Received: 18 June 2025 | Accepted: 15 October 2025

DOI: 10.1111/2041-210X.70187

Methods in Ecology and Evolution

RESEARCH ARTICLE

Vision transformers for age prediction from facial images in a wild primate

Julien P. Renoult[1] | Romain Karpinski[2] | Loïc Sauvadet[3] | Mélodie Kreyer[3] | Richard Mbadoumou[3] | Berta Roura-Torres[4] | Alice Baniel[4] | Marie J.E. Charpentier[4,5]

[1]CEFE, University Montpellier, CNRS, EPHE, IRD, Montpellier, France

[2]CNRS, Inria, LORIA, Nancy, France

[3]Projet Mandrillus, Fondation Lekedi pour la Biodiversité, Bakoumba, Gabon

[4]Institut des Sciences de l'Evolution de Montpellier (ISEM), UMR5554-University of Montpellier/CNRS/IRD/EPHE, Montpellier, France

[5]Department for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Konstanz, Germany

Correspondence
Julien P. Renoult
Email: julien.renoult@umontpellier.fr

Funding information
GENCI-IDRIS, Grant/Award Number: 2023-AD011014579; SEEG Lekedi and SEE-LIFE initiative (INEE-CNRS); Leakey Foundation, Grant/Award Number: S202210309; Max Planck Society; Agence Nationale de la Recherche, Grant/Award Number: ANR-20-CE02-0005-01

Handling Editor: Russell Dinnage

Abstract

1. Accurate estimation of individual ages is crucial for studies in ecology, behaviour and conservation. However, when birth dates are unknown, estimating chronological ages often relies on post-mortem morphological analyses or invasive and cumbersome techniques. Here we investigate the potential of deep learning applied to photographic portraits for non-invasive chronological age prediction.
2. Comparing the predictive capabilities of several recent deep learning models with 25,500 portraits of wild mandrills collected on 284 individuals of known ages in situ, we show that the foundational transformer models DINOv2 largely outperformed convolutional networks (notably ResNext, ConvNeXt, EfficientNetv2) and the other popular transformer model VOLO.
3. To gain insight into the model's predictions, we first examine the influence of the background. Although the model relied on background information for its predictions, this did not lead to a significant improvement in overall accuracy: there was no meaningful difference between predictions when age estimates were from images with or without background. Second, we show that inter-individual variation in prediction errors is partly explained by biological factors. At the individual scale, the prediction error was consistent through time: when individuals appeared older than their chronological age when young, they also consistently appeared older throughout their life. In addition, we found that offspring of older mothers appeared older compared to those of younger mothers, consistent with previous findings on the link between offspring development and maternal age in this species.
4. Altogether, these results indicate that the most modern artificial intelligence methods offer a simple, low-cost and non-invasive approach for chronological age estimation and that the difference between chronological and estimated ages could be used by behavioural ecologists to study individual growth, pace of development and biological aging processes.

Julien P. Renoult and Romain Karpinski contributed equally to this work.

KEYWORDS
artificial intelligence, behavioural ecology, chronological age, DCNN, deep learning, DINOv2, Mandrillus sphinx, prediction errors

1 | INTRODUCTION

Predicting the chronological age of wild animals plays a pivotal role in behavioural ecology and conservation science. Virtually all biological traits are affected by age including growth, survival and reproductive success, making it an essential metric when studying individual fitness, population demography and ecological dynamics (Chaloupka & Musick, 2017; Jarman et al., 2015). In wild animals, age predictions help estimate population growth rates and survival probabilities, which inform management strategies such as setting sustainable catch quotas in fisheries or conducting population viability analyses in conservation efforts (Beissinger & Westphal, 1998; Rughetti, 2016; Sutherland & Norris, 2002). However, assessing directly the chronological age of wild animals remains highly challenging. Traditional approaches often face limitations when applied to free-ranging animals because they are often unreliable, require invasive sampling, which can be impractical or ethically problematic in the wild, or necessitate longitudinal approaches that demand substantial time and resources.

Historically, the main method to predict chronological age has relied on morphological markers, such as fish otoliths and dental cementum in mammals (Campana & Thorrold, 2001; Wittwer-Backofen et al., 2004). However, these morphological methods are not universal and generally require capturing animals or collecting invasive samples, which are all problematic for long-term studies on elusive or endangered species. More recently, molecular biomarkers have offered alternative approaches to estimate age. DNA methylation, the addition of methyl groups to DNA molecules, for example, has been shown to correlate with age (Horvath, 2013). However, methylation clocks often require specific calibration, as methylation dynamics vary significantly across taxa (Bewick et al., 2017). Additionally, the high cost of methylation sequencing methods, data processing and analytical tools constitutes a substantial barrier to their use in ecological studies. Another molecular age biomarker is telomeres—repetitive DNA sequences at the end of chromosomes—which gradually shorten with age in many species (Vaiserman & Krasnienkov, 2021). The rate of telomere attrition is, however, highly sensitive to genetic factors, stress, environmental exposures and life-history traits, which limits its use to predict chronological age (Vaiserman & Krasnienkov, 2021). One major drawback of using molecular age biomarkers is their reliance on high-quality DNA, which often requires invasive sampling of fresh tissue. This poses significant challenges for long-term studies in field conditions, where non-invasive or minimally invasive sampling methods are preferred to alleviate stress on organisms and ensure repeated measurements over time. Methods that minimize animal disturbance, are scalable and can be universally applied across taxa are thus needed for predicting chronological age in wild animals.

One such promising approach is the use of photographic data to predict chronological age. Photographic imaging is inexpensive and does not, or little, disturb animals. In humans, facial image analysis for age prediction has been an active area of research for decades, advancing from manual feature extraction to automated, artificial intelligence (AI)-driven systems (Angulu et al., 2018). Recent strides in AI have revolutionized facial age estimation by leveraging vast amounts of image data to capture complex aging patterns (Bobrov et al., 2018; Meng et al., 2024). For example, a convolutional neural network (CNN) trained with images of the corner of the eye—the facial region most sensitive to aging in humans—was able to predict chronological age with a mean absolute error (MAE) of 2.3 years for a sample of individuals aged 20 to 80 years (Bobrov et al., 2018). On the same sample, the authors reported a MAE of 2.7 years for an estimation based on DNA methylation. In wild animals, to our knowledge a single study has used an AI-based method to predict age from images (Zang et al., 2022). A CNN trained on photographic portraits of pandas was able to predict age with a MAE of 2.4 years for individuals aged between 0 and 38 years.

Beyond CNNs, Vision Transformers (ViTs) have been recently used to estimate age from photographs. ViTs represent a groundbreaking shift in computer vision by applying transformer architecture—originally developed for text analysis—to visual data (Dosovitskiy et al., 2020). Unlike traditional CNNs that operate hierarchically on image pixels to capture spatial features, ViTs break images into smaller patches, process each patch as a sequence (similar to words in a sentence), and model relationships between these patches to capture complex patterns and contextual dependencies (Raghu et al., 2021). This allows ViTs to learn more intricate, global representations of visual information, which are especially beneficial for nuanced tasks where capturing subtle variations in texture, shape and environmental context is essential. This explains why the state-of-the-art in age prediction from human face images is currently achieved with a ViT model, VOLO-D1, with a MAE of 4.2 years on a test dataset including a homogeneous age distribution between 0 and 70 years (Kuprashevich & Tolstykh, 2023).

The objective of this study was twofold. First, we aimed to evaluate the performance of state-of-the-art AI methods in predicting chronological age from photographic portraits in a wild primate: the mandrill (*Mandrillus sphinx*). Portrait images have been collected on 284 individuals of known exact ages for more than 10 years, within the framework of a long-term field research project in Gabon. We compared the performance of different AI architectures, including CNNs and ViTs, with a particular focus on DINOv2 models (Oquab et al., 2023). DINOv2 is a family of ViT models recently developed by META AI and is often presented as the first foundation model, the equivalent of ChatGPT for images. DINOv2 has recently shown very high performance in a diversity

of tasks, but has seldom been applied in ecology and evolutionary biology (e.g. see Maddigan et al., 2024).

Second, we investigated how the model made its predictions, and analysed how factors related to both the data and the biology of the studied mandrills influenced the model's predictive accuracy. Leveraging techniques of Explainable AI, we investigated the contribution of face and background pixels in an image to age prediction. This allowed us to determine if the model exhibited shortcut learning, which occurs when an AI model learns a simple, non-robust rule to solve a task instead of the complex, intended logic. A classic example of shortcut learning is a model trained to distinguish between wolves and huskies, which learned to associate snow in the background with 'wolf' because most wolf photos in the training set were taken in the snow (Ribeiro et al., 2016). While exploiting background information might not necessarily be an issue for estimating the age of wild mandrills, it would limit the model's generalization and performance in other settings, such as predicting the age of primates in a homogeneous environment like a laboratory. Last, we studied whether biological factors contribute to explain variation in prediction errors. If so, we predicted that intra-individual age predictions across pictures should show less variation compared to inter-individual predictions for individuals of the same age. We also predicted that individuals who appear older than their chronological age at a given time should continue to do so for a certain period. Last, we predicted that age errors should convey information regarding early or late development in the studied mandrills. In captive mandrills, infants born to high-ranking and older mothers are heavier (i.e. higher weight) but not taller (based on crown-rump length measures) than those born to low-ranking or younger mothers (Setchell et al., 2001). Here, using portraits collected on infants, we analysed the relationship between the relative error in predicting infants' chronological age and their mother's age and rank.

