2023-09-25

Age-constrained ear recognition: the EICZA dataset and SASE baseline model

W. Qin, L. Etter, A. Simukanga, C. Gill, M. Betke. 2023. "Age-constrained Ear Recognition: The EICZA Dataset and SASE Baseline Model" IEEE IJCB 2023.

https://hdl.handle.net/2144/47479

Age-constrained Ear Recognition:
The EICZA Dataset and SASE Baseline Model

Wenda Qin[1,B], Lauren Etter[2,B], Alinani Simukanga[1,Z], Christopher J. Gill[2,B], and Margrit Betke[1,B]
[1]Department of Computer Science, [2]Department of Global Health
[B]Boston University and [Z]University of Zambia
{wdqin, laetter, cgill, betke}@bu.edu, alinani10@gmail.com

Abstract

Using the ear as a biometric identifier, particularly for children in healthcare settings, has an important advantage over using the face – the privacy of the person can be protected better. However, aging and the resulting appearance differences, known to be challenges for face recognition models, have not been addressed for ear recognition yet. To address this limitation, we curated a publicly available dataset, which we call Ears of Infant Cohort in Zambia with Aging *(EICZA)* [1] *. The dataset contains 3,330 ear images of 177 subjects, each photographed multiple times between the ages of 6 days and 9 months, when ear growth is most significant. For the task of age-constrained ear recognition, i.e., recognizing a person who has aged since the model was trained, we propose a new ear recognition model, called SASE for* ***S****elf-****A****ttention-based* ***S****equential* ***E****ar image analysis. The model takes a sequence of ear images at early but different ages as input (instead of a single image) and processes them with a feature extraction network and a Transformer encoder. Trained with a large margin cosine loss function, the model is encouraged to learn a feature representation that distinguishes subjects from each other. Our experiments show that accounting for age enables our model to outperform other models that do not in recognizing ears that have grown and look different in later time periods.*

1. Introduction

Ear images can be used as biometric identifiers with unique benefits compared to traditional face or fingerprint systems. They provide better privacy, as they are usually not as recognizable by humans as faces are. An ear photograph can easily be captured with the camera of a mobile device, and this can be done contact-free, unlike collecting a fingerprint. The ear as a biometric identifier also has the advantage that its acquisition "does not depend on the cooperativeness of the person one is trying to recognize" [18]. Therefore, numerous works, e.g. [18, 31, 60] have explored the potential of ear-based biometric systems, including deep-learning models for ear recognition [4, 16, 20, 59] in unconstrained conditions (also called "in the wild").

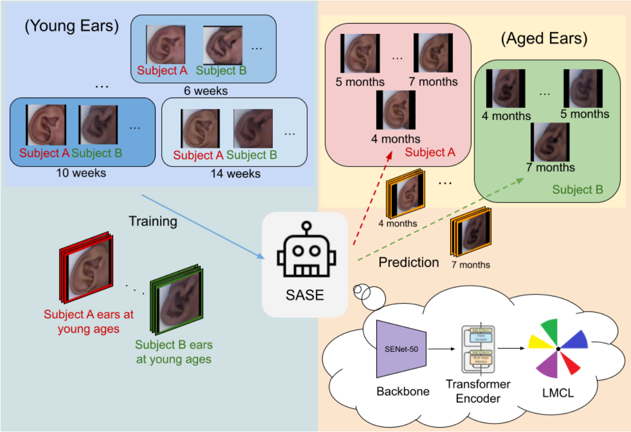

Figure 1. Visualization of our proposed age-constrained ear recognition task and our **S**elf-**A**ttention-based **S**equential **E**ar image analysis (SASE) model. During training, SASE takes ear images of younger ages as inputs, while during inference, SASE identifies a person from one or more ear images taken at a later age. Using a backbone network and Transformer, SASE learns a subject representation with the large margin cosine loss (LMCL) function.

In this paper, we propose a new recognition task for ears, that is, identifying a person based on ear images that look different due to aging, see Fig. 1. This task is suitable for healthcare settings, particularly for identifying children.

Our task relates to the biometrics task of age-invariant face recognition (AIFR) [21, 33, 56], which has the goal of identifying people irrespective of their age. To overcome the challenges of larger intra-class than inter-class differences – people at the same age often look more similar to each other than their own much younger selves – methods

[1] Funding by the National Institutes of Health, grant 1R21TW010939-01, is gratefully acknowledged.

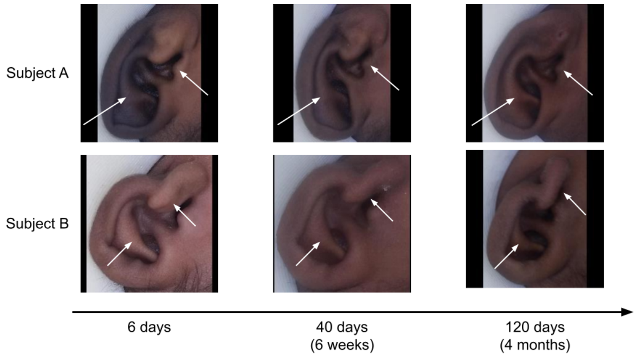

Figure 2. Ears of subjects at different ages. The arrows show some of the significant differences among ears of different ages.

aim to compute age-*invariant*, subject-discriminatory features. Models are typically trained on face images at earlier ages and tested on a single image at the current age. Our proposed task is similar but for a different target, i.e., ears: Given one or more images of the ears of a subject at an older age, an ear recognition model is asked to identify the subject after having been trained only on ear images that were recorded when the subject was younger. We propose to call this the age-*constrained* recognition task.

