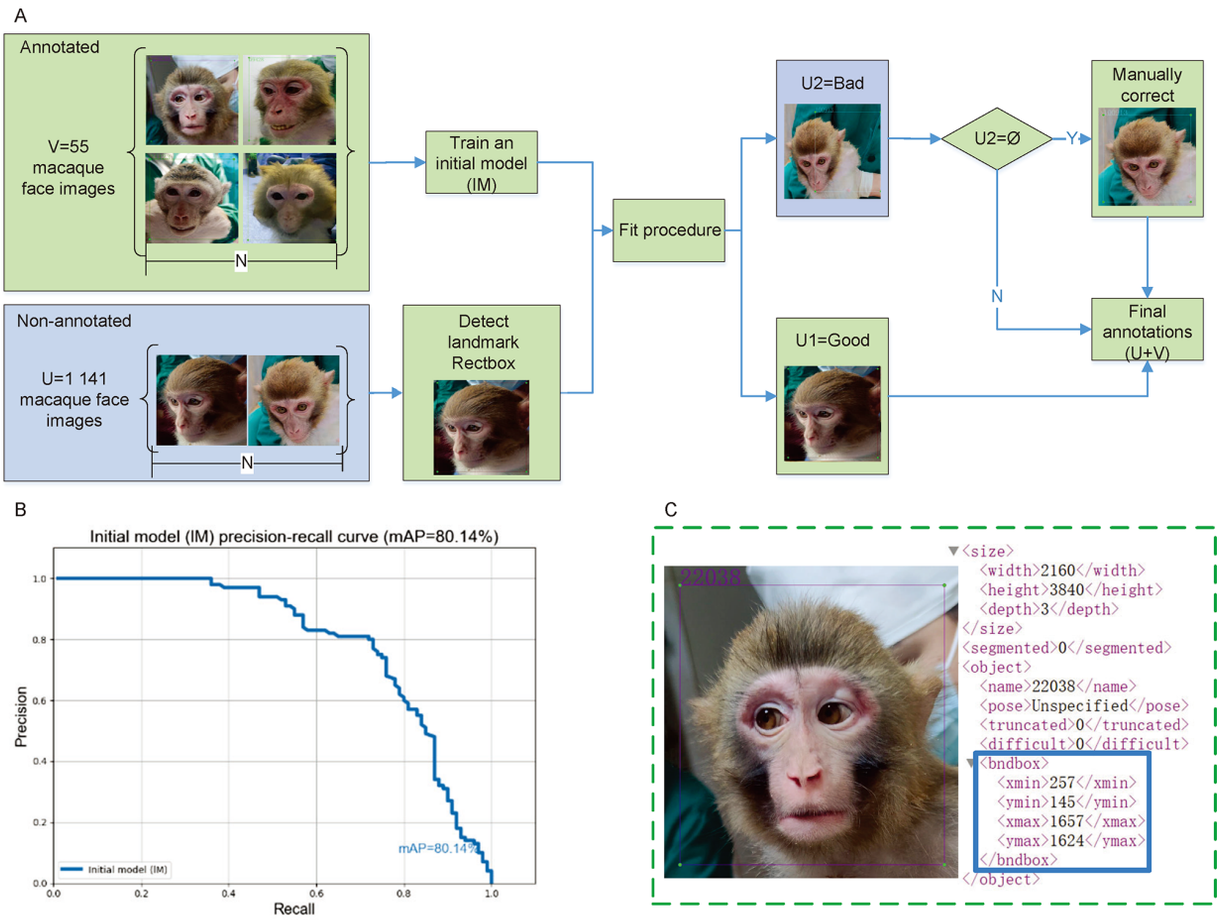

Figure 3 Semi-automated annotation and facial landmark

A: Flowchart of image data annotation processes. B: Precision-recall curve for initial model on annotated subset *V*. C: Example of facial annotation (purple) for macaque labeled "22038". Image shows RectBox coordinates highlighted in blue, which delineate the facial region.

Table 1 Algorithm for semi-automatic data annotation

Input:
Annotated subset V: 55 manually labeled macaque facial images.
Non-annotated subset U: 1 141 macaque facial images.
Output:
Complete annotations for U.
Steps:
1: Train the initial model (IM) using the annotated subset V.
2: Use the IM with the YOLO-landmark detector to detect RectBox for U.
3: Apply a fitting procedure to the detected RectBox in U.
4: Manually review and classify RectBox annotations in U into two categories: "Good" U_1: Accurate annotations. "Bad" $U_2 = U - U_1$: Incorrect or incomplete annotations.
5: if $U_2 \neq \varnothing$, do
6: Apply manual corrections to the RectBox annotations in U_2.
7: if $U_2 = \varnothing$, end if
8: Final annotations for $(U + V)$ are obtained.

match the size of the resized maps. (3) Each weight map was multiplied element-wise using its corresponding feature map. (4) The results of these multiplications were summed to generate a fused feature map. The ASFF-3 formula can be rewritten as:

$$ASFF-3 = X^{1\to3} \otimes \alpha^3 + X^{2\to3} \otimes \beta^3 + X^{3\to3} \otimes \gamma^3 \quad (5)$$

This fusion process enriched X^3 with small-object features while suppressing large- and medium-object features, thus enhancing small object detection. ASFF-1 and ASFF-2 followed a similar principle, focusing on their respective object sizes to improve detection accuracy. The ASFF structure, which is simple and computationally efficient (Yang, 2022), is ideal for applications requiring small object detection in complex backgrounds, such as captive macaque facial recognition, and is well-suited for deployment on mobile devices.

Complete Intersection over Union loss (CIoU)

The YOLOX model employed IoU Loss to measure localization errors. IoU is a straightforward metric for