2 | MATERIALS AND METHODS

2.1 | Data

Mandrill portraits were retrieved from the Mandrillus Face Database (MFD), created and managed by the Mandrillus Project which studies, since 2012, the socio-ecology of the only natural population of mandrills habituated to human presence (project approved by an authorization from the CENAREST Institute, Gabon; permit number: AR017/22/MESRSTTCA//). MFD includes photographic portraits collected between January 2012 and December 2022 on 410 mandrills of all ages and both sexes who are all individually recognized (Tieo et al., 2023). Pictures were taken directly in the forest by field assistants. As distance to the camera, illumination, pose and facial expression varied between pictures due to field conditions, pictures of MFD are thus qualified as 'non-standardized'. The quality of each portrait was manually scored between 0 (worst images) and 3 (best images; see Tieo et al., 2023 for details). In this study, we discarded images with quality score 0, as well as images in profile view (FaceView = 0 in MFD). We further excluded individuals with an estimated error in their date of birth that exceeded 14 days. The final dataset includes 25.5 K square portrait images taken on 284 individuals (129 males, 155 females; Figure 1; Figure S1). On average, each individual was represented by 90 images (±72.5 SD; range: 1–533). The date of birth was known with certainty (accuracy <1 day) for 231 individuals (81%; Figure S2). For the remaining 53 individuals, their date of birth was estimated, with an error rate ≤14 days, based on observational data on their mother's reproductive cycle and growth patterns. The age at the time of shooting ('chronological age' hereafter) ranged from 0 to 16.4 years (mean: 2.9 ± 2.6 SD) for males, and from 0 to 23.6 years (mean: 5.5 ± 3.9 SD) for females (Figure S1).

2.2 | Models

For a conservative use of computational resources, we proceeded to identify the best model through successive and nested steps, rather than through an extensive grid search. We first identified the best architecture, then tested the effect of age normalization with this best architecture, then the effect of the age range during training with the best architecture and normalization scheme, and finally, the influence of the training task on the best configuration obtained previously. All experiments were performed on a V100 GPU supercomputer using the PyTorch library.

2.2.1 | Architectures

The CNN architectures tested included the classic VGG16 model (Simonyan & Zisserman, 2014) and more recent models: ResNeXt (Xie et al., 2017), EfficientNetv2 (Tan & Le, 2021) and ConvNeXt (Liu et al., 2022). The vision transformers included three variants of VOLO (D1, D3 and D5; Yuan et al., 2022) and three variants of DINOv2 (SMALL, MEDIUM and LARGE; Oquab et al., 2023), with variants differing in depth and the number of parameters (see Table 1). We performed transfer learning, that is, we fine-tuned models pre-trained on either a dedicated dataset (for DINOv2, see Oquab et al., 2023) or on the ImageNet 1 k (Russakovsky et al., 2015) dataset (other models). All models and weights were retrieved from the PyTorch Image Models repository (Wightman, 2019), except VGG16 which was retrieved directly from PyTorch. The 'head' consisted of a single-neuron, regression layer grafted to the backbone architecture. In preliminary experiments, we tested the effect of adding up to four dense layers (the first of size 512, 256, 128 or 64, followed by 0 to 3 additional dense layers), between the backbone architecture and the regression neuron. For none of the models did adding dense layers improve the accuracy of predictions (results not shown).

2.2.2 | Age range in the training set

We tested the effect of varying the age range during training on the accuracy of predictions. For example, we investigated whether the

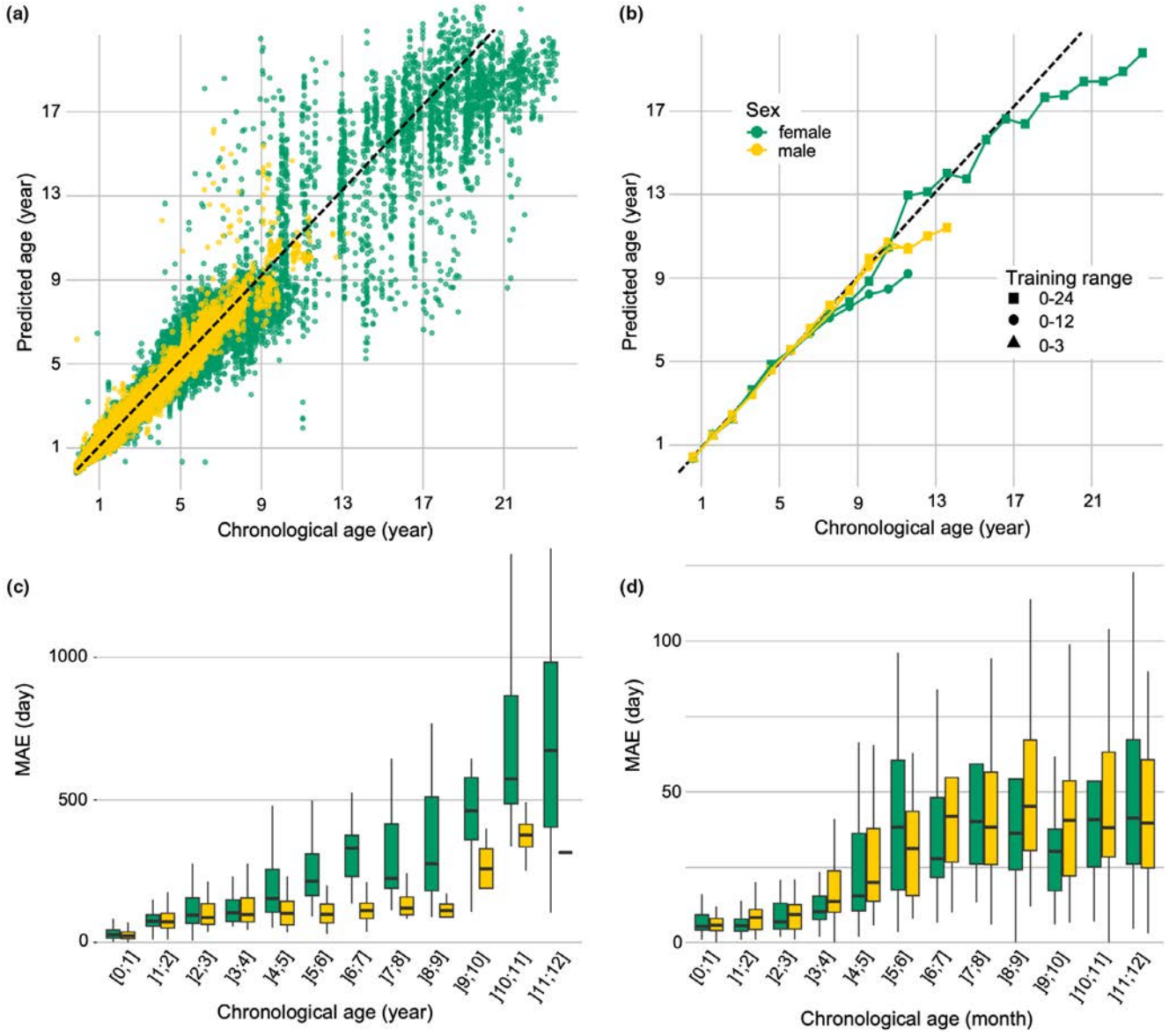

FIGURE 1 Accuracy in age prediction. Model: DINOv2-Fine tuned-LARGE. Females in green, males in yellow. (a) Relation between predicted and chronological ages, all images displayed. Training range: 0–24 years. (b) Comparison of the relation between predicted and chronological ages across training age ranges. Data pooled by 1-year intervals of chronological age. (c) Distribution of accuracy (MAE) by 1-year intervals for individuals aged 0–12 years. MAE calculated on validation sets (fivefold cross-validation: Each image had one predicted value). For a given age interval, MAE was averaged across images of a given individual. Age range for training set: 0–12 years. (d) same as (c) but: Age range 0–1 years and MAE pooled by 1-month intervals. Age range for training set: 0–3 years.

model performed better in predicting the age of young individuals (0–3 years) when the training dataset only included this age range, or when it also included older individuals (e.g. 0–12 years or 0–24 years).

2.2.3 | Training task

In humans, four deep learning-based approaches have been used to predict the age from portrait images: classification (e.g. Lapuschkin et al., 2017; Nada et al., 2020), regression (e.g. Dornaika et al., 2020; Shen et al., 2018), ranking (i.e. predicting the age as ranks; e.g. Shin et al., 2022) or distribution learning (age is predicted as an imprecise tag; e.g. Wen et al., 2020). In this study, we chose to formulate the age prediction task as a regression problem, which provides several key advantages over alternatives. First, the deep regression predicts continuous values, allowing the model to estimate any age within a range. This aligns naturally with the way age functions as a continuous variable. In contrast, other methods require discretizing age as ranks or into distinct categories or bins, which lead to a loss of information. Contrary to regression models, in particular, classification models cannot approximate age ranges from unseen classes (Kuprashevich & Tolstykh, 2023). Second, regression accounts for the inherent order of ages. The model understands that age 10 is younger than

TABLE 1 Comparison of accuracy in age prediction across architectures.