The task is valuable for real-life applications, such as healthcare systems to identify patients over long time spans (e.g., decades), or infants during the first few weeks or months, when their ears grow most rapidly (see the intra-class infant ear differences in Fig. 2). Even for general ear recognition applications, a model that can handle differences caused by aging may also be able to handle recognition challenges caused by physical alterations (piercings) and accessories (earrings).

To investigate the task of age-constrained ear recognition, we curated a dataset we call EICZA for **E**ars of **I**nfant **C**ohort in **Z**ambia with **A**ging. EICZA contains 3,330 ear images of 177 subjects with multiple ear photos of the same subject taken at different ages. We experimented with four existing feature extraction models as baselines, fine-tuning them on EICZA. We then propose a model that directly addresses the issue of ear aging. We call our model SASE for **S**elf-**A**ttention-based **S**equential **E**ar image analysis. It is based on the concept of processing sequential image input with a deep learning architecture. SASE consists of a pre-trained backbone Squeeze-and-Excitation Network [9, 27] that extracts a sequence of features that are then interpreted by a Transformer [51]. We use a large margin cosine loss function [54] to both cluster intra-subject feature embeddings and separate inter-subject embeddings. We tested SASE on two existing benchmark datasets (ear images without age labels and face images with age labels), as well as our EICZA (ear images with age labels).

In summary, this paper contributes to the existing literature as follows:

- We curated the ear recognition dataset EICZA. It contains 3,330 images of 177 infant participants, collected from infants in hospitals in Lusaka, Zambia. The ear images of the same subject are captured at different ages, ranging from 6 days to 9 months old.
- We propose the task of age-constrained ear recognition – recognizing a person who has aged since the model was trained. To address this challenging task, the model must learn about the concept of aging having only seen younger versions of the query person's ear.
- We evaluated four existing baseline models on EICZA for the traditional age-neutral task of recognizing ears, as well as the proposed age-constrained task. The results show that the age-constrained task is extremely difficult – recognition accuracy drops drastically, e.g., $>50\%$ points (pp), when a model tries to identify a grown ear at a later age.
- We propose an ear recognition model called SASE that takes aging into account. In addition to adopting a representation-specific loss function, i.e., the large margin cosine loss [54], SASE takes a sequence of ear images at different ages as input and adopts a Transformer encoder [51] to obtain context information from ear features of earlier ages to interpret the query ear image. The experimental results show that our proposed method significantly outperforms the baseline methods in the constrained-age task setting.

The curated dataset and our code are publicly available at **https://github.com/wdqin/eicza**, allowing further analysis of the dataset and the development of ear recognition models.

2. Related Work

The idea of using the ear as a biometric is old [8]. We here discuss the main datasets and models that have been used for ear recognition research and also describe face recognition work that addresses the subjects' ages.

2.1. Ear Recognition Datasets

The earliest publicly available recognition datasets consist of ear photos captured under laboratory conditions, with well-controlled views of the ears and lighting, e.g., IITD [31], IITK [38], UND [57], and AMI [48]. Under these conditions, recognition systems based on hand-crafted features worked well, and so the research focus moved to address ear recognition "in the wild," that is, under more difficult unconstrained conditions. Datasets such as UBEAR, AWE, USTB-Helloear, EarVN, and UERC [40, 18, 60, 26, 19] were published accordingly. Some of these are large-scale collections of ear images [19, 26, 60], which enable the training of deep learning models, now frequently adopted for recognition tasks. We use the newest dataset, UERC [19], as a benchmark for our experiments.

2.2. Ear Recognition Models

Early ear recognition models relied on hand-crafted features for ear identification and verification [1, 29, 53, 37, 11, 23, 42], e.g., Adaboost with the Haar feature [1]. When tested on images collected under laboratory conditions, these methods reportedly reach accuracy levels as high as 95% [1, 23, 37, 53]. However, when challenged with ear images captured "in the wild," with noise and variations in illumination, imaging angle, and size, these models suffer great performance drops [18]. Deep learning has been employed to address the task of unconstrained ear recognition, initially with basic convolutional neural networks (CNNs) [12, 41, 50, 24, 2]. Methodological contributions to ear recognition include a U-Net-like encoder-decoder structure [16], an ensemble of different CNNs to detect the ear in the image [20], a comparison of various image feature extraction models [5, 17], the use of a center loss function [59], and a process to concatenate image features from both local patches of the ear images and the whole image [47]. Recent work proposed a vision transformer for ear recognition [4] and showed the benefits of pre-training [43].