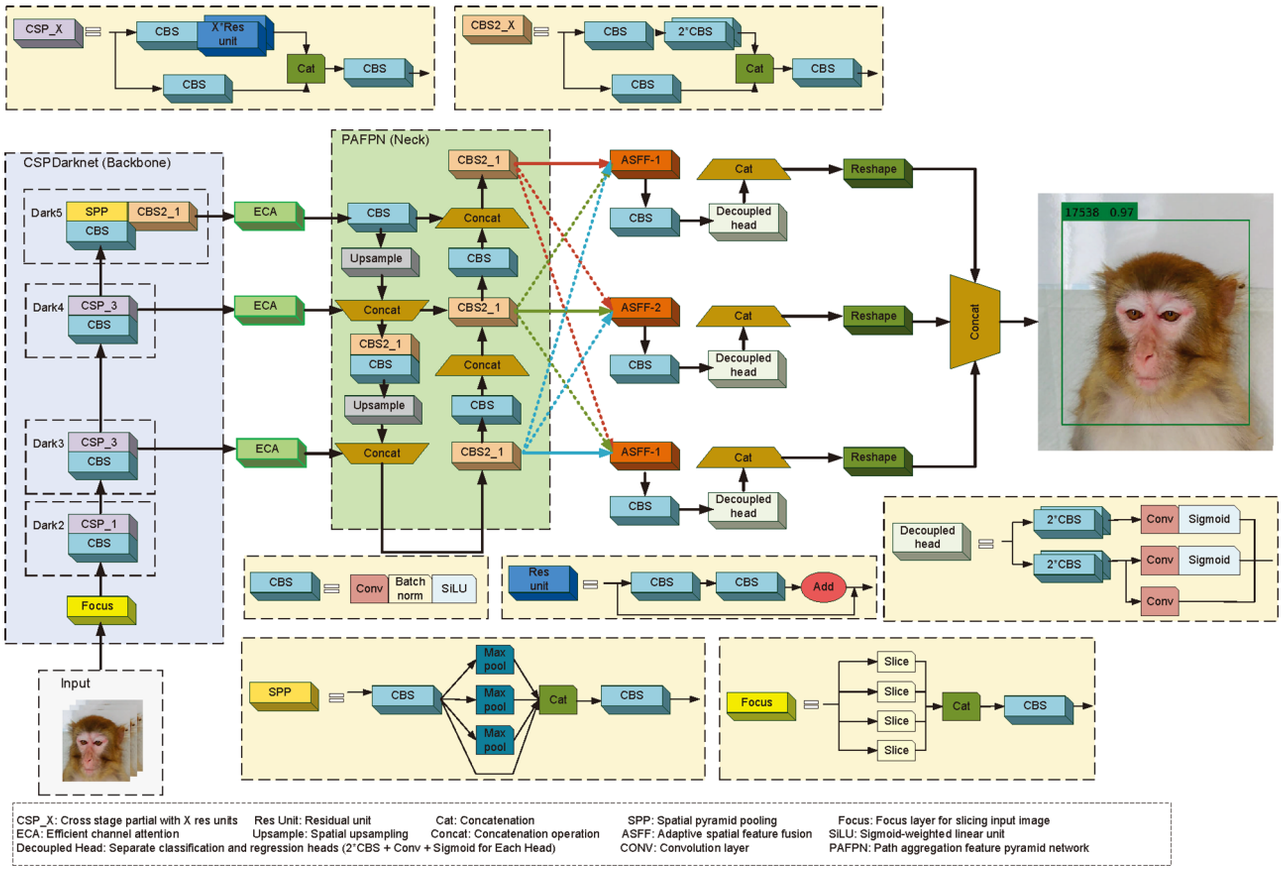

Figure 4 Architecture of ACE-YOLOX model for macaque facial recognition

quantifying the overlap between predicted and ground truth boxes. Here, various localization loss functions were compared, including Generalized Intersection over Union loss (GIoU) (Rezatofighi et al., 2019), Efficient Intersection over Union loss (EIoU) (Zhang et al., 2022), CIoU (Zheng et al., 2020), and Distance Intersection over Union loss (DIoU) (Zheng et al., 2020). CIoU was selected for its enhancement of facial recognition of captive macaques and for addressing the integration of deep and shallow network information. CIoU was designed to overcome the limitations of GIoU by incorporating additional geometric factors, defined as:

$$CIoU = 1 - IoU + \frac{p^2(b_1 + b_2)}{c^2} + \alpha v \tag{6}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{7}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{b_1}}{h^{b_1}} - arctan\frac{w^{b_2}}{h^{b_2}}\right)^2 \tag{8}$$

where b_1 is the center point of the prediction box, b_2 is the center point of the ground truth, and c is the diagonal distance of the smallest box covering the two frames. CIoU incorporates the Euclidean distance between the centers of the predicted and ground truth bounding boxes to enhance localization precision. When the predicted box is entirely within the ground truth box, differences in their positions and center coordinates are accounted for in the CIoU calculation. Parameters w^{b_1}, h^{b_1}, w^{b_2}, and h^{b_2} represent the widths and heights of the prediction and ground truth boxes, respectively. By introducing v, the aspect ratio between the predicted and object boxes is also considered when the centers of the two boxes coincide, leading to a more precise positioning frame and improved detection accuracy. Furthermore, the classification loss in this approach was calculated using cross-entropy (C-E).

RESULTS

Model training

All experiments were conducted under the conditions outlined in Table 2. The facial image dataset of 1 196 macaques was randomly divided into training, validation, and test sets using an 8:1:1 ratio. This partition resulted in 143 520 facial images allocated to the training set, while the validation and test sets each contained 17 940 images. Additionally, an independent test set derived from the living environments of captive macaques was used to evaluate model generalization. This independent dataset was excluded from training to ensure an unbiased assessment of the performance of the model. Furthermore, to ensure fair and consistent performance comparison, all models —including ACE-YOLOX and the classical object detection models (YOLOX, Faster-RCNN, SSD, and CenterNet) —were trained and evaluated on the same macaque facial image dataset.

The performance and behavior of the ACE-YOLOX and YOLOX models are strongly influenced by hyperparameters. In this study, seven key hyperparameters were optimized for both models, including learning rate, batch size, and momentum, among others (Table 3). The hyperparameters were adjusted to balance detection accuracy, training efficiency, and model generalization. This optimization ensured the model performed robustly across various conditions.