Architecture	Rank	MAE	CS@*l* 1/12	CS@*l* 1	CS@*l* 3
VGG16	6	252.0	20.6	82.6	93.7
ResNeXt	5	245.5	19.4	79.9	**94.9**
EfficientNetv2	8	273.0	18.3	77.9	92.9
ConvNeXt	10	309.3	19.6	77.2	92.3
VOLO-D1	4	234.8	20.3	**82.7**	93.8
VOLO-D3	2	224.8	22.5	82.4	93.8
VOLO-D5	9	292.2	16.9	78.5	92.5
DINOv2-Frozen- MEDIUM	11	323.7	18.5	75.1	91.7
DINOv2-Fine tuned-SMALL	7	255.5	22.3	81.2	93.2
DINOv2-Fine tuned-MEDIUM	3	229.6	23.7	82.4	94.2
DINOv2-Fine tuned-LARGE	1	**213.4**	**27.5**	82.4	92.7

Note: Accuracy was calculated on a single validation set (the same for all trainings) corresponding to ⅕ of the entire image dataset, with image quality 1–3 and all ages included (0–24 years). For each metric, the highest accuracy (in days) is in bold. MAE: mean absolute error. CS@*l*: cumulative score at threshold *l* years. Rank calculated from MAE.

age 20 and older than age 5. Standard classification does not account for the order between classes, treating them as independent. Third, regression eliminates the need to define age bins, which can be arbitrary and may not reflect natural groupings in the data. Fourth, regression uses loss functions that directly measure the difference between predicted and chronological ages, leading to more accurate optimization. In contrast, categorical loss functions (like the cross-entropy loss) may not penalize the model appropriately for near-miss predictions (e.g. predicting age 10 when the actual age is 11). For these reasons, and following the most recent studies in this field (e.g. Qin et al., 2023; Zha et al., 2024), we treated the age prediction task as a regression problem and used the mean absolute error (MAE) as a loss function.

In addition, we tested whether age prediction was improved when coupling regression with other tasks. We compared the following tasks performed in parallel to the regression task: sex classification (cross-entropy loss, two categories), image quality classification (cross-entropy loss, three categories), sex and image quality classifications and age verification. The age verification task consists of taking another image of the same individual taken the same day and training the model to predict that the age difference between the two images is 0.

2.3 | Training

Age was predicted in years as we found no improvement when normalizing age within the range [0–1] (either using sigmoid or max normalization; Tables S1 and S2). Portrait images were resized to 224 × 224 pixels using bilinear interpolation, and their RedGreenBlue channels z-normalized. The dataset was then split into fivefolds, ensuring that multiple images of the same individual were all in the same fold to avoid overestimation of accuracy. For the architecture comparison step, we trained models on fourfolds and evaluated performance on the remaining fold only once. For other experiments, we performed fivefold cross-validation and present results averaged over the five validation folds. The training datasets were augmented using various image transformations (see Table S3). All training experiments lasted 30 epochs, which was sufficient for all models to reach convergence. For each training run, we selected the model with the lowest validation loss. The batch size was set to 256. Weights were optimized using the AdamW optimizer. For each architecture, the best learning rate (fixed across all epochs) was searched for over a short training period of five epochs. All weights (both of the backbone and the head) were optimized except in one training scheme for which only the weights of the head were optimized (DINOv2-Frozen- MEDIUM).

2.4 | Evaluation

Model accuracy was evaluated using two types of metrics: the Mean Absolute Error (MAE) and the Cumulative Scores (CS). MAE calculates the absolute difference between the predicted and the actual age in the validation set. CS describes the percentage of images whose absolute error is less than or equal to a tolerance level *l*. We present results using different tolerance levels varying from 1 month to 3 years. For example, CS@1/12 = 50 means that 50% of predictions have an error (MAE) of less than 1 month (1/12 year). All accuracies are presented after converting the age in days.

In the discussion we will compare the performance of our best model with the state-of-the-art in age prediction from photographic portraits in humans. To do so, we need to convert mandrill age into human age. We fitted a regression model linking mandrill age to human age, based on a comparison of developmental periods

between the two primate species: infancy (weaning), childhood (first molar eruption), juvenescence (growth spurt: takeoff velocity), adolescence (growth spurt: peak velocity) and adulthood. For each period, key age milestones were retrieved from Bogin and Smith (1996) for humans, and from Setchell et al. (2001) and Wickings and Dixson (1992) for mandrills.

2.5 | Influence of the background

We used the Integrated Gradients method (Sundararajan et al., 2017) as implemented in the Captum v0.8.0 library to visualize and quantify the contribution of individual pixels to age predictions. Integrated Gradients attributes importance to each input feature by integrating gradients along a path from a baseline input to the actual input, producing an attribution score per pixel that reflects its relative contribution to the model's prediction. We calculated attribution scores for pixels in the face and background regions separately. To do so, we trained a model to segment the images by generating a mask of the face (including the fur). First, a subset of 429 images was manually annotated to create a training dataset. This dataset was then used to fine-tune the Segment Anything Model v2 (Ravi et al., 2024), which was subsequently employed to semi-automatically annotate all images of one validation fold ($n=4522$ images). The fine-tuned encoder of SAM2 was leveraged as a feature extractor, with a newly added decoder component to generate the final segmentation masks. We compared attribution scores for the face and background areas for the first fold of the cross-validation procedure. Then, we compared predictions when training the best-performing model on the original portraits or those with a masked background (i.e. background pixels set to 0).

2.6 | Error variability analysis

We tested the effect of sex and image quality on accuracy using a generalized linear mixed effect model, with a Gamma distribution family and a log link function, with MAE being considered as the response variable and sex (two-level categorical variable), image quality (three-level categorical variable) and individual chronological age (continuous z-transformed variable) as three fixed effects. Mandrill's identity was considered as a random effect. This model was fitted with three different datasets with different ranges of chronological ages: 0–3 years, 0–12 years and 0–24 years.

We analysed the inter-individual variation in age prediction accuracy using alternatively the standard deviation (SD) and the coefficient of variation (CV) of MAE calculated for a given age interval (e.g. 1-month or 1-semester intervals, see Results' section). MAE was averaged across images of a given individual, for a given age interval, and SD and CV were then computed across individuals. We fitted a generalized linear model, using a Gamma distribution family and a log link function, considering either SD or CV as a response variable and sex, chronological age, the interaction between these two variables and the number of pictures (total number of images included in the age interval) as fixed effects. Only individuals with at least three pictures (see justification in Figure S3), and only age intervals with at least three individuals were included in these analyses. The same models were fitted on different datasets using different ranges of chronological ages (0–3 years, 0–12 years or 0–24 years) and age intervals (1-month, 1-trimester or 1-year intervals depending on the age range).

2.7 | Predicted versus chronological age throughout life

We examined whether an individual whose predicted age was higher than their chronological age at a given age tended to consistently display higher predicted than chronological ages when aging. Such a result is expected if, for example, variation in error is partly due to variation in growth rate or developmental pace across individuals. Because the developmental precocity or delay due to prenatal and early maternal effects could quickly fade during an individual's development, we studied individuals aged between 0 and 1 year separately from those older than 2 years.

For the 0–1 year-old mandrills (infants thereafter), portraits were pooled in 1-month age classes. We randomly selected two photos of the same individual taken at a time interval Δt, with $\Delta t \in \{1, 2, 4, 6, 8\}$ months. For $\Delta t = 4$, for example, the selected pairs of photos could depict a given individual aged 4 and 8 months or an individual aged 2 and 6 months. We randomly drew 1000 pairs of photos from all possible pairs, each time randomly selecting first an individual, then a photo of that individual in each of the two age classes separated by Δt. From these 1000 pairs, we calculated the probability that an individual whose predicted age was higher than their chronological age at time t still had a predicted age higher than their chronological age at time $t + \Delta t$. We tested the significance of this probability using a permutation test: for each Δt, the probability was compared to a null distribution (one-tailed test) calculated by repeating the above procedure 1000 times but randomly selecting two individuals at each iteration to form the photo pairs separated by Δt.

For individuals aged 2 years and older, we performed similar analyses but with a different handling of the time intervals. Here, photos were pooled in 1-year age classes; t corresponded to portraits of individuals aged between 2 and 3 years, and Δt corresponded to {1, 2, 4, 6} years. With $\Delta t = 4$, for example, the possible pairs of photos concerned individuals aged between 2 and 3 years and photos of the same individuals aged 6 and 7 years old. The difference between the tests with 0–1 year olds and those aged 2 years and above was due to the constraint of having a minimum of three different individuals and more than three photos per individual for each of the two age classes separated by Δt.

2.8 | Maternal age influence

We tested the hypothesis that the error in predicting an infant's chronological age can be explained by the mother's age at birth.

We analysed the relative (not absolute) error (model trained with individuals aged 0–3 years) of 4367 portraits representing 167 infants aged 0–1 year, with a known mother. Low-quality images (face_qual = 0 or 1) were excluded. Predictions were pooled for each individual over 3-day intervals, and intervals with fewer than three images per individual were discarded (results were qualitatively similar when predictions were pooled over 7-day or 15-day intervals). This resulted in a dataset of 705 pooled predictions (167 infants, mean number of images per prediction: 6.1 ± 5.0 SD; range: 3–55). Maternal age at birth was retrieved from long-term demographic records available on the study population (as per: Charpentier et al., 2020). Maternal rank at birth was obtained using outcomes of approach–avoidance behaviours collected during ad libitum and focal observations. Normalized David's scores were first computed annually to quantify the social dominance for all females aged 4 years and older. Ranks were then calculated from the proportion of other females dominated by each mother, with continuous values ranging from 0 (lowest-ranking mother, who dominated none of the others) to 1 (highest-ranking mother). We fitted a linear mixed-effects model to explain the relative prediction error using maternal rank, maternal age at birth, sex of the infant and standardized chronological age as fixed effects. Infant identity was included as a random effect, and a variance structure (varExp) was applied to account for heteroscedasticity increasing with age.