2.3. Face Recognition in the Presence of Aging

As for ear recognition, early face recognition systems, which relied on handcrafted features [3, 34], have been replaced by deep learning models (e.g., [36, 45, 49, 55]) with the emergence of large face datasets [7, 9, 22]. Various works proposed new ways of representing faces [14, 15, 35, 39, 54, 61]. We found the latent space representation with the large margin cosine loss [54] particularly useful for our work. Most relevant to our research is the literature on age-invariant face recognition [6, 13, 21, 33, 56, 58], in particular, the use of sequential face inputs [56].

3. The EICZA Dataset

To date, the computer vision community has not worked with ear image datasets that have age labels of the same subjects, although an unnamed, not-much-explored dataset exists, i.e., https://doi.org/10.5281/zenodo.5676103. We curated this raw dataset by making the images suitable for processing with deep learning architectures and organizing them in training/validation/testing batches that are appropriate for addressing the age-constrained recognition task. We name this curated dataset EICZA for Ears of Infant Cohort in Zambia with Aging. EICZA consists of the ear images of infants who were seen at an urban clinic, the Chawama Clinic, in Lusaka, Zambia, during routine well-child visits [46], which occurred at ages 6 days, 6, 10, and 14 weeks, as well as monthly from age 4 to 9 months (Fig. 3). A total of 227 infants were enrolled in this longitudinal study after consent for the release of the images was taken from the participants' next of kin. Consent forms were presented in two local languages and approved by the two universities involved in the study. None of the participants took part in all 10 imaging sessions. Attrition was notable for sessions that did not co-occur with vaccination visits (after 14W).

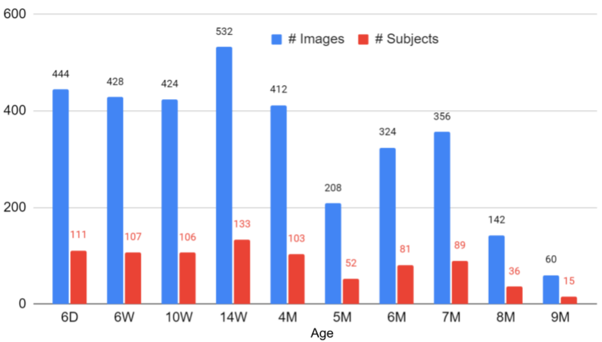

Figure 3. Histogram of the number of imaged infants (red) and the number of images (blue) per session. Each infant was imaged during at most 10 sessions from age 6 days (D) to age 9 months (M). Infants were enrolled in the longitudinal study either at age 6D or age 14 weeks (W).

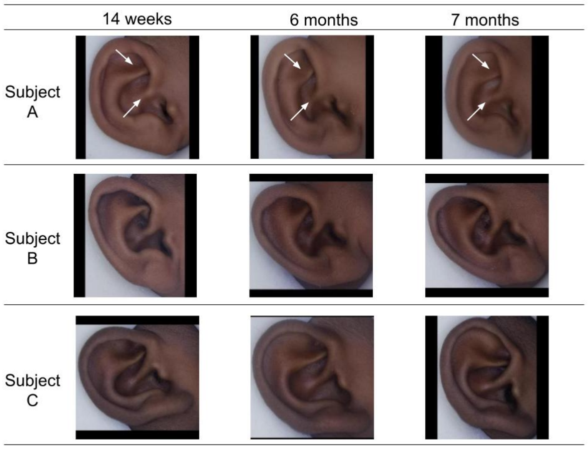

Figure 4. Examples of curated ear images from three subjects at different ages. The white arrows point to ear areas that changed significantly as subject A aged.

For most participants, four ear images, two right and two left, were taken at each imaging session. In most sessions, additional images with the ear rotated by a 30° angle were taken. The images were taken with a white background under appropriate lighting. As a step in our curation process, we cropped and resized the raw images to a resolution of 224×224 pixels for ease of image feature extraction. Processed sample images of three subjects are shown in Fig. 4. Ear growth was significant until 6 months of age [46] (compare the images in the first two columns of Fig. 4).

Table 1. Properties of our EICZA dataset compared to the existing ear datasets (without age labels) and three face datasets with age information. EICZA includes more subjects and twice as many images as the first published face dataset with age labels FG-NET.