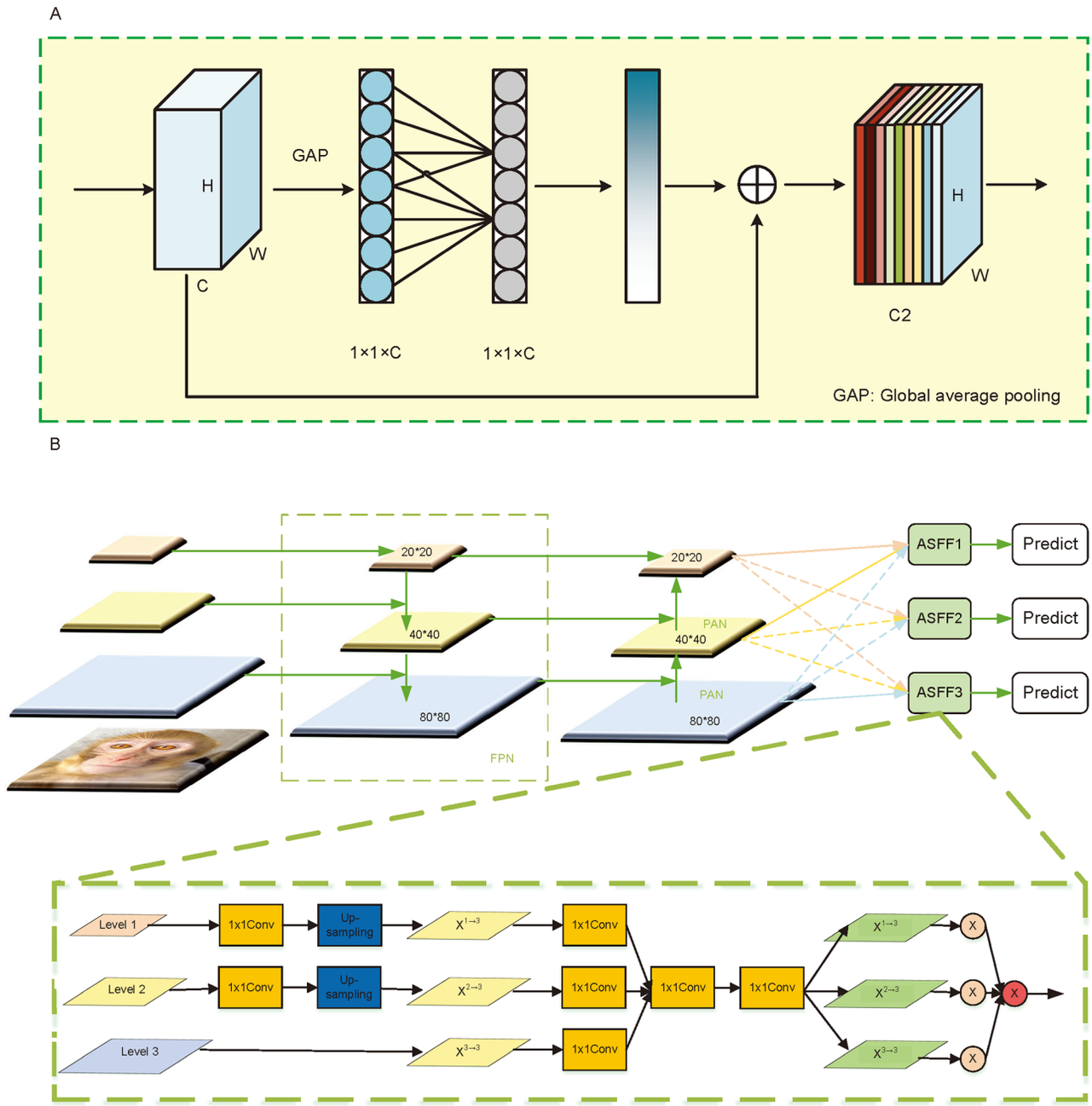

Figure 5 Structural diagram of two blocks inserted into ACE-YOLOX

A: Efficient Channel Attention block. B: Path Aggregation Feature Pyramid Network and Adaptive Spatial Feature Fusion.

Evaluation indicators

The evaluation metrics of the model included precision, recall, mAP, and F1-score, determined using the following equations:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$mAP = \frac{1}{N}\sum_{c=1}^{N} AP_c \tag{11}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{12}$$

where true positive (TP) is the number of positively categorized samples, false positive (FP) is the number of negatively categorized samples, false negative (FN) is the number of positively categorized samples, mAP is the average AP for different categories, and N is the number of categories of the tested samples. In this experiment, macaque ID categories needed to be detected; therefore, N=1 196. The F1-score balanced the effects of precision and recall.

Performance comparison of attentional mechanisms

To evaluate the impact of different attention mechanisms on the YOLOX model, three attention blocks were investigated. The ECA block captures channel-wise dependencies by efficiently recalibrating feature responses. The CBAM integrates both channel and spatial attention mechanisms, enabling the model to adaptively recalibrate feature responses across channels and spatial locations. The SE block emphasizes channel-wise dependencies solely by computing channel attention maps based on global feature statistics,

Table 2 Experimental setup

Configuration	Parameter
Operating System	Windows 10
CPU	Intel Core i9-10900KF
GPU	NVIDIA GeForce RTX3080
RAM	64 GB
Parallel environment	CUDA 11.6
Development environment	PyTorch 1.9.0
Programming language	Python 3.8
Development Framework	Torch 1.10.0, Torchvision 0.11.1

CPU, central processing unit; GB, gigabytes; GPU, graphics processing unit; RAM, random-access memory.

Table 3 Hyperparameter values for ACE-YOLOX and YOLOX model training

Hyperparameters	Value
Initial learning rate	0.01
Batch size	8
Optimizer	SGD
Weight decay	1e-4
Momentum	0.9
Input size	640×640
Epochs	300

SGD, stochastic gradient descent.

thereby enabling the model to selectively emphasize informative channels. These attention mechanisms were added to the end of the backbone network, ensuring that recalibrated features were passed to the PAFPN for further processing. The three methods were trained on the macaque face dataset, with the results listed in Table 4.

Among the three mechanisms, the YOLOX-ECA model demonstrated superior performance, achieving a mAP of 92.21%, recall of 66.26%, F1-score score of 76.5%, and precision of 94.3%. Notably, the mAP of the YOLOX-ECA model was 3.5%, 1.81%, and 3.3% higher than the original YOLOX, YOLOX-SE, and YOLOX-CBAM models, respectively. Furthermore, YOLOX-ECA outperformed the other models in terms of precision, recall, and F1-score, indicating its effectiveness in improving model detection performance.

Performance comparison of loss function

The YOLOX model traditionally employs the IoU loss function, a widely used and effective approach for bounding box regression. However, it lacks the ability to effectively capture target shape perception, limiting its suitability for tasks requiring precise localization. To address this, alternative loss functions, including CIoU, DIoU, EIoU, and GIoU, were evaluated for their impact on detection accuracy. As shown in Table 5, the YOLOX-CIoU model achieved the highest mAP (92.07%), representing a 3.36%, 0.58%, 0.60%, and 1.89% improvement compared to the baseline YOLOX, YOLOX-DIoU, YOLOX-EIoU, and YOLOX-GIoU models, respectively. However, the YOLO-DIoU and YOLOX-EIoU models showed slightly higher precision than the YOLOX-CIoU model. This outcome may be attributed to the heightened emphasis on bounding box alignment and center point distances, which can potentially introduce trade-offs that impact the accuracy of macaque facial recognition.

The CIoU function offers significant advantages, including enhanced training stability, improved optimization, and better target shape perception, translating into superior detection performance and generalization capabilities. Based on these findings, the CIoU function was selected as a replacement for IoU, as it provided consistent improvements in accuracy without the potential limitations observed with the DIoU and EIoU functions.

Table 4 Impact of attention mechanism block on model performance

Method	Precision (%)	Recall (%)	F1-score (%)	mAP (%)
YOLOX	90.70	64.51	74.00	88.71
YOLOX-SE	93.25	61.41	72.83	90.40
YOLOX-CBAM	93.91	63.14	73.50	88.91
YOLOX-ECA	94.30	66.26	76.50	92.21

Table 5 Performance evaluation of four loss functions

Method	Precision (%)	Recall (%)	F1-score (%)	mAP (%)
YOLOX	90.70	64.51	74.00	88.71
YOLOX-CIoU	92.33	75.96	82.83	92.07
YOLOX-DIoU	93.31	71.44	79.50	91.49
YOLOX-EIoU	93.98	68.76	78.33	91.47
YOLOX-GIoU	92.63	67.13	77.20	90.18

Ablation experiment

To validate the effectiveness of network improvement on overall model performance, an ablation experiment was conducted. This analysis examined the contributions of key enhancements, including attention mechanisms, loss function refinements, and feature pyramid optimizations. The results are shown in Table 6 and Figure 6A.