All statistical analyses were performed on R (version 4.3.0) using the functions glm (stats package, v.3.6.2), lmer and glmer (lme4 package, v.1.1.33), for generalized, linear mixed-effects and generalized mixed-effects models, respectively.

3 | RESULTS

3.1 | Model optimization

3.1.1 | Comparison between architectures

The comparison of architectures was performed using the entire age range (0–24 years). The lowest Mean Absolute Error (MAE) of 213 days was achieved with the DINOv2 transformer, specifically with the LARGE variant which contained the highest number of layers and parameters (Table 1, Table S4). The second-best architecture was VOLO-D3 but the precision improvement brought by DINOv2-LARGE was substantial: an average of 11.4 days. In general, transformers outperformed CNNs: VGG, ResNeXt, EfficientNetv2 and ConvNeXt ranked lower, being only surpassed by the smallest variant of DINOv2 (SMALL) and by DINOv2-Frozen-MEDIUM for which the regression layer alone was fine-tuned. It is notable that the simple VGG16 architecture was more accurate for this task than other newer CNNs (EfficientNetv2 and ConvNeXt). The cumulative scores confirmed this ranking when the age threshold value *l* was less than 1 year. For $l > 1$, other models seemed to be competitive. However, when datasets (for both training and evaluation) were restricted to good quality images (quality score 2 and 3), DINOv2-LARGE outperformed all alternatives for all considered metrics (Table S5). Only DINOv2-LARGE was thus used in subsequent analyses.

3.1.2 | Age range for training

As expected, the mean accuracy was higher (i.e. MAE lower) when predicting the age of infants and juveniles (age range of the validation set: 0–3 years, second column in Table 2) compared to larger age ranges (0–12 or 0–24 years, Figure 1a). Restricting training to the age range corresponding to the one we sought to predict yielded more accurate predictions than when the training age range also included older individuals. For example, to predict ages between 0 and 12 years, it was preferable to include only individuals aged 0 to 12 years in the training set (MAE = 135.9) rather than individuals aged 0 to 24 years (MAE = 165.1; Table 2, Tables S6–S9).

Training task and image quality.–Multi-tasking did not increase accuracy in age prediction. On the contrary, adding another task to age prediction led to an increase in MAE of at least 7.9 days ('age' vs. 'age + sex' in Table S10). Across all datasets, image quality influenced accuracy, with the highest MAE for the lowest quality (quality score: 1) and the lowest MAE for the highest quality (quality score: 3; Table S11).

3.1.3 | Conversion in human age

Based on a comparison of the timing of entry into the five major developmental periods in humans and mandrills (infancy, childhood, juvenescence, adolescence and adulthood; Table S12), we found that mandrill age = 0.46 × human age − 0.94 for males and mandrill age = 0.37 × human age − 0.86 for females (Figure S4). Thus, 5 years in humans is equivalent to 1.36 and 0.99 years for a male and female

TABLE 2 Effect of age range in the training set.

Train (left) & validation (right) max age		MAE	CS@/ 1/12	CS@/ 1	CS@/ 3
24	24	**218.2 ± 36.4**	**25.6 ± 2.1**	**83.2 ± 4.5**	**96.2 ± 1.3**
24	12	165.1 ± 25.2	25.2 ± 2.4	89 ± 3.5	98.9 ± 0.5
12	12	135.9 ± 17.2	**30.8 ± 2.4**	**90.8 ± 2.6**	**99.1 ± 0.7**
24	3	70.4 ± 7.4	39.3 ± 3	99.2 ± 0.4	**100 ± 0**
12	3	55.9 ± 6.8	49.2 ± 2.8	99.5 ± 0.4	**100 ± 0**
3	3	**46.4 ± 6.1**	**53.4 ± 3.1**	**99.7 ± 0.1**	**100 ± 0**

Note: The age range is given by the first two columns. For example, Train = 24 and Validation = 12 means that the model has been trained using portraits of individuals aged 0–24 years, and validated portraits of individuals aged 0–12 years. Accuracy was calculated as the mean (± standard deviation) of the five validation sets. The highest accuracy is in bold. Image quality: 1–3. Model: DINOv2-Fine tuned-LARGE. MAE: mean absolute error, in days. CS@*l*: cumulative score at threshold*l*/years.

mandrill, respectively. Therefore, CS@5 in humans should be compared to CS@1 in mandrills (see Section 4).

3.2 | Influence of the background

A visual inspection of the attribution score heatmaps reveals the occurrence of occasionally high scores located in the background (Figure 2a). To quantify this effect, we created a background segmentation mask (Figure 2b). On average, the background occupied 11% of the total image area. Among images with at least 1% background (76%), the ratio of attribution scores for the face (including the fur) and the background was, on average, 0.96. This indicates that the age prediction was primarily based on the face. However, a correlation exists between this ratio and the ratio of the face area to the background area: the more space the background occupies in the image, the more it is taken into account for determining the mandrills' age (Figure 2c). Nevertheless, due to its generally very small area, the background did not influence the model's overall performances. The MAE of age predictions with the background (132.61 days) and without the background (132.23 days) was not significantly different (Wilcoxon signed rank test with continuity correction: $V = 3.6e6$, $p = 0.1827$). There was no effect of sex on the difference between predictions with and without the background (Figure 2d, Table S13). We found a significant but very small effect of age: the error difference increases by 2.00 days for every one standard deviation increase in chronological age, which corresponds to 920.7 days (Figure 2d, Table S13). Overall, even if the background was used by the model to make predictions, we can confidently conclude that it did not bias model predictions.

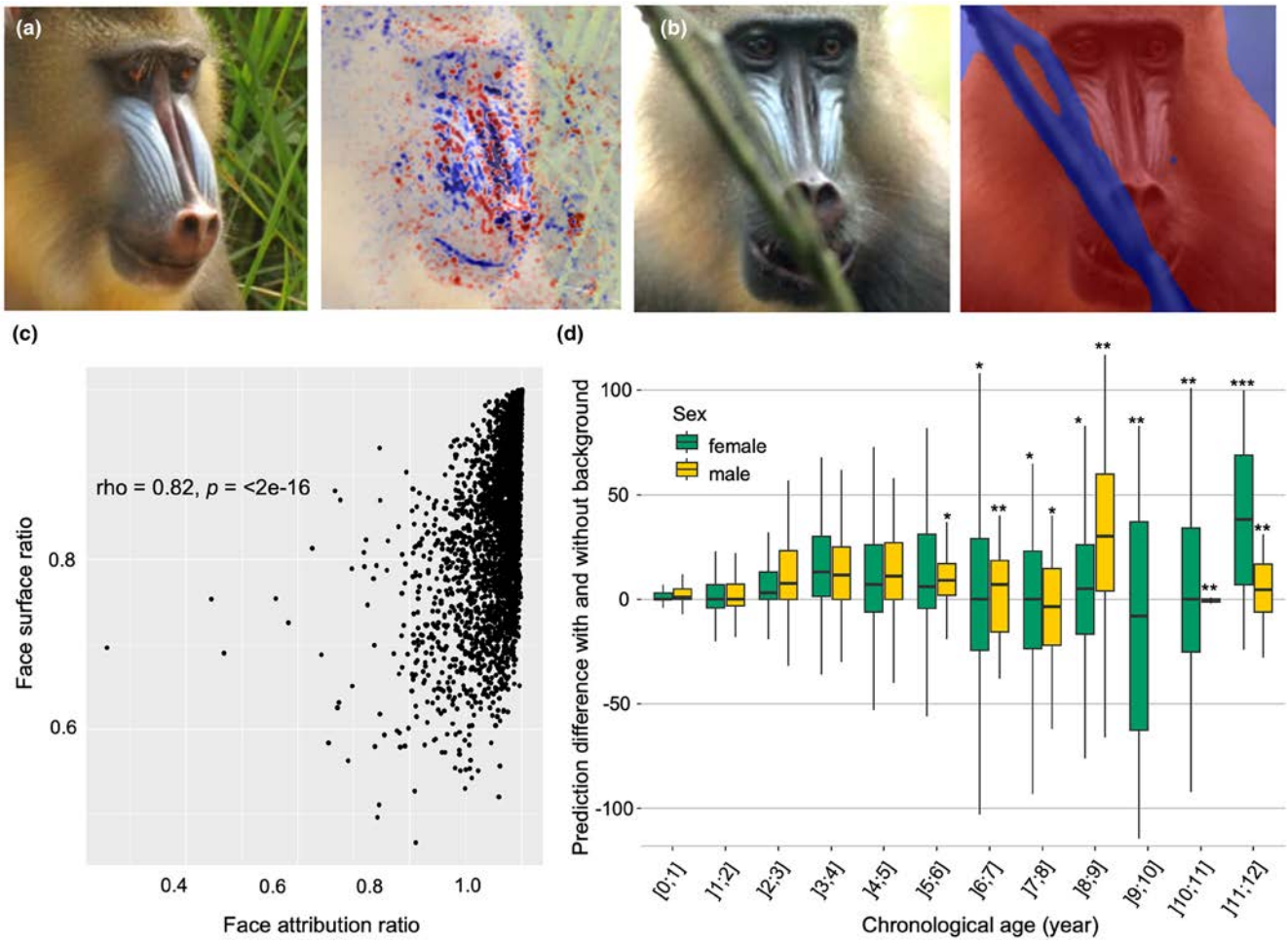

FIGURE 2 Analysis of background-based shortcut learning. The analyses were performed on the first fold of the fivefold cross-validation procedure. (a) Example of an image (left) with particularly high attribution scores (right) for background features. Red pixels contribute to increasing the predicted age, while blue pixels decrease it. (b) Example of a 'face' (in red) and a 'background' (in blue) segmentation: Original image (left) and segmentation masks (right). The face includes the fur, while the background may also contain, as in this case, occluding foreground elements. (c) Correlation between the ratio of attribution scores for the face versus the background and the ratio of the face versus background surface area. The rho value indicates the Spearman rank correlation score, together with its *p*-value. (d) Difference between predictions with and without the background for different age classes between 0 and 12 years. The relative difference is calculated for each image. * classes with fewer than 100 images; ** classes with fewer than 50 images; *** class with fewer than 5 images.