	Bio- metric	# of subjects	# of images	Images per subject	Age label
IITD-2 [31]	Ear	221	793	3.59	
UND [57]	Ear	952	3,480	3.66	
UBEAR [40]	Ear	126	4,430	35.16	
UERC 2019 [19]	Ear	3,706	11,804	3.19	
EarVN 1.0 [26]	Ear	164	28,412	173.24	
FG-NET [32]	Face	82	1,002	12.22	✓
MORPH-II [44]	Face	13,000	55,134	4.24	✓
CACD [10]	Face	2,000	163,446	81.72	✓
EICZA (our)	**Ear**	**177**	**3,330**	**18.81**	✓

The original study [46] collected 3,544 ear images of 227 infants. For our task of age-constrained ear recognition, we need ear images of the same participant, taken at 3 different ages at least. This constraint reduced our dataset to 3,330 images of 177 infants for whom at least $\sim 3 \times 4 = 12$ images (4 photos each session) had been collected. For these 177 infants, an average of 18.8 images are available in total, and 4.7 images at a particular age. The images in the original dataset were labeled by patient identity and age. To curate the dataset for training and testing of recognition models in a neutral or age-constrained manner, we organized the images into training, validation, and testing sets based on the participant's identity and the index of the session that a participant **actually** attended (i.e., came back to the clinic and took part in the ear collection at the required age). In particular, the set of images taken at the kth imaging session of participant p is the set $S_k^{(p)} = \{I_{k,1}^{(p)}, I_{k,2}^{(p)}, \ldots, I_{k,u}^{(p)}\}$, where the maximum number u of images is typically 4 and the maximum session index k ranges from 3 to 8 sessions per patient. Section 5 describes how we split the per-patient and per-session sets $S_k^{(p)}$ of images into training, validation, and testing data splits.

Table 1 provides a summary of the statistics of the curated EICZA dataset and other datasets. We list the existing ear datasets as well as face datasets with age labels. Among the face datasets, we selected FG-NET [32] for our experiments, since it is publicly available and does not have label noise. We conducted both age-neutral and age-constrained experiments with it. Among the ear datasets, we chose the most recently published dataset, UERC 2019 [19], which includes the largest number of subjects among ear datasets, for our experiments. Given its lack of age labels, we can only conduct age-neutral experiments with it.

4. Proposed SASE Model

The motivation of our model design was to train a subject-discriminatory representation that utilizes previously collected ear images of the same subject, i.e., ears at younger ages. Accordingly, Our proposed model, SASE for Self-Attention-based Sequential Ear image analysis, consists of three parts, a feature-extracting backbone network, which adopts the Squeeze-and-Excitation Network (SENet) [27], a Transformer encoder [52], which allows the extracted feature to obtain context information from ears of early ages, and a per-subject representation by "center vectors," computed using the Large Margin Cosine Loss (LMCL). A summary of our proposed SASE model is visualized in Figure 5.

4.1. Sequential Input and Transformer Encoder

The task of the SASE model is to identify a person p by a sequence of l ear images of the same person. Concretely, our model is a classification model that, given an input sequence of images $[I_{k-l+1}^{(p)}, ..., I_{k-1}^{(p)}, I_k^{(p)}]$, predicts a person's identity $\hat{p}$ with probability $\Pr(\hat{p}|[I_{k-l+1}^{(p)}, ..., I_{k-1}^{(p)}, I_k^{(p)}])$. Since the input images to the model are always from the same person, it is convenient to drop the superscript (p) and re-write $I^{(p)}$ as I. Specifically, I_k is an ear image taken during session S_k.

During training, I_{k-i} is an ear image randomly chosen from images of session S_{k-i} of a younger age $k-i$ where i ranges from 1 to $l-1$. If $(k-i) < 1$, we sample images from k instead. Practically, during inference, the model is not able to access images of the query subject at a younger age, so we sample images from S_k instead during the evaluation process. In our experiments, we chose $l = 3$, i.e., a sequence of three images as input.

The backbone is a Squeeze-and-Excitation Network (SENet) [27] extractor which extracts the image features from the image sequence:

$$[f_{I_{k-l+1}}, ..., f_{I_{k-1}}, f_{I_k}] = \text{Backbone}([I_{k-l+1}, ..., I_{k-1}, I_k]). \tag{1}$$

For convenience, we re-write f_{I_k} as f_k. $f_k \in \mathbb{R}^{d_f}$, where the dimension d_f is 2,048 for a SENet. In practice, we first pre-trained the SENet with the VGG-Face v2 dataset [9] and then fine-tune the pre-trained backbone with EICZA.

The transformer encoder module takes over the feature sequence $[f_{k-l+1}, ..., f_k]$ and outputs the self-attended feature f'_k from sequence $[f'_{k-l+1}, ..., f'_k]$. Unlike f_k, which only contains information about itself, f'_k also encodes context from the other images in the input sequence of the same person, i.e., from $[f_{k-l+1}, ..., f_{k-1}]$.

There are two main parts of the transformer encoder, the self-attention layer, and the feed-forward network layer. The self-attention layer computes the f'_k with a "scaled dot-product attention" mechanism:

$$Q = FW_Q, \;\; K = FW_K, \;\; V = FW_V, \;\; \text{and}$$

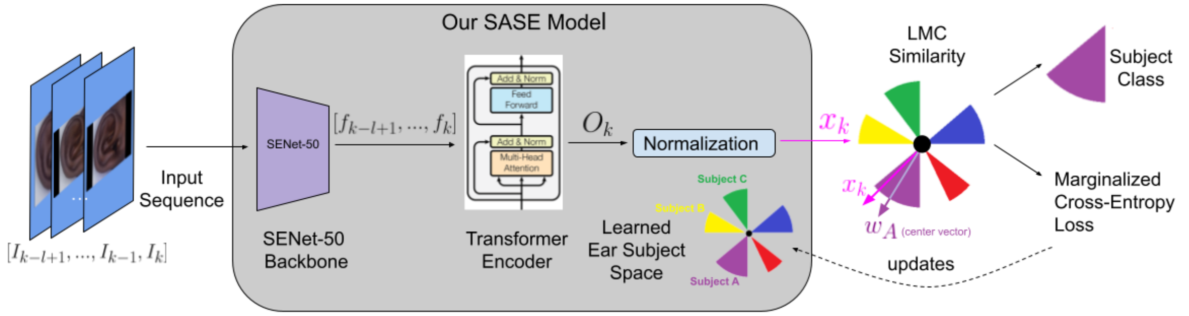

Figure 5. A visualization of our SASE model. Taking an input sequence of l ear images $[I_{k-l+1}$, the model extracts their image features $[f_{k-l+1}, \ldots, f_k]$ with a backbone network and sends them through a Transformer encoder to compute an output feature O_k representing the input ear. The identity of the subject is predicted by comparing the cosine similarities between each element w of a set of learnable center vectors and the normalized output feature representation x_k, and choosing the most similar ("LMC similarities"). During training, a marginalized cross-entropy loss is computed to fine-tune the backbone, Transformer encoder, and ear subject space (the color wheel).

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V,$$

where $F \in \mathbb{R}^{(l+1)\times d_f}$ is the concatenated matrix form of $[f_{k-l+1}, ..., f_k]$, $W_Q, W_K, W_V \in \mathbb{R}^{(d_f)\times d_k}$ are learnable matrices, and $d_k = 256$.

The transformer encoder also adopts a multi-head attention structure, which computes a different Attention(Q, K, V) at the same time:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_n)W_O,$$

where head$_i$ = Attention(Q_i, K_i, V_i) that $i \leq n$. $W_O \in \mathbb{R}^{nd_k \times d_f}$, $n = 8$ in our case.

The computed MultiHead(Q, K, V) is sent to a feed-forward network which can be considered as a neural network of fully-connected layers. The output is $F' = [f'_{k-l+1}, ..., f'_k]$. Additionally, the self-attention layer and the feed-forward network can stack up with each other. Our stack size is two.

We note that techniques such as layer normalization and residual connection are also applied in the transformer encoder [52].

By sending the f'_k through two additional fully-connected layers, we obtain O_k that encodes information from images of different ages. We normalize O_k by dividing by its norm and obtain $x_k = \frac{O_k}{\|O_k\|}$, which allows us to compute the Large Margin Cosine Loss (LMCL) function [54], described next.

4.2. Large Margin Cosine Loss (LMCL)

The Large Margin Cosine Loss [54] can be considered as a special form of cross-entropy loss between the predictions and the target classes. A visualization of the latent ear space trained with LMCL is given in Figure 5. Each feature in the space can be considered as a unit vector with its own direction (cosine value). So whether a feature belongs to a certain subject class (the colored fan in the figure) is decided by how small the angle θ is between the direction of feature vector x_k and the direction of the class center unit vector w. For classification, a set of parameters, the "centers" w's, are learned to represent each class (subject) during the process. During inference, the cosine similarities are computed between the x_k and w's to decide to which subject p feature vector x_k belongs. The goal of the optimizer is to maximize the cosine similarity between x_k and the center of our target center w_p. A hyperparameter M controls the size of margins (gaps between colored fans in the figure) between different class areas, and hyperparameter s controls the magnitude of feature vectors. The Large Margin Cosine Loss is defined as follows

$$\ell_{\text{LMC}} = \frac{1}{N}\sum_{i}^{N} -\log \frac{e^{s(\cos(\theta_{y_i}, i) - M)}}{e^{s(\cos(\theta_{y_i}, i) - M)} + \sum_{p \neq y_i} e^{s\cos(\theta_{p,i})}}, \tag{2}$$

where N is the number of training samples, i stands for the ith sample, y_i stands for the ground truth subject of the sample, $\cos(\theta_p, i) = w_p^T x_i$ (with $\|w^T\| = \|x_i\| = 1$), and the learnable vector w_p of the pth subject, which works as a "center" to represent subject p in latent "ear space."

5. Proposed Evaluation Methodology

Cross-validation Setup. We used 4-fold cross validation in the experiments with the EICZA dataset. In each fold, the ear image data are divided into three disjoint sets for training, validation, and testing. The model is first trained with the images in the training set. Training is then continued on the validation data while regularly storing "snapshots" of the learned parameters. Among these snapshots, the snapshot with the best performance on the validation set is selected as the best model for the current fold. The chosen snapshot model is then tested on the testing data. Once the four folds have been processed, we report the average accuracy of the four best models on the

Table 2. **Cross-Validation Setup for Age Constrained Recognition Experiments.** For 2 of 4 folds the sets of images included in training, validation, and testing are shown, where $S_k^{(p)}$ is the set of images of person p obtained at session k (age increases by session). In each fold, all images are used for training (green), except the images taken during sessions 2–n for a group of m persons (for fold 1, the first m persons; for fold 2, the second m, etc). The newest images are used for testing.