Each improvement step contributed positively to model accuracy. The baseline YOLOX model, without modifications, achieved a precision of 90.70%, recall of 64.51%, F1-score of 74.00%, and mAP of 88.71%. Incorporating ECA, CIoU, and ASFF resulted in significant performance gains, with precision of 95.57%, recall of 94.86%, F1-score of 95.16%, and mAP of 98.88%. These results indicate that each enhancement effectively improved model performance. The combined implementation of all three blocks yielded the best overall results, significantly enhancing the detection capabilities of the network.

Heatmap visualizations (Figure 6B) further illustrated the improvements in attention distribution across macaque facial features for each model variant tested in the ablation study. The standard YOLOX model (top row, first column) exhibited dispersed and inconsistent attention across the face, indicating suboptimal feature detection. Incremental modifications, such as the inclusion of CIoU, ECA, and ASFF, progressively refined the focus on critical facial landmarks, particularly the eyes and nose. For example, integrating the CIoU block alone (top row, second column) enhanced

Table 6 Ablation experiment results

Method	Improvement strategy			Precision (%)	Recall (%)	F1-score (%)	mAP (%)
	ECA	ASFF	CIoU				
YOLOX	–	–	–	90.70	64.51	74.00	88.71
	√	–	–	94.30	66.26	76.50	92.21
	–	√	–	92.52	80.54	85.17	93.76
	–	–	√	92.33	75.96	82.83	92.07
	√	√	–	91.65	78.95	84.17	94.79
	√	–	√	95.60	75.20	83.30	93.70
	–	√	√	93.23	84.28	87.33	95.61
	√	√	√	95.57	94.86	95.16	98.88

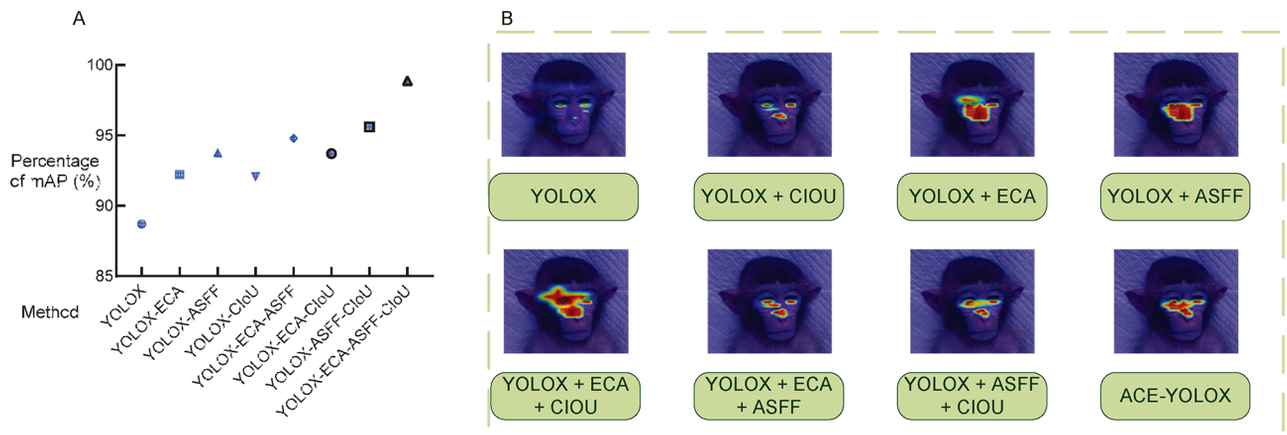

Figure 6 Ablation experiment results

A: mAPs of different models during ablation experiments. B: Heatmap visualization of attention focus of different models in ablation experiments on macaque facial features. Warmer colors (red and yellow) indicate higher attention regions, while cooler colors (purple) indicate lower attention.

attention on central facial features, while ECA and ASFF (top row, third, and fourth columns) further intensified attention on these critical areas. The integration of ECA, CIoU, and ASFF blocks (bottom row) resulted in the most precise attention map, emphasizing high-saliency areas with clear delineation. The final ACE-YOLOX configuration (bottom row, rightmost column) consolidated these enhancements and demonstrated an optimal attention pattern that captured macaque facial structures with the highest accuracy.

The alignment between quantitative metrics and heatmap visualizations reinforces the effectiveness of these enhancements. Each network modification advanced feature localization and attention focus, thereby enhancing the overall precision of the model in macaque facial recognition tasks.

Comparison with other classic object detection models

To assess generalization performance and potential overfitting, the macaque facial dataset was divided into training and validation sets, and loss curves were analyzed across different object detection models. Figure 7 presents the training and validation loss curves for ACE-YOLOX, YOLOX, Faster-RCNN (Ren et al., 2017), SSD (Liu et al., 2016) and CenterNet (Duan et al., 2019) over 300 epochs. ACE-YOLOX exhibited smooth and consistent convergence in both the training and validation sets, indicative of reliable generalizability and an absence of overfitting. The model achieved convergence at approximately 200 epochs, where both training and validation losses stabilized, confirming that the model reliably learned macaque facial features without significant performance degradation on the validation set. In comparison, other models exhibited varied convergence patterns. YOLOX demonstrated a stable convergence pattern but reached a slightly higher validation loss than ACE-YOLOX, suggesting reduced effectiveness in feature extraction for this dataset. Faster-RCNN and SSD displayed higher initial losses and slower convergence, reflecting challenges in capturing the specific features of macaque faces, likely due to their higher parameter counts and complexity. CenterNet, optimized for detecting small objects, failed to fully converge and exhibited notable fluctuations in validation loss, suggesting moderate overfitting and limited generalization capability.

Quantitative comparisons further highlighted the superiority of ACE-YOLOX over other models (Table 7). The model achieved a mAP of 98.88%, recall of 94.86%, and F1-score of 95.16%, significantly outperforming YOLOX, Faster-RCNN, SSD, and CenterNet by 10.17%, 13.79%, 14.02%, and 16.52%, respectively. These improvements reflect the enhanced ability of ACE-YOLOX to detect and localize macaque facial features accurately.

Beyond detection accuracy, ACE-YOLOX maintained a high level of efficiency. With a model size of 37.62 MB, it was slightly larger than the original YOLOX model, but significantly smaller than Faster-RCNN (132.39 MB) and comparable to SSD and CenterNet. As an important indicator of computational complexity, parameter count was relatively low (10.24 million parameters) compared to that of Faster-RCNN (138.36 million parameters) and SSD (26.29 million parameters), indicating a balance between small model size and low parameter count with high accuracy, making it highly suitable for mobile deployment.