3.3 | Influence of biological factors

3.3.1 | Sex

Figure 1 reveals that MAE increases more rapidly with age for females than for males. This difference between sexes appears at the age of 4, and from 6 years onward, MAE for females is almost double that for males (Figure 1c). Generalized linear models confirmed this pattern of sex differentiation (Table S11). When fitting data over 0–3 years of age, MAE was not significantly different between males and females; however, when fitting data over 0–12 or 0–24 years (thus including both 0–3 year and older individuals), MAE was significantly smaller for males than for females (Table S11).

3.3.2 | Inter-individual variation

Inter-individual variation of MAE increased with chronological age, especially in females (Figures 1a,c and 3). For all ranges of chronological ages (0–3, 0–12 or 0–24) and age intervals tested (i.e. pooling mean individual's MAE by month, trimester or year), we showed a significant effect of chronological age, independently of the number of pictures considered, on the standard deviation (SD) of accuracy of predicted ages, but not on the coefficient of variation (CV; Figure S5; see details in Table S14). Consequently, while the absolute variation in prediction accuracy decreased as chronological age increased (SD), the variation relative to chronological age remained relatively stable as individuals aged (CV). This result indicates that the decrease in accuracy with age was a natural consequence of inter-individual variation in predicted age, which increased with age, rather than a limitation of our models to accurately predict the age of older individuals (e.g. because of fewer training data).

3.3.3 | Intra-individual variation

For a given age interval, we compared the average difference between every possible pair of photos collected on the same individual with the average difference between an equivalent number of randomly drawn pairs of photos taken on different individuals (Figure 4). For the 0–1 year period and a 15-day interval, the intra-individual and inter-individual variations were on average 17.6 days and 25.1 days, respectively. For the 2–3 year period and a 1-month interval, these differences were 59.7 days and 104.4 days, respectively. Finally, for the 4–12 year period and a 1-month interval, the intra-individual and inter-individual differences were on average 149.0 days and 245.6 days, respectively. These analyses confirmed that inter-individual variation was always greater than intra-individual variation.

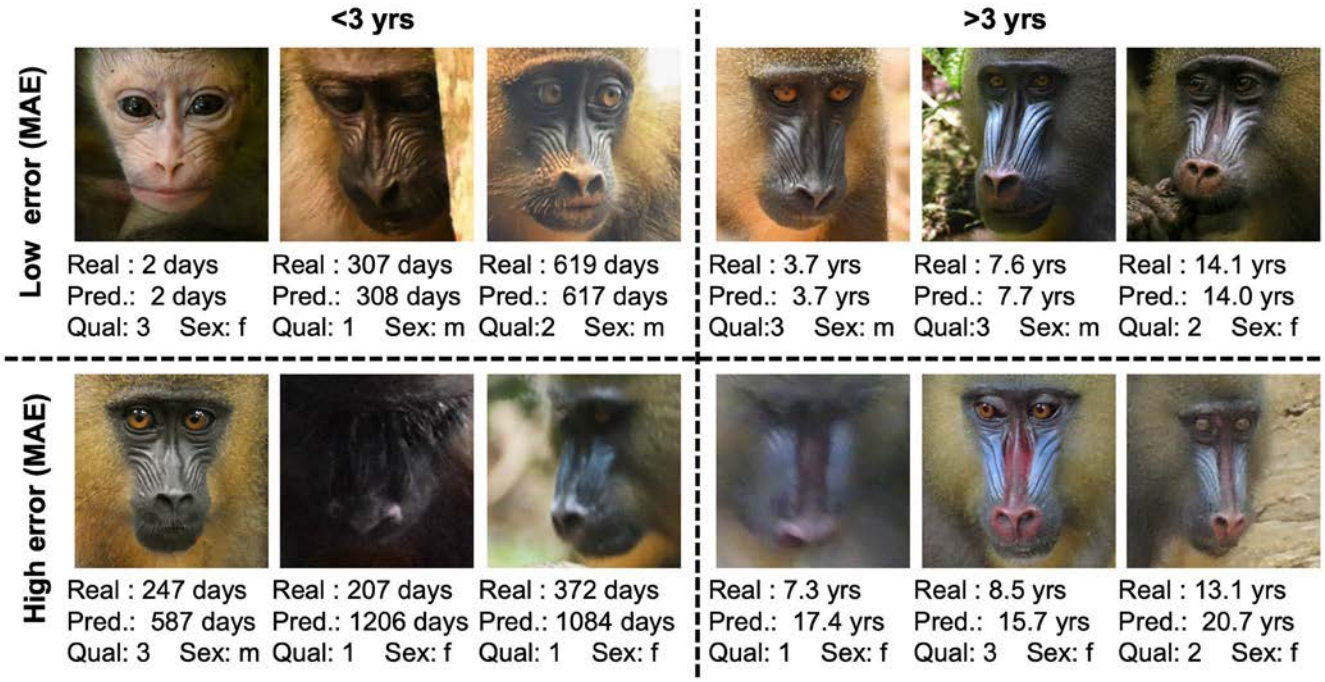

FIGURE 3 Example of images with extremely low and extremely high error (MAE). Predictions obtained with the best models: DINOv2-LARGE trained with 0–3, 0–12 and 0–24 year-old individuals to predict the age of other individuals of 0–3, 0–12 and 0–24 year old, respectively, with the age predicted in years (secondarily converted in days for <3 years individuals here). The figure illustrates the influence of external factors on age prediction, like the quality of the image ('Qual') and the sex of the individual (e.g. old females were predicted to be older than they were), but high error can occur even for high-quality images of young males. Such errors partly reflect inter-individual variation in developmental progress (e.g. the 372 day-old female in the bottom row shows a blue tint that typically appears in older individuals).

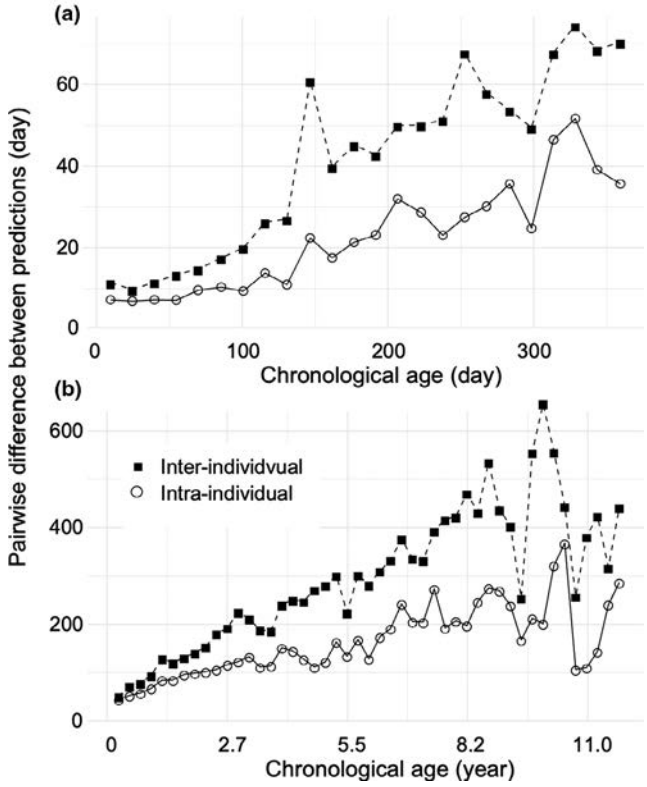

FIGURE 4 Comparison of intra- and inter-individual variations in accuracy. Each point represents the pairwise difference in predicted age calculated between same and different-individual pictures, averaged for a given age interval. (a) Individuals aged 0–1 year, with pairwise differences averaged over 15-day intervals. (b) Individuals aged 0–12 years, with pairwise differences averaged by trimesters.