Fold	Training Set	Validation Set	Testing Set
1	$S_1^{(1)}, \ldots, S_1^{(m)}$	$S_2^{(1)}, \ldots, S_2^{(m)}$	$S_3^{(1)}, \ldots S_n^{(m)}$
	$S_1^{(m+1)}, \ldots, S_1^{(2m)}$	$S_2^{(m+1)}, \ldots, S_2^{(2m)}$	$S_3^{(m+1)}, \ldots, S_n^{(2m)}$
	$S_1^{(2m+1)}, \ldots, S_1^{(3m)}$	$S_2^{(2m+1)}, \ldots, S_2^{(3m)}$	$S_3^{(2m+1)}, \ldots, S_n^{(3m)}$
	$S_1^{(3m+1)}, \ldots, S_1^{(4m)}$	$S_2^{(2m+1)}, \ldots, S_2^{(4m)}$	$S_3^{(2m+1)}, \ldots, S_n^{(4m)}$
2	$S_1^{(1)}, \ldots S_1^{(m)}$	$S_2^{(1)}, \ldots S_2^{(m)}$	$S_3^{(1)}, \ldots S_n^{(m)}$
	$S_1^{(m+1)}, \ldots S_1^{(2m)}$	$S_2^{(m+1)}, \ldots S_2^{(2m)}$	$S_3^{(m+1)}, \ldots S_n^{(2m)}$
	$S_1^{(2m+1)}, \ldots S_1^{(3m)}$	$S_2^{(2m+1)}, \ldots S_2^{(3m)}$	$S_3^{(2m+1)}, \ldots S_n^{(3m)}$
	$S_1^{(3m+1)}, \ldots, S_1^{(4m)}$	$S_2^{(3m+1)}, \ldots, S_2^{(4m)}$	$S_3^{(3m+1)}, \ldots, S_n^{(4m)}$

fold-specific testing sets.

To ensure that, during testing, the model is not queried with an ear image of a previously unseen person, we organize our cross-validation scheme by subjects. For a population of $4m$ subjects, each cross-validation fold has a testing set that contains the images of a different group of m subjects. We ensure that images of these subjects are also included in the training and validation sets of the fold. To enhance the training set, we allow images of subjects outside the group of m to be included.

For the task of age-neutral recognition, we can randomly select a person's images into training, validation, and testing sets *irrespective of age labels*. For the task of age constrained recognition, however, we need to devise a data split procedure that takes the age labels into account, as described next.

Setup for Age Constrained Recognition Experiments. Our cross validation procedure groups images by age. For ease of description, we refer to an imaging session rather than the specific age of the person at that session. In EICZA, we have $S_1, \ldots, S_n$ sessions, where the number n of sessions varies among infants (n is at least 3 and at most 8). For the scenario that the ear is used as a biometric for access to health care records, the use case would be to train the model with images taken when the subject is younger ("gallery images" in biometrics) and to query the model with an image when the subject is older ("probe" in biometrics). Thus, we set up our testing set to include the images of more recent sessions $S_3, \ldots, S_n$ and reserve a subject's images in the first session S_1 for training and in the second session S_2 for hyperparameter validation. (For other use cases, the procedure could be changed so that the younger ears are in the query images and the older used for training.)

Table 2 illustrates the dataset split for the 4-fold cross validation experiments with our age constrained recognition model for the first two folds. In the first fold, the images of the third and subsequent imaging session of the first m subjects make up the testing set (blue), and the images of the second session of these m subjects make up the validation set (red). The images in the first session are included in the training set, as well as all data from all sessions of the remaining $3m$ subjects (green). Including the latter is done to help the model learn the general concept of aging, including the former is done so the model learns how the ears of the m subjects look initially.

For the age-neutral experiments, we directly use one group for testing, one group for validation, and the rest for training.

6. Setup of Experiments

Dataset Experiments. To process the EICZA data, we used the procedure described in Section 5 with 177 individuals. Consequently, the ears of 43 or 44 subjects are evaluated during validation and testing for each fold. The number n of sessions to be at least 3 and at most 8, and a [Training:Validation:Testing] split of [2755:248:327] images on average, corresponding to a [83%:7%:10%] of the total number of images, 3330. During experiments, we noticed that images of 6-day ears cause large intra-subject differences compared to the rest of the images, leading to a significant drop in accuracy. Thus, we also trained and tested the ear models without including the 6-day ear data and report their performance.

We experimented with the UERC dataset [19], which contains 11,804 ear images of 3,706 subjects. The main part of this dataset is a subset of the Extended Annotated Web Ears (AWEx) dataset [18], which contains 3,300 ear images of 330 subjects. We used a subset of the UERC 2019 public dataset, including all subjects that have at least 10 ear images, which is a total of 442 subjects and 6,218 images. Since the UERC data do not include age labels, we followed the age-neutral evaluation procedure as described in Section 5. Because the dataset is larger than EICZA, we were able to use five cross-validation folds. When training

SASE on UERC, which requires a sequence of input images, we simply input copies of the same image.