Real-time performance was also a key advantage, with ACE-YOLOX achieving 23.67 FPS, confirming its suitability for rapid inference in image and video-based macaque facial

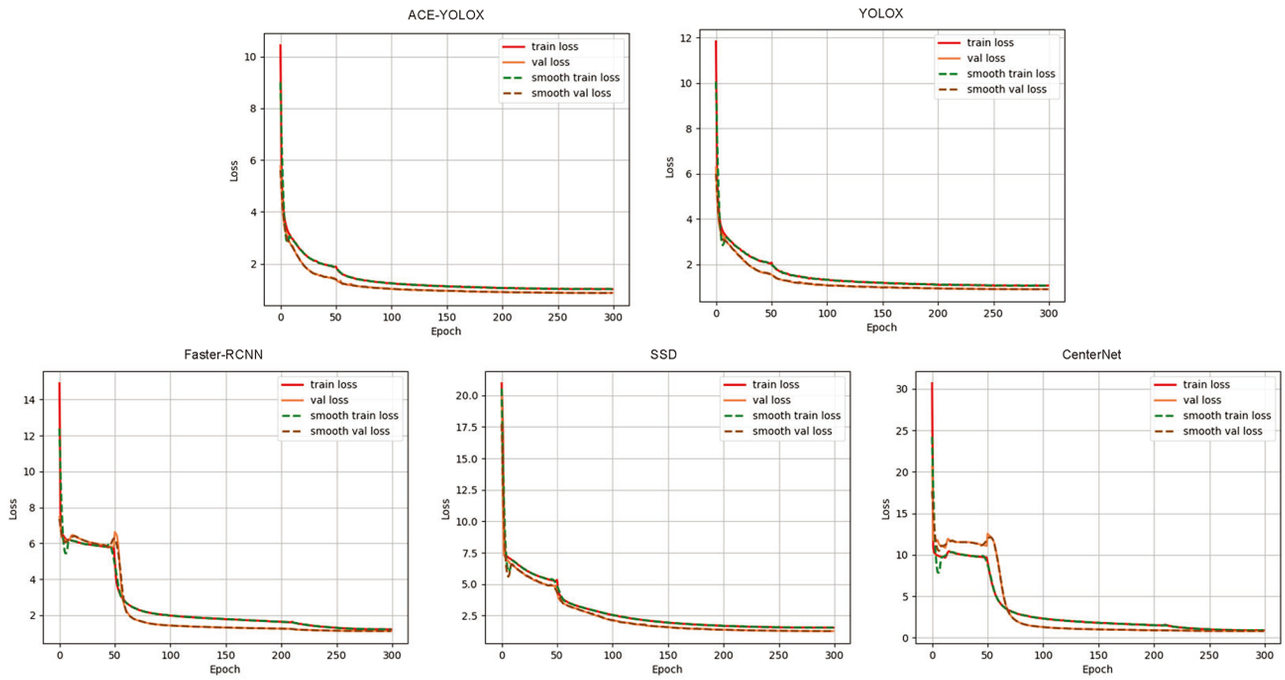

Figure 7 Training and validation loss curve of ACE-YOLOX and classic object detection models

Table 7 Performance evaluation of different object detection models

Method	mAP (%)	Recall (%)	F1-score (%)	Model Size (MB)	Parameter (M)	FPS
YOLOX	88.71	64.51	74.00	34.3	8.97	22.14
Faster-RCNN	85.09	60.18	72.13	132.39	138.35	7.61
SSD	84.86	58.27	69.52	35.78	26.28	16.23
CenterNet	82.63	57.86	65.37	31.16	23.49	13.71
ACE-YOLOX	98.88	94.86	95.16	37.62	10.24	23.67

MB, megabytes; M, million.

recognition. With its lightweight architecture, efficient parameterization, and high detection accuracy, ACE-YOLOX is well-suited for applications requiring fast and reliable identification, particularly on smartphones and other resource-limited platforms.

Evaluation of ACE-YOLOX performance

ACE-YOLOX was evaluated for macaque facial recognition and compared with YOLOX, Faster R-CNN, SSD, and CenterNet using a test dataset from captive macaques. The ability of each model to detect and recognize individual macaques from various angles and in images with multiple individuals was tested, as shown in Figure 8. Bounding boxes indicate detected macaques, with numerical labels representing identification codes and decimal values denoting confidence scores. Results showed that ACE-YOLOX consistently maintained high confidence in identifying individual macaques across varied viewpoints, demonstrating superior detection accuracy and reliability. For instance, ACE-YOLOX correctly identified macaques 06091, 11394, and 06036 with high confidence (0.93, 0.91, and 0.90, respectively), demonstrating superior detection robustness compared to the baseline YOLOX and other classic models. In contrast, Faster R-CNN, SSD, and CenterNet displayed lower confidence values and instances of misdetections or overlapping labels, particularly in non-frontal views. These findings underscore the improved ability of ACE-YOLOX to detect and differentiate individuals across complex scenarios, addressing challenges present in real-world macaque identification tasks.

Figure 9 presents the P-R curve for ACE-YOLOX, corresponding to a mAP of 98.88%. The P-R curve remains close to the upper-left corner, indicating both high precision and high recall across various thresholds. This result demonstrates a strong capability for accurate detection and minimal error rates, even in challenging test conditions. These evaluations establish ACE-YOLOX as an accurate and computationally efficient approach for real-time macaque facial recognition tasks within complex and dynamic settings.

Application of ACE-YOLOX

The ACE-YOLOX system was deployed as a macaque facial recognition tool on a smartphone, enabling real-time detection and identification, as shown in Figure 10A. The Image Acquisition Module enables users to capture a photograph or video of the face of a macaque using the smartphone camera or to select an existing image or video from the gallery. The Image Processing Module preprocesses the selected input, optimizing it for detection tasks. The Facial Detection Module analyzes the preprocessed input, identifying macaque faces and overlaying annotations that include bounding boxes, identification labels, and confidence scores. These results are presented through the Display Module, providing clear visual outputs for recognizing individual macaques. Figure 10B illustrates an example where the application successfully identifies a macaque from an image, annotating the result with identification details and confidence metrics.

Figure 8 Evaluation of recognition models (ACE-YOLOX, YOLOX, Faster-RCNN, SSD, and CenterNet) for multi-macaque detection and facial orientation analysis

DISCUSSION

This study introduces ACE-YOLOX, a lightweight, real-time facial recognition model specifically designed for non-invasive identification of individual captive macaques in group-housed settings without the use of invasive tracking devices. Building upon the YOLOX framework, ACE-YOLOX incorporates ECA and CIoU blocks to enhance accuracy while minimizing computational complexity. Furthermore, the integration of ASFF within the PAFPN architecture strengthens multiscale feature extraction, ensuring a balance between detection efficiency, accuracy, and model size. Experimental evaluations demonstrated that ACE-YOLOX achieved a mAP of 98.88%, a recall of 94.86%, an F1-score of 95.16%, a compact model size of 37.62 MB, and a parameter count of 10.24 million, while maintaining a frame rate of 23.67 FPS. These metrics meet the high accuracy and real-time processing requirements for practical deployment on edge devices such as smartphones, enabling practical and scalable macaque identification in captive environments.