3.3.4 | Predicted versus chronological age throughout life

We calculated the probability that an individual whose predicted age was higher than their chronological age at a given time also showed a predicted age still higher when they were older. For individuals aged 0–1 year old, the probability that two photos simultaneously had a predicted age higher than the chronological age was significantly greater when the photos were of the same individual than when they were of different individuals, provided the photos were taken less than 3 months apart (Figure 5a). When the interval between the two photos was greater than 4 months, however, this probability became non-significant. For older individuals, when an individual's predicted age was higher than its chronological age at 2–3 years old, it was also consistently and significantly higher when aged 4, 5, 6, 7 or even 9 years old (larger intervals could not be tested due to insufficient data, Figure 5b).

3.3.5 | Maternal age

The linear mixed-effects model revealed significant positive effects of maternal age at offspring birth ($\beta = 0.50 \pm 0.15$, $p = 0.0009$) and chronological age ($\beta = 4.96 \pm 1.03$, $p < 0.0001$) on the relative prediction error (Table S15). Neither maternal rank ($\beta = 0.59 \pm 2.17$, $p = 0.787$) nor sex ($\beta = -1.57 \pm 1.37$, $p = 0.252$) had significant effects. In other words, older mothers tend to produce infants who appear older than their actual age when they are 0–1 years old, while younger mothers tend to produce offspring who appear younger for age, independently of their social rank.

4 | DISCUSSION

In this study, we aimed to evaluate the performance of state-of-the-art AI methods in predicting chronological age from a large database of photographic portraits collected on wild mandrills and to investigate the influence of biological factors on predictions. Our results indicate a very good performance of deep learning models in predicting chronological age in mandrills, particularly the vision transformers DINOv2. For example, for an individual aged 6 months, the best model provided predictions with an average error of 33 days, while for individuals aged 5 years, the average error was 6 months. Although these errors may appear large, they are relatively low, both when compared to benchmarks in human age prediction and considering that they reflect real biological variability.

4.1 | Age prediction in mandrills

Overall, DINOv2 appears highly performant in predicting chronological age of female and male mandrills of different ages. Firstly, the accuracy is close to the state-of-the-art in human age prediction. In computer sciences, it is generally difficult to compare performances across training schemes because of the multiple factors interfering with the results. The training and validation datasets, in particular, have different characteristics (e.g. age range, imbalance level) that strongly influence the results. Additionally, recent studies tend to establish benchmarks by predicting apparent age (i.e. the age perceived by humans) rather than chronological age, to estimate model performance independently of biological variations (Qin et al., 2023). The evaluation metrics used vary between the tasks tested (e.g. regression vs. classification) but they are not standardized even for a given task. Here, we will base our comparison on results presented by Kuprashevich and Tolstykh (2023), who tested their own model as well as the main state-of-the-art models on the same datasets and using the same two metrics: MAE and CS@5.

With DINOv2-LARGE and within the age range of 0–12 years, we obtained a MAE of 135.9 days, which is equivalent to approximately 3.3 years using our formula to shift between human ages and mandrill ages and a CS@1 of 90.8 (i.e. 90.8% of predictions have MAE <1 year; CS@5 in humans is equivalent to CS@1 in mandrills). With the entire range of 0–24 years, we obtained a MAE of 218.2, equivalent to approximately 3.9 years in humans, and a CS@1 of 83.2.

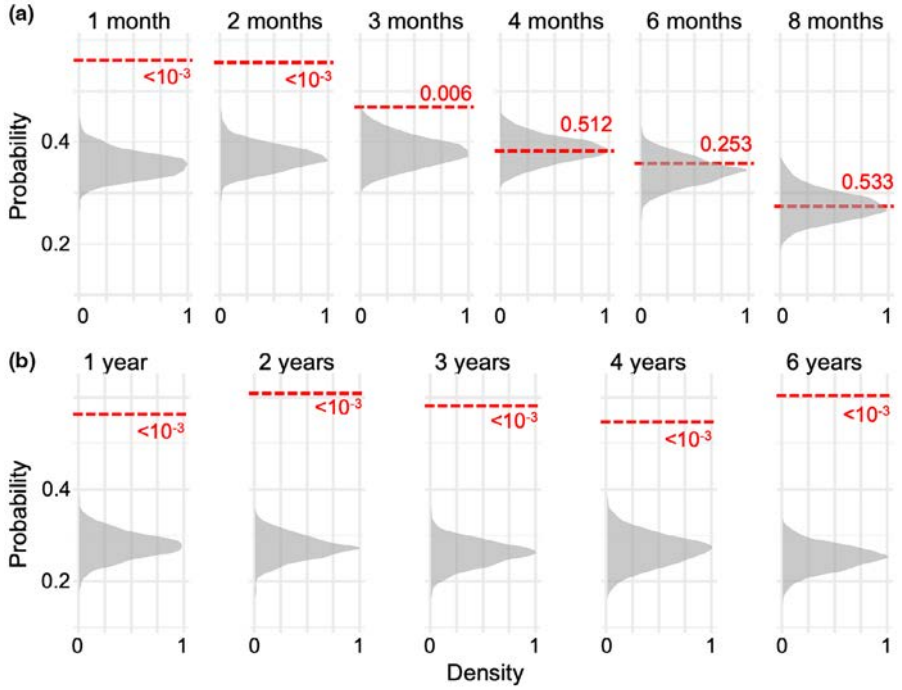

FIGURE 5 Predicted versus chronological ages throughout life. The dashed red line indicates the probability that an individual whose predicted age was higher than their chronological age at a given age had a predicted age still higher when they were older, for (a) individuals aged 0–1 year old and for (b) older individuals. The grey density represents the null distribution of this probability, obtained by repeating the previous probability calculus 1000 times while randomly selecting two photos from two different individuals at each iteration, instead of using two photos of the same individual as in the probability represented by the red line. *p*-values are provided in red (one-tailed tests based on randomly paired individuals with 1000 iterations). Each panel represents a different time interval between the two compared photos.

Using the UTKFace dataset (16.2 k images; Zhang et al., 2017), CORAL (a rank-consistent, deep ordinal regression model; Cao et al., 2020) reached a MAE of 5.4 years, while MWR (a moving-window deep ordinal regression model; Shin et al., 2022) reached a MAE of 4.4 years. These errors are therefore larger than ours. UTKFace did not, however, include young individuals (only people aged [21, 60]), which lowers the average accuracy. With the IMDB-Clean dataset (102 k images; Lin et al., 2022), the FP-Age model (a face parsing-based network that learns semantic information at different scales; Lin et al., 2022) had MAE = 4.68/CS@5 = 63.78, and VOLO-D1 had MAE = 4.39/CS@5 = 67.71 in single-task prediction, and MAE = 4.22/CS@5 = 68.68 in multi-task prediction (age and gender prediction; Kuprashevich & Tolstykh, 2023). This dataset represents celebrities of all ages but with a strong excess within the range [25–45]. Finally, with the LAGENDA dataset (67.2 k images; Kuprashevich & Tolstykh, 2023), VOLO-D1 achieved MAE = 4.19/CS@5 = 69.36 in single-task prediction and MAE = 4.11/CS@5 = 70.11 in multi-task prediction (Kuprashevich & Tolstykh, 2023). In this dataset, the chronological age is uniformly distributed [0, 70]. In our dataset, the oldest females are 24 years old, which is equivalent to 67 years in human age. The age range is thus similar; however, because we have many more young individuals than older ones (Figure S1), our accuracies are certainly overestimated compared to those of these last models, which currently represent the state-of-the-art in human age prediction. Regardless, these results show that our performance is comparable to the state-of-the-art in age prediction in humans.

Importantly, the model did rely on facial features to predict age and the influence of the background remained marginal. Removing the background did not significantly modify the model's performance. However, we found that the background could occasionally influence the prediction, particularly in older individuals. This finding is surprising, and we do not have a clear explanation at this stage. The studied mandrill group is wild, moves constantly across a wide range composed of forest undergrowth, and experiences only weak seasonality. Moreover, photos are taken at all times of the day. Our results nevertheless suggest that there may be systematic, previously unrecognized biases in the way individuals, especially older ones, position themselves within their environment. Although background-based shortcut learning is often perceived as noise, its analysis may in fact reveal previously unnoticed signals (see Xiao et al., 2020) and, in the context of ecology, shed light on individual behaviour.

Several potential limits to our study, stemming from the nature of our data, must nevertheless be acknowledged. First, our training and validation datasets are not totally independent. While we ensured that photos of the same individual are not present in both datasets, it does not account for the fact that individuals exhibit various degrees of relatedness. In a previous study on this population, we showed that genetic relatedness is correlated with the distance between faces in the latent space of a CNN trained for individual re-identification (Charpentier et al., 2020). Furthermore, temporal autocorrelation within the training set (i.e. the presence

of multiple photos of the same individual at different ages) could allow the model to learn statistical shortcuts, by which the model recognizes features associated with age-related milestones rather than learning the overall, continuous aging process. This is not necessarily a problem, but this suggests that the learning strategy may be different from, and that the model's performance may not be generalizable to other studies with portraits of entirely different individuals.

4.2 | Biological factors influence prediction errors

In humans, the main reason to explain why model performance has reached a ceiling is the effective discrepancy between chronological age and apparent age (Agustsson et al., 2017). The apparent age, or how old an individual looks, correlates with various health outcomes and can even predict mortality, independently from chronological age (Christensen et al., 2009). The discrepancy between chronological and apparent ages has been used as a biological measure with significant meanings in developmental and aging studies (Salih et al., 2023). Multiple lines of evidence suggest that the discrepancy we obtained is also partly explained by biological factors.