We also experimented with the FG-NET dataset [32], which contains face images of 82 subjects aged from 0 to 69 years old. Due to the limited number of images for certain subjects, we did not apply cross-validation. However, we did apply the age-neutral and age-constrained evaluation procedures in a "single fold experiment."

Model Comparisons. In addition to SASE, we experimented with three feature extraction models, SqueezeNet [28], ResNet-50 [25], and SENet (version SEResNet-50) [27].

Input Preprocessing. All images are resized and padded to 224 × 224 pixels as input of the models. During training, with 0.5 probability, each input image is either processed in this form or in its horizontally-flipped form. Similarly, with equal probability, the input image is rotated by a degree within the range of $0° - 30°$. This preprocessing method ensures a richer data augmentation.

Hyper-parameter Tuning. We used a subset of images to determine the learning rate and number of times all training data should be processed during backpropagation without incurring overfitting. We determined a learning rate of 10^{-5} for models based on ResNet and SENet, a learning rate of 5×10^{-5} for SqueezeNet-based models, and the number of epochs in all experiments to be 200. The hyperparameters s and M are set to 64 and 0.35, respectively.

Network Initialization. We initialized the ResNet and SENet feature extractor parameters by pre-training the models first with the MS1M dataset [22], and then with the VGG-Face2 dataset [9]. We initialized SqueezeNet with the parameters pre-trained on ImageNet.

Snapshot Retrieval. We used a batch size of 16 for models with time series input and 32 for models with single image input. Each iteration during training processes a batch. When training on the validation data, after every 200 iterations, a snapshot is stored. This yields approximately 400 snapshots per cross-validation fold, from which the best-performing model (per fold) is determined.

Computing Environment. We implemented our experiment in Python 3.6 with Pytorch 1.10 and adopted Adam [30] as the optimizer. We used a Quadro RTX 6000 graphics card. Each 4-fold cross-validation experiment takes ∼8,000 MB memory and ∼20 hours (∼ 5 h per fold).

7. Results

Main Results on EICZA. Our experimental results show that our proposed model, SASE, when applied to EICZA outperforms baseline models by large margins, see Table 3. This applies to the three experimental conditions of age-neutral and age-constrained, trained with and without Day 6 images, i.e., an average accuracy of 69%, 33%, and 50%, respectively, and uses the cross-validation procedure described in Section 5 and experimental setup in Section 6. We note that the standard deviation across cross-validation folds is about 6% points, on average. We also conducted an experiment that shows that the smaller the age difference between train and probe images is, the higher the model performance becomes. In particular, SASE's age-constrained performance on EICZA improves to 43% (with Day 6) and 58% (without Day 6) if train and test images were from consecutive imaging sessions (if two sessions apart, 37% (w D6) vs. 52% (w/o D6). Among the three deep models SqueezeNet, ResNet-50, and SENet, the fine-tuned SENet performs the best (i.e., 28% on age-constrained testing without Day 6 images in the training set), which justifies our choice of SENet as the backbone feature extractor in SASE.

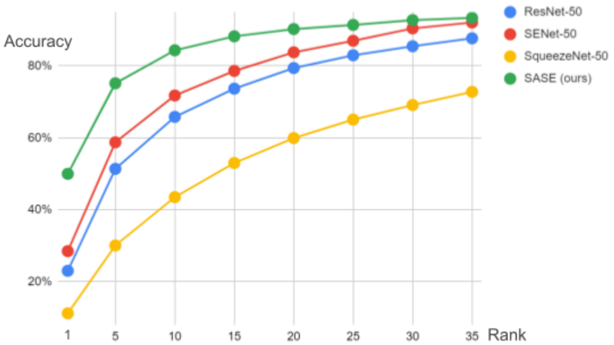

Figure 6. Cumulative Match Characteristic (CMC) curves for the four tested ear recognition models. The proposed SASE model performs consistently the highest accuracy for ranks 1 through 35. The performance is computed for EICZA data without the age 6D images included in the training set.

Model Comparison with CMC. We computed Cumulative Match Characteristic (CMC) curves for the four tested ear recognition models from rank-1 to rank-35 accuracy, see Fig. 6. The plot shows that our proposed SASE model consistently performs with the highest accuracy for these ranks. We note the large jump in accuracy levels from rank 1 (50%) to rank 5 (75%) and rank 10 (84%).