Despite advances in primate facial recognition, challenges persist in adapting models across different settings due to species-specific traits and environmental variables. For example, species-specific models developed for guenons (Allen & Higham, 2015) and gorillas (Loos & Pfitzer, 2012) rely on distinct facial traits that restrict broader applicability (Beery et al., 2019; Norouzzadeh et al., 2018). Similarly, ACE-YOLOX is optimized for recognizing individual macaques within captive breeding facilities, where stable environmental conditions and the presence of familiar individuals facilitate high detection reliability. However, in field conditions or mixed-

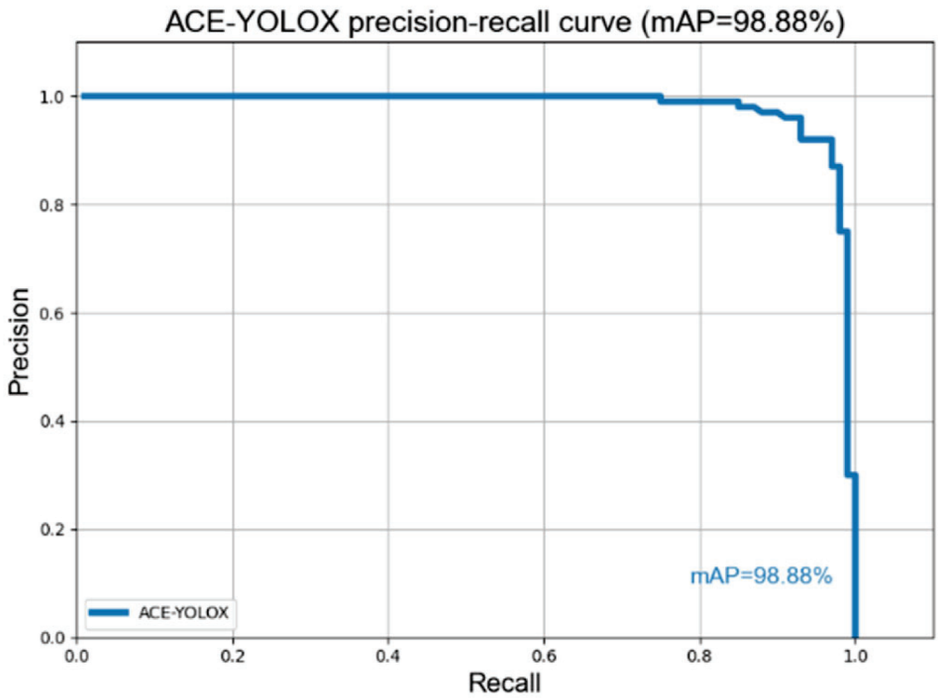

Figure 9 Precision-recall curve of ACE-YOLOX model

species habitats, increased variability in facial features and behavioral patterns introduces significant recognition challenges. Future investigations could focus on developing adaptive recognition frameworks capable of accommodating environmental variability while maintaining high identification accuracy, although the primary focus of research remains the precise and efficient identification of captive macaques.

Another limitation of ACE-YOLOX is its dependence on a predefined group of individuals for training (Sun et al., 2017). When group compositions change, retraining is necessary to integrate newly introduced macaques, limiting flexibility in dynamic settings (Goodfellow, 2016). Techniques from human facial recognition, such as few-shot learning (Koch et al., 2015), offer a potential solution for rapid incorporation of new individuals with minimal additional data. Implementing such techniques would enhance model flexibility by reducing the need for repeated retraining, allowing seamless adaptation to evolving group compositions. Environmental factors also present challenges for macaque recognition (Branson et al., 2009). In captive settings, visual obstructions from enclosure fencing can interfere with recognition accuracy. While the present approach does not require background removal, future implementations, particularly those incorporating dynamic data collected from outside enclosures, could benefit from background subtraction techniques similar to those used in MonkeyTrail (Liu et al., 2022). These methods would enhance macaque feature isolation, improving detection accuracy in visually complex environments. Future work could explore these methods to improve adaptability while reducing the need for frequent retraining, addressing constraints related to species-specificity and static group identification.

ACE-YOLOX demonstrated effectiveness in real-time recognition of individual captive macaques, making it particularly effective for monitoring groups in controlled environments. Expanding its application in large-scale breeding facilities or more complex research contexts may require additional safeguards to enhance reliability. Integrating a proofreading module into the mobile recognition system could enable manual validation of automated classifications, reducing the risk of misidentification (Brabham, 2008). This feature would be especially valuable in scenarios with heightened classification challenges, potentially reducing the need for frequent retraining and enhancing overall accuracy (Estellés-Arolas et al., 2012).

From an ethical and epistemic standpoint, the integration of deep learning into animal behavior studies introduces unique considerations. While automated recognition optimizes efficiency and scalability (Mittelstadt et al., 2016), reliance on machine-driven recognition may inadvertently limit opportunities for serendipitous discoveries and weaken the depth of observational insight gained through direct human engagement (Crawford, 2021). These trade-offs highlight a broader challenge in balancing innovation with the intangible yet critical aspects of traditional ethological approaches. Although ACE-YOLOX eliminates the need for invasive tracking devices, aligning with ethical standards in animal welfare (Gönen & Alpaydın, 2011), its application should be complemented by strategies to address these limitations. Incorporating participatory observations alongside automated recognition or integrating interpretable models could help mitigate the epistemic and ethical costs while maintaining the benefits of automation (Paulet et al., 2024).

In conclusion, ACE-YOLOX offers a robust and efficient solution for real-time identification of individual captive macaques, with smartphone-based deployment enabling efficient management of group-housed primates in research facilities and zoological institutions. By improving identification accuracy and operational efficiency, this approach enhances animal welfare while reducing the labor demands associated with traditional monitoring methods. Beyond individual recognition, the model establishes a foundation for future applications in facial expression recognition (Liu et al., 2023), cognitive psychology, and social behavior in macaques. More broadly, this research highlights the potential applicability of artificial intelligence in wildlife conservation, ecological