First, we found that predictions across multiple portraits of the same individual are more consistent than predictions obtained across different individuals of the same chronological age. This suggests that the prediction errors are neither primarily due to randomness or noise in the model, nor are they driven by differences in image quality or pose, but might instead reflect real biological variation across individuals.

Second, we found that within-individual predictions remain coherent over time: individuals who appear older than their chronological age at a given time also appear older later on, with one notable exception. Indeed, in mandrill infants this result disappears when the interval between two photos was greater than 4 months. Environmental factors encountered during infancy may reshape genetic influences or prenatal maternal effects, explaining why facial developmental precocity or delay in the early months does not persist beyond several months during this first year of life. After 2 years of age, however, any developmental precocity or delay relative to the rest of the population, as estimated through facial traits, appears to be fixed and potentially maintained for life. In humans, livestock and laboratory animals, environmental factors experienced during infancy can also have long-lasting effects on development, either through epigenetic or physiological mechanisms (Champagne, 2011; Meaney, 2001). These effects can persist throughout life if the factors act during specific developmental windows (Knudsen, 2004).

It is important to note, however, the limits of this second line of evidence taken in isolation, as some form of shortcut learning could alternatively explain these results. If the model learns to associate a static, non-aging-related characteristic with a particular age range, individuals in the validation set who possess a similar, unchanging feature will have their age systematically under- or overestimated. For example, if the model learns that a short snout is a feature of younger mandrills, when it sees a new, older mandrill in the validation set that happens to have a naturally short snout, the model might consistently predict a younger age for this individual throughout its life.

Third, variation in prediction errors reflects key aspects of mandrill developmental biology. In females, MAE increases noticeably around 6–8 years of age (Figure 1), which corresponds to the typical age of adult body size attainment in mandrills (Setchell et al., 2001). Similarly, males exhibit a sharp rise in MAE around 10 years of age, an age when somatic growth plateaus. Notably, the standard deviation of MAE remains low in males between 5 and 9 years, a period marked by rapid growth acceleration. This suggests that during this window, body size and by extension, facial morphology, serve as a reliable proxy for chronological age. Furthermore, sex differences in MAE become apparent at approximately 4–5 years of age, which corresponds to the onset of visible sexual size dimorphism. Collectively, these findings align with well-established observations in humans that the rapid and hormonally driven morphological changes of adolescence are associated with highly predictable age-related variation (Marshall & Tanner, 1969). Once somatic development slows down or ceases, however, aging leads to greater inter-individual variability in physical traits. Again, this pattern parallels what is observed in humans, where the transition from adolescence to adulthood is marked by increasing and then plateauing variance in biological aging (Işıldak et al., 2020; Kuznetsov et al., 2024).

Fourth, the influence of maternal age, but not rank, on infant age errors corroborates previous findings on captive unweaned mandrills using morphological measurements obtained during captures. Specifically, infants born to older females tend to be heavier than those born to younger mothers (Setchell et al., 2001) and, in our study, also appear older for their age. This suggests that weight differences among individuals may be perceptible from facial features alone and are detected by our AI-based tool.

Altogether, these biological validations suggest that our non-invasive deep learning tool, which uses non-standardized photographic portraits taken in the wild, can provide a valuable biological clock that could inform about developmental and aging processes in mandrills, and more broadly, in nonhuman primates.

4.3 | Vision transformer for behavioural ecology

Our model comparison confirms that within the AI landscape, foundation models based on vision transformers are emerging as the models of choice for image analysis. Beyond their architectural innovation, vision transformers shift the processing paradigm from local feature aggregation (characteristic of CNNs) to early global representation learning via self-attention mechanisms. One potential drawback for this high capacity is an increased risk of overfitting, which may reduce performance compared to smaller models, when

training data is limited (Dosovitskiy et al., 2020). However, when applied to the Mandrillus Face Database, we did not observe evidence of overfitting when training DINOv2 models, as we monitored the accuracy and loss learning curves on both the training and validation sets (results not shown). This shows that DINOv2 can be fine-tuned even on datasets of a few tens of thousands, and possibly a few thousand images.

Nevertheless, two bodies of evidence suggest that, despite its large number of parameters, DINOv2 experiences trade-offs in its learning capacities, and thus that all parameters seem necessary to perform optimally in the age prediction task. First, the accuracy was improved when the training age range was only slightly larger than the age range to predict, while it was reduced when the training range was too large. While this result highlights that the features used to predict age likely differ among the different age groups studied (0–3, 3–12 and over 12 years), it also indicates that it is difficult for DINOv2 to perform well simultaneously across all age groups (for similar findings, see Xia et al., 2020). Second, we found that single-task learning (STL) outperformed multi-task learning (MTL) in our deep regression model for age prediction. The explanation is probably similar for these two bodies of evidence: a narrow training age range and STL both allow the model to focus exclusively on age-specific features without interference from other tasks.

The implications of our work extend far beyond age prediction and primates. The ability of ViTs to extract biologically meaningful features from non-standardized images holds promise for many behavioural ecology applications. These models could be repurposed to track changes in health, detect hormonal status (e.g. oestrus), identify injuries or parasitic burdens or classify behavioural states in various taxa. Overall, this study supports the integration of AI-driven phenotyping in long-term ecological research. By linking external phenotypes with internal physiological markers and life-history traits, such models offer a non-invasive, scalable alternative to traditional methods, bridging the gap between field observation and biological inference.

AUTHOR CONTRIBUTIONS

Julien P. Renoult and Marie J.E. Charpentier conceived the ideas and designed the methodology; Loïc Sauvadet, Mélodie Kreyer, Richard Mbadoumou, Berta Roura-Torres, Alice Baniel and Marie J.E. Charpentier collected the data (both images and birth dates); Romain Karpinski performed the deep learning experiments, Julien P. Renoult and Marie J.E. Charpentier analysed the model predictions; Julien P. Renoult led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

The collaboration between JPR, MJEC and RK was funded by the Programme National de Recherche en IA (PNRIA). We are grateful to the past and present field assistants of the Mandrillus Project for their daily data collection. We also thank the SODEPAL-COMILOG society (ERAMET group) for their long-term logistical support. This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011014579 to JPR and MJEC). The Mandrillus Project has been funded by several grants that allowed long-term data collection including: SEEG Lekedi and SEE-LIFE initiative (INEE-CNRS), the Leakey Foundation (S202210309), the Max Planck Society (all to MJEC) and the Agence Nationale de la Recherche (ANR-20-CE02-0005-01; to JPR). This is a Mandrillus Project Publication number 37.

CONFLICT OF INTEREST STATEMENT

We have no conflicts of interest to declare.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70187.

DATA AVAILABILITY STATEMENT

The images analysed are available in the Mandrillus Face Database, accessible online at https://zenodo.org/records/7467318 (Tieo et al., 2023). The code to reproduce the training and testing of the AI models, and the R code and data files (AI model predictions) for analysing the error are accessible online at https://zenodo.org/records/17376792 (Renoult & Karpinski, 2025).

STATEMENT ON INCLUSION

Our study is part of the Mandrillus Project, a long-term monitoring programme based in Gabon. This programme is dedicated to training Gabonese students, including doctoral candidates, and employs field assistants from the village that hosts the project's facilities. This study is co-authored by both the French and Gabonese data collection managers and field assistants, three of whom are residents of Gabon.

ORCID

Julien P. Renoult https://orcid.org/0000-0001-6690-0085

REFERENCES

Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., & Rothe, R. (2017). Apparent and real age estimation in still images with deep residual regressors on appa-real database.

Angulu, R., Tapamo, J. R., & Adewumi, A. O. (2018). Age estimation via face images: A survey. *EURASIP Journal on Image and Video Processing*, *2018*(1), 1–35.

Beissinger, S. R., & Westphal, M. I. (1998). On the use of demographic models of population viability in endangered species management. *The Journal of Wildlife Management*, *62*, 821–841.

Bewick, A. J., Vogel, K. J., Moore, A. J., & Schmitz, R. J. (2017). Evolution of DNA methylation across insects. *Molecular Biology and Evolution*, *34*(3), 654–665.

Bobrov, E., Georgievskaya, A., Kiselev, K., Sevastopolsky, A., Zhavoronkov, A., Gurov, S., Rudakov, K., Tobar, M. d. P. B., Jaspers, S., & Clemann, S. (2018). PhotoAgeClock: Deep learning algorithms for development of non-invasive visual biomarkers of aging. *Aging*, *10*(11), 3249–3259.

Bogin, B., & Smith, B. H. (1996). Evolution of the human life cycle. *American Journal of Human Biology*, *8*(6), 703–716.

Campana, S. E., & Thorrold, S. R. (2001). Otoliths, increments, and elements: Keys to a comprehensive understanding of fish populations? *Canadian Journal of Fisheries and Aquatic Sciences*, *58*(1), 30–38.

Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, *140*, 325–331.

Chaloupka, M., & Musick, J. A. (2017). Age, growth, and population dynamics. *The Biology of SEA Turtles*, *I*, 233–276.

Champagne, F. A. (2011). Maternal imprints and the origins of variation. *Hormones and Behavior*, *60*(1), 4–11.

Charpentier, M. J., Harté, M., Poirotte, C., de Bellefon, J. M., Laubi, B., Kappeler, P., & Renoult, J. P. (2020). Same father, same face: Deep learning reveals selection for signaling kinship in a wild primate. *Science Advances*, *6*(22), eaba3274.