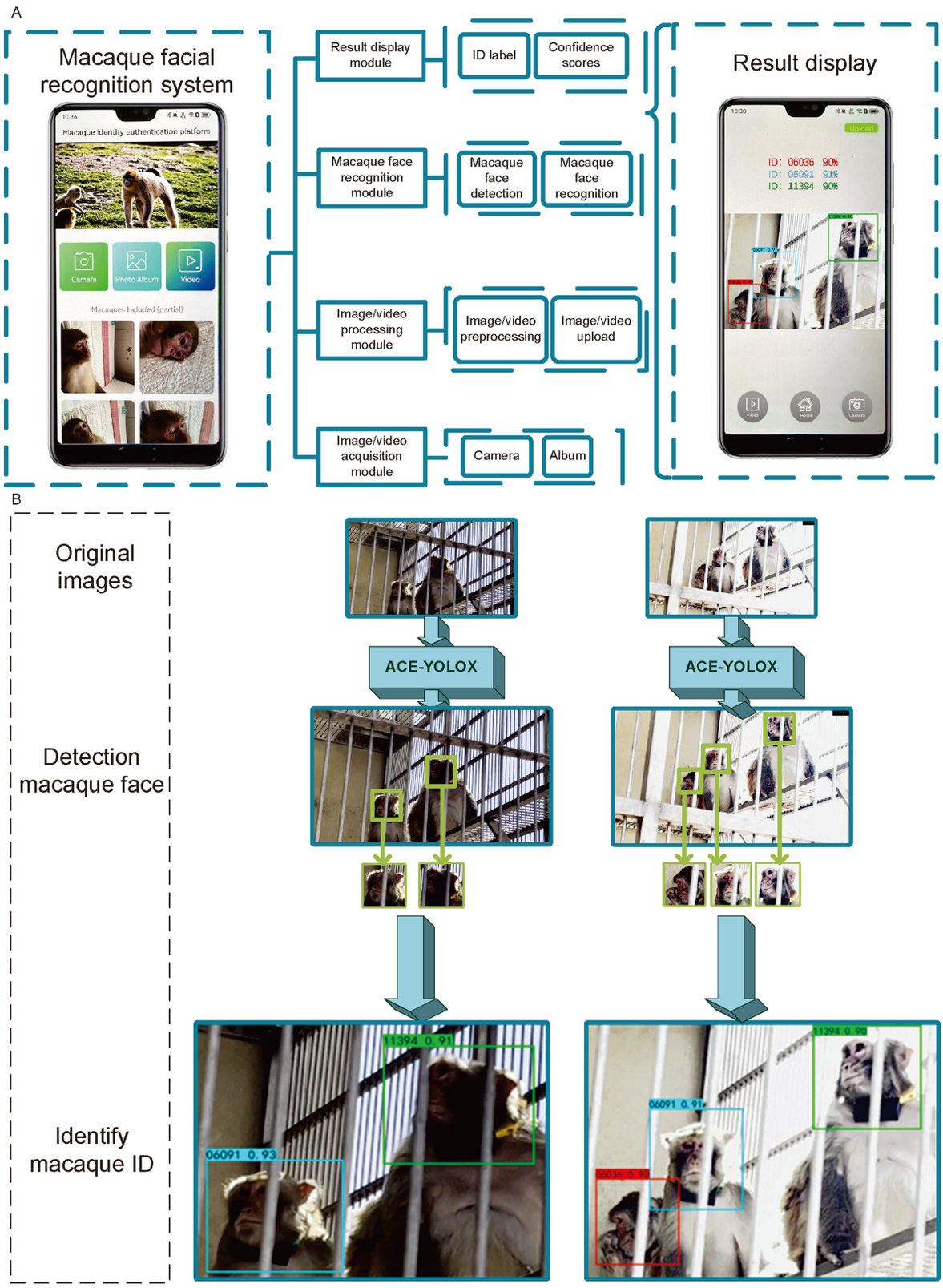

Figure 10 Structure and workflow of macaque facial recognition application on a smartphone

A: Structure of macaque facial recognition smartphone application. B: Detection and identification of individual macaques on a smartphone.

monitoring, and species management, demonstrating the transformative role of deep learning in advancing non-invasive methodologies for behavioral and ecological studies.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

D.D.W. and J.J.Z. conceptualized the project, designed the experiments, and supervised the project. Y.G. and B.L.Z. collected facial photos of macaques and analyzed the data. J.J.Z. and Y.G. developed and optimized the algorithm. D.D.W., J.J.Z., and Y.G. wrote and edited the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors sincerely thank Prof. Bin Su, Prof. Long-Bao Lv, Dr. Jian-Hong Wang, Dr. Xue-Rui Zeng, Dr. Xiao-Mei Yu, and Dr. Ying-Zhou Hu from the Kunming Institute of Zoology for their invaluable assistance in conducting this study. We also extend our gratitude to the staff members of the National Research Facility for Phenotypic & Genetic Analysis of Model Animals (Primate Facility) (https://cstr.cn/31137.02.NPRC) for their technical support and assistance with data collection and analysis.

REFERENCES

Ait-Saidi A, Caja G, Salama AAK, et al. 2014. Implementing electronic identification for performance recording in sheep: I. Manual versus semiautomatic and automatic recording systems in dairy and meat farms. *Journal of Dairy Science*, **97**(12): 7505–7514.

Allen WL, Higham JP. 2015. Assessing the potential information content of multicomponent visual signals: a machine learning approach. *Proceedings of the Royal Society B: Biological Sciences*, **282**(1802): 20142284.

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

Beery S, Morris D, Yang SY. 2019. Efficient pipeline for camera trap image review. arXiv preprint arXiv: 1907.06772.

Billah M, Wang XH, Yu JT, et al. 2022. Real-time goat face recognition using convolutional neural network. *Computers and Electronics in Agriculture*, **194**: 106730.

Bochkovskiy A, Wang CY, Liao HYM. 2020. YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934.

Brabham DC. 2008. Crowdsourcing as a model for problem solving: an introduction and cases. *Convergence: The International Journal of Research into New Media Technologies*, **14**(1): 75–90.

Branson K, Robie AA, Bender J, et al. 2009. High-throughput ethomics in large groups of *Drosophila*. *Nature Methods*, **6**(6): 451–457.

Cai C, Li JQ. 2013. Cattle face recognition using local binary pattern descriptor. *In*: Proceedings of 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. *Kaohsiung*, **China**: IEEE,1–4.

Chen P, Swarup P, Matkowski WM, et al. 2020. A study on giant panda recognition based on images of a large proportion of captive pandas. *Ecology and Evolution*, **10**(7): 3561–3573.

Corrêa MDS, Catai AM, Milan-Mattos JC, et al. 2019. Cardiovascular autonomic modulation and baroreflex control in the second trimester of pregnancy: a cross sectional study. *PLoS One*, **14**(5): e0216063.

Crawford K. 2021. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Haven: Yale University Press.

Crouse D, Jacobs RL, Richardson Z, et al. 2017. LemurFaceID: a face recognition system to facilitate individual identification of lemurs. *BMC Zoology*, **2**(1): 2.

Duan KW, Bai S, Xie LX, et al. 2019. CenterNet: keypoint triplets for object detection. *In*: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 6568–6577

Ernst A, Küblbeck C. 2011. Fast face detection and species classification of African great apes. *In*: Proceedings of 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Klagenfurt: IEEE, 279–284.

Estellés Arolas E, González-Ladrón-De-Guevara F. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science*, **38**(2): 189–200.

Fernandez-Duque M, Chapman CA, Glander KE, et al. 2018. Darting primates: steps toward procedural and reporting standards. *International Journal of Primatology*, **39**(6): 1009–1016.

Floyd RE. 2015. RFID in animal-tracking applications. *IEEE Potentials*, **34**(5): 32–33.

Freytag A, Rodner E, Simon M, et al. 2016. Chimpanzee faces in the wild: log-euclidean CNNs for predicting identities and attributes of primates. *In*: Proceedings of 38th German Conference on Pattern Recognition. Hannover: Springer, 51–63.

Gao SH, Cheng MM, Zhao K, et al. 2021. Res2Net: a new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(2): 652–662.

Gao ZL, Xie JT, Wang QL, et al. 2019. Global second-order pooling convolutional networks. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 3019–3028.

Ge Z, Liu ST, Wang F, et al. 2021. YOLOX: exceeding yolo series in 2021. arXiv preprint arXiv: 2107.08430.

Gokcekus S, Firth JA, Regan C, et al. 2021. Recognising the key role of individual recognition in social networks. *Trends in Ecology & Evolution*, **36**(11): 1024–1035.

Gönen M, Alpaydın E. 2011. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, **12**: 2211–2268.

Goodfellow I. 2016. Deep Learning. Cambridge: MIT Press.

Guo ST, Xu PF, Miao QG, et al. 2020. Automatic identification of individual primates with deep learning techniques. *iScience*, **23**(8): 101412.

Harding JD. 2017. Nonhuman primates and translational research: progress, opportunities, and challenges. *ILAR Journal*, **58**(2): 141–150.

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

Hemsworth PH, Mellor DJ, Cronin GM, et al. 2015. Scientific assessment of animal welfare. *New Zealand Veterinary Journal*, **63**(1): 24–30.

Hu J, Shen L, Sun G. 2018a. Squeeze-and-excitation networks. *In*: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 7132–7141

Hu J, Shen L, Albanie S, et al. 2018b. Gather-excite: exploiting feature context in convolutional neural networks. *In*: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc. , 9423–9433.

Koch G, Zemel R, Salakhutdinov R. 2015. Siamese neural networks for one-shot image recognition. *In*: Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR, 1–30.

Li HY. 2019. Channel locality block: a variant of squeeze-and-excitation. arXiv preprint arXiv: 1901.01493.

Lin M, Chen Q, Yan SC. 2013. Network in network. arXiv preprint arXiv: 1312.4400.

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

Liu ST, Huang D, Wang YH. 2019. Learning spatial fusion for single-shot object detection. arXiv preprint arXiv: 1911.09516.

Liu W, Anguelov D, Erhan D, et al. 2016. SSD: single shot MultiBox detector. *In*: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 21–37.

Liu XH, Gan L, Zhang ZT, et al. 2023. Probing the processing of facial expressions in monkeys via time perception and eye tracking. *Zoological Research*, **44**(5): 882–893.

Loos A, Pfitzer M. 2012. Towards automated visual identification of primates using face recognition. *In*: Proceedings of 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP). Vienna: IEEE, 425–428.

Marsot M, Mei JQ, Shan XC, et al. 2020. An adaptive pig face recognition approach using Convolutional Neural Networks. *Computers and Electronics in Agriculture*, **173**: 105386.

Mittelstadt BD, Allo P, Taddeo M, et al. 2016. The ethics of algorithms:

mapping the debate. *Big Data & Society*, **3**(2): 2053951716679679.

Norouzzadeh MS, Nguyen A, Kosmala M, et al. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(25): E5716–E5725.

Paulet J, Molina A, Beltzung B, et al. 2024. Deep learning for automatic facial detection and recognition in Japanese macaques: illuminating social networks. *Primates*, **65**(4): 265–279.

Pereira N. 2022. PereiraASLNet: ASL letter recognition with YOLOX taking mean average precision and inference time considerations. *In*: Proceedings of 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP). Vijayawada: IEEE, 1–6

Redmon J, Divvala S, Girshick R, et al. 2016. You only look once: unified, real-time object detection. *In*: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 779–788.

Redmon J, Farhadi A. 2018. YOLOv3: an incremental improvement. arXiv preprint arXiv: 1804.02767.

Ren SQ, He KM, Girshick R, et al. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(6): 1137–1149.

Rezatofighi H, Tsoi N, Gwak J, et al. 2019. Generalized intersection over union: a metric and a loss for bounding box regression. *In*: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 658–666.

Rose C, De Heer RC, Korte S, et al. 2012. Quantified tracking and monitoring of diazepam treated socially housed cynomolgus monkeys. *Regulatory Toxicology and Pharmacology*, **62**(2): 292–301.

Roy AM, Bhaduri J, Kumar T, et al. 2023. WilDect-YOLO: an efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecological Informatics*, **75**: 101919.

Sarker IH. 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, **2**(6): 420.

Schofield D, Nagrani A, Zisserman A, et al. 2019. Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, **5**(9): eaaw0736.

Schofield DP, Albery GF, Firth JA, et al. 2023. Automated face recognition using deep neural networks produces robust primate social networks and sociality measures. *Methods in Ecology and Evolution*, **14**(8): 1937–1951.

Sheehan MJ, Straub MA, Tibbetts EA. 2014. How does individual recognition evolve? Comparing responses to identity information in *Polistes* species with and without individual recognition. *Ethology*, **120**(2): 169–179.

Song S, Liu TH, Wang H, et al. 2022. Using pruning-based YOLOv3 deep learning algorithm for accurate detection of sheep face. *Animals*, **12**(11): 1465.

Sun C, Shrivastava A, Singh S, et al. 2017. Revisiting unreasonable effectiveness of data in deep learning era. *In*: Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 843–852.

Tan MX, Pang RM, Le QV. 2020. EfficientDet: scalable and efficient object detection. *In*: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 10778–10787.

Tibbetts EA, Dale J. 2007. Individual recognition: it is good to be different. *Trends in Ecology & Evolution*, **22**(10): 529–537.

Tzutalin. 2015. LabelImg. Git code. https://github.com/tzutalin/labelImg.

Ueno M, Kabata R, Hayashi H, et al. 2022. Automatic individual recognition of Japanese macaques (*Macaca fuscata*) from sequential images. *Ethology*, **128**(5): 461–470.

Wang QL, Wu BG, Zhu PF, et al. 2020. ECA-Net: efficient channel attention for deep convolutional neural networks. *In*: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 11531–11539.

Wang Z, Bovik AC, Sheikh HR, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13**(4): 600–612.

Witham CL. 2018. Automated face recognition of rhesus macaques. *Journal of Neuroscience Methods*, **300**: 157–165.

Woo S, Park J, Lee JY, et al. 2018. CBAM: convolutional block attention module. *In*: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 3–19.

Xu BB, Wang WS, Guo LF, et al. 2021. Evaluation of deep learning for automatic multi-view face detection in cattle. *Agriculture*, **11**(11): 1062.

Yang L, Yuan GW, Zhou H, et al. 2022. RS-YOLOX: a high-precision detector for object detection in satellite remote sensing images. *Applied Sciences*, **12**(17): 8707.

Yang S, Luo P, Loy CC, et al. 2016. WIDER FACE: a face detection benchmark. *In*: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 5525–5533.

Yao YG. 2022. Towards the peak: the 10-year journey of the National Research Facility for Phenotypic and Genetic Analysis of Model Animals (Primate Facility) and a call for international collaboration in non-human primate research. *Zoological Research*, **43**(2): 237–240.

Zhang YF, Ren WQ, Zhang Z, et al. 2022. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing*, **506**: 146–157.

Zheng ZH, Wang P, Liu W, et al. 2020. Distance-IoU loss: faster and better learning for bounding box regression. *In*: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 12993–13000.