Christensen, K., Thinggaard, M., McGue, M., Rexbye, H., Aviv, A., Gunn, D., van der Ouderaa, F., & Vaupel, J. W. (2009). Perceived age as clinically useful biomarker of ageing: Cohort study. *BMJ (Clinical Research Ed.)*, *339*, b5262.

Dornaika, F., Bekhouche, S. E., & Arganda-Carreras, I. (2020). Robust regression with deep CNNs for facial age estimation: An empirical study. *Expert Systems with Applications*, *141*, 112942.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv Preprint arXiv:2010.11929.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*, 1–20.

Işıldak, U., Somel, M., Thornton, J. M., & Dönertaş, H. M. (2020). Temporal changes in the gene expression heterogeneity during brain development and aging. *Scientific Reports*, *10*(1), 4080.

Jarman, S. N., Polanowski, A. M., Faux, C. E., Robbins, J., De Paoli-Iseppi, R., Bravington, M., & Deagle, B. E. (2015). Molecular biomarkers for chronological age in animal ecology. *Molecular Ecology*, *24*(19), 4826–4847.

Knudsen, E. I. (2004). Sensitive periods in the development of the brain and behavior. *Journal of Cognitive Neuroscience*, *16*(8), 1412–1425.

Kuprashevich, M., & Tolstykh, I. (2023). *Mivolo: Multi-input transformer for age and gender estimation*. 212–226.

Kuznetsov, D. V., Liu, Y., Schowe, A. M., Czamara, D., Instinske, J., Pahnke, C. K., Noethen, M. M., Spinath, F. M., Binder, E. B., & Diewald, M. (2024). Age-associated genetic and environmental contributions to epigenetic aging across adolescence and emerging adulthood. bioRxiv, 2024–06.

Lapuschkin, S., Binder, A., Muller, K.-R., & Samek, W. (2017). *Understanding and comparing deep neural networks for age and gender classification*. IEEE.

Lin, Y., Shen, J., Wang, Y., & Pantic, M. (2022). Fp-age: Leveraging face parsing attention for facial age estimation in the wild. In *IEEE Transactions on Image Processing*. IEEE.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A convnet for the 2020s*. 11976–11986.

Maddigan, P., Ehrhardt, O., Lensen, A., & Shaw, R. C. (2024). *Re-Identification of Individual Kākā: An Explainable DINO-Based Model*. 1–6.

Marshall, W. A., & Tanner, J. M. (1969). Variations in pattern of pubertal changes in girls. *Archives of Disease in Childhood*, *44*(235), 291–303.

Meaney, M. J. (2001). Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations. *Annual Review of Neuroscience*, *24*(1), 1161–1192.

Meng, D., Zhang, S., Huang, Y., Mao, K., & Han, J.-D. J. (2024). Application of AI in biological age prediction. *Current Opinion in Structural Biology*, *85*, 102777.

Nada, A. A., Alajrami, E., Al-Saqqa, A. A., & Abu-Naser, S. S. (2020). Age and gender prediction and validation through single user images using CNN. *International Journal of Academic Engineering Research*, *4*, 21–24.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., & El-Nouby, A. (2023). *Dinov2: Learning robust visual features without supervision*. arXiv Preprint arXiv:2304.07193.

Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., & Deng, W. (2023). SwinFace: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. In *IEEE transactions on circuits and systems for video technology*. IEEE.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, *34*, 12116–12128.

Ravi, N., Gabeur, V., Hu, Y. T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., & Feichtenhofer, C. (2024). Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714.

Renoult, J. P., & Karpinski, R. (2025). Vision transformers for age prediction from facial images in a wild primate- Code & Data. *Zenodo Repository*. https://zenodo.org/records/17376792

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.

Rughetti, M. (2016). Age structure: An indicator to monitor populations of large herbivores. *Ecological Indicators*, *70*, 249–254.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Salih, A., Nichols, T., Szabo, L., Petersen, S. E., & Raisi-Estabragh, Z. (2023). Conceptual overview of biological age estimation. *Aging and Disease*, *14*(3), 583–588.

Setchell, J. M., Lee, P. C., Wickings, E. J., & Dixson, A. F. (2001). Growth and ontogeny of sexual size dimorphism in the mandrill (*Mandrillus sphinx*). *American Journal of Physical Anthropology*, *115*(4), 349–360.

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., & Yuille, A. L. (2018). *Deep regression forests for age estimation*. 2304–2313.

Shin, N.-H., Lee, S.-H., & Kim, C.-S. (2022). *Moving window regression: A novel approach to ordinal regression*. 18760–18769.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv Preprint arXiv:1409.1556.

Sundararajan, M., Taly, A., & Yan, Q. (2017). In D. Precup & Y. W. Teh (Eds.), *Proc. 34th International Conference on Machine Learning* (pp. 3319–3328). PMLR.

Sutherland, W. J., & Norris, K. (2002). Behavioural models of population growth rates: Implications for conservation and prediction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *357*(1425), 1273–1284.

Tan, M., & Le, Q. (2021). *Efficientnetv2: Smaller models and faster training*. 10096–10106.

Tieo, S., Restrepo-Ortiz, C. X., Roura-Torres, B., Sauvadet, L., Harté, M., Charpentier, M. J., & Renoult, J. P. (2023). The Mandrillus face database: A portrait image database for individual and sex recognition, and age prediction in a non-human primate. *Data in Brief*, *47*, 108939.

Vaiserman, A., & Krasnienkov, D. (2021). Telomere length as a marker of biological age: State-of-the-art, open issues, and future perspectives. *Frontiers in Genetics*, *11*, 630186.

Wen, X., Li, B., Guo, H., Liu, Z., Hu, G., Tang, M., & Wang, J. (2020). *Adaptive variance based label distribution learning for facial age estimation*. 379–395.

Wickings, E., & Dixson, A. (1992). Development from birth to sexual maturity in a semi-free-ranging colony of mandrills (*Mandrillus sphinx*) in Gabon. *Reproduction*, *95*(1), 129–138.

Wightman, R. (2019). PyTorch Image Models. GitHub Repository, https://github.com/rwightman/pytorch-image-models https://doi.org/10.5281/zenodo.4414861

Wittwer-Backofen, U., Gampe, J., & Vaupel, J. W. (2004). Tooth cementum annulation for age estimation: Results from a large known-age validation study. *American Journal of Physical Anthropology*, *123*(2), 119–129.

Xia, M., Zhang, X., Weng, L., & Xu, Y. (2020). Multi-stage feature constraints learning for age estimation. *IEEE Transactions on Information Forensics and Security*, *15*, 2417–2428.

Xiao, K., Engstrom, L., Ilyas, A., & Madry, A. (2020). *Noise or signal: The role of image backgrounds in object recognition*. arXiv preprint arXiv:2006.09994.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). *Aggregated residual transformations for deep neural networks*. 1492–1500.

Yuan, L., Hou, Q., Jiang, Z., Feng, J., & Yan, S. (2022). Volo: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 6575–6586.

Zang, H.-X., Su, H., Qi, Y., Feng, L., Hou, R., He, M., Liu, P., Xu, P., Yu, Y., & Chen, P. (2022). Ages of giant panda can be accurately predicted using facial images and machine learning. *Ecological Informatics*, *72*, 101892.

Zha, K., Cao, P., Son, J., Yang, Y., & Katabi, D. (2023). Rank-n-contrast: Learning continuous representations for regression. In A. Oh, et al. (Eds.), *Proc. Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023* (pp. 10–16).

Zhang, Z., Song, Y., & Qi, H. (2017). *Age progression/regression by conditional adversarial autoencoder*. 5810–5818.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Number of images and individuals in relation to chronological age and sex.

Figure S2. Cumulative number of images and individuals in relation to accuracy in date of birth and sex.

Figure S3. Effect of the number of images on the intra-individual variation in accuracy.

Figure S4. Relation between human age and mandrill age.

Figure S5. Inter-individual variation in predicted ages.

Table S1. Effect of age normalization on age prediction accuracy.

Table S2. Effect of age normalization on age prediction accuracy, dataset restricted to image quality scores 2–3.

Table S3. Image augmentation applied during model training.

Table S4. Comparison of accuracy in age prediction across architectures.

Table S5. Comparison of accuracy in age prediction across architectures, dataset restricted to image quality scores 2–3.

Table S6. Effect of age range in the training set.

Table S7. Accuracies based on age categorized into 1-year intervals, for individuals aged 0–24 years.

Table S8. Accuracies based on age categorized into one-semester intervals, for individuals aged 0–12 years.

Table S9. Accuracies based on age categorized into 15-day intervals, for individuals aged 0–1 years.

Table S10. Effect of the training task.

Table S11. Results of statistical models fitting the accuracy as a function of sex and image quality.

Table S12. Developmental periods in humans and mandrills.

Table S13. Results of statistical models fitting the difference in prediction with and without the background as a function of sex and age.

Table S14. Results of statistical models fitting the inter-individual variation in accuracy as a function of age.

Table S15. Results of the statistical model fitting the relative prediction error as a function of maternal age and social rank.

How to cite this article: Renoult, J. P., Karpinski, R., Sauvadet, L., Kreyer, M., Mbadoumou, R., Roura-Torres, B., Baniel, A., & Charpentier, M. J. (2025). Vision transformers for age prediction from facial images in a wild primate. *Methods in Ecology and Evolution*, *00*, 1–15. https://doi.org/10.1111/2041-210X.70187