Results on Other Datasets. Since EICZA is the first dataset of its kind and there are no other datasets of ear images with age labels, we cannot apply SASE on other datasets for benchmark comparisons. Nonetheless, we tested SASE on UERC [19] and FG-NET [32] in the age-neutral settings, for which SASE outperforms the baseline models, and on FG-NET in the age-constrained setting, for which it underperforms by 3% points (but outperforms the fine-tuned SENet model, its own backbone, by 7% points). We must stress that our performance levels are lower than what has been reported in the literature for these datasets for state-of-the-art models, but this fact is beside the point we are making here. Our experimental results apply to the age-constrained recognition task we define in this paper, and

Table 3. Average cross-validation recognition accuracy of SASE compared to four baseline models on three datasets

Dataset	UERC [19]	FG-NET [32] (Aging Faces)		Our EICZA (Aging Ears)		
	without	Age Neutral	Age Constrained	Age Neutral	Age Constrained Train/Test	
Model	Ear Ages	Train/Test	Train/Test	Train/Test	with Day 6	without Day 6
SqueezeNet [28]	26.88%	17.85%	7.24%	52.30 %	8.23%	11.14%
ResNet-50 [25]	36.72%	82.84%	**55.92%**	61.30%	13.84%	22.98%
SENet [27]	41.86%	78.89%	46.05%	68.11 %	18.85%	28.46%
SASE (Our Model)	**42.56%**	**82.90%**	52.96%	**69.49%**	**33.14%**	**49.98%**

Table 4. Ablation study results for our SASE model, removing both LMCL (L) and sequential input (S) or one at a time, with (w) and without (w/o) Day 6 training. The percentage point improvement (Impr.) is computed over the baseline SENet (SASE w/o L&S).

Model SASE	Accuracy w Day 6	Impr. [pp]	Accuracy w/o Day 6	Impr. [pp]
w/o L&S	18.9%		28.5%	
w/o L	25.5%	6.7	41.6%	13.1
w/o S	28.1%	9.2	44.0%	15.6
w L&S	33.1%	14.3	50.0%	21.5

the task requires a different experimental methodology than what previous works have used. Our purpose in reporting the numbers in Table 3 is to illustrate differences in task difficulty and baseline models.

Ablation Study of the Proposed Model SASE. In order to show that the design of using sequential inputs and the LMCL-based embedding space are both effective components of our SASE model, we conducted an ablation study that removed either component and then tested performance on the EICZA dataset (Table 4). Without sequential inputs and LMCL, our proposed model is degraded into a SENet feature extractor and fully-connected layers for classification output (row 1). As we can see from rows 2 and 3, use of sequential input and LMCL increases the model performance by 6.7% points and 9.2% points, respectively, when trained on Day 6 data. When the two designs are combined together, our model achieves average accuracy of 33.1%, which is even higher than only using either of them. When we conduct the same ablation experiments on SASE trained without day 6 images, the improvements over the baseline model are even more significant.

Length of Input Sequence. We experimented with different input sequence lengths l on EICZA with and without Day 6 images. We tested $l = 1$ to 5 (Table 5). Length $l = 1$ yields the lowest and $l = 3$ yields the highest accuracy for the testing set for both dataset versions, validating our approach to use sequences of three input images, rather than single images. For the validation sets, we found differences in which length provides highest average accuracy levels ($l = 5$ vs. 3), indicating that length 3 may not necessarily yield the best results in other settings. However, the benefit of processing sequential image input likely generalizes.

Table 5. Accuracy of SASE with different length l on EICZA.

SASE	with Day 6		w/o Day 6	
	Val.	Test	Val.	Test
$l = 1$	57.47%	30.40%	65.97%	44.48%
$l = 2$	62.70%	32.87%	70.19%	45.72%
$l = 3$	56.64 %	**33.14%**	**71.47%**	**49.98%**
$l = 4$	55.63%	26.51 %	71.03%	42.79%
$l = 5$	**62.87%**	32.83 %	69.77%	43.09%

Table 6. The result of the outsider experiment for model SASE with the training set that includes the challenging 6-day data.

Thresh-	Insider	Outsider Detection			
old T	Accuracy	Accuracy	Precision	Recall	F-1
0.2	30.25%	50.16%	50.08%	99.84%	66.70%
0.3	**36.40%**	**54.91%**	**53.59%**	**73.35%**	**61.93%**
0.4	**55.49%**	**54.68%**	**60.42%**	**27.12%**	**37.43%**
0.5	70.69%	51.57%	67.44%	6.06%	11.12%
0.6	90.91%	50.13%	64.71%	0.57%	1.14%

Outsider Experiment. In each fold of the 4-fold cross-validation, we selected 1/4 of the subjects as “outsiders” and randomly sampled ear images from them for an amount equal to the original testing images. If the highest score for a subject predicted by the model is lower than the threshold T, the model predicts the input ear as an ear of a previously-unseen “outsider.” The results in Table 6 show there is a trade-off between high insider and outsider detection. The “sweet spot” for T is between 0.3 and 0.4.

8. Discussion and Conclusion

Our work has shown that ear aging adds a substantial challenge to the problem of ear recognition. While the performance margins of SASE are large compared to the baseline models (20 percent points and more), the absolute accuracy level of the best model (50%) is too low yet for SASE to serve as a biometric method. This encourages future work by the IJCB community. We suggest exploring how to handle the issue of missed sessions, maybe working with the age labels directly rather than the session index k.

We suggest that the curated dataset be used for research on ear-based identity verification for access to electronic health records of infants. After all, being able to match a child to its immunization records was one of the goals that motivated the ear image data collection in Zambia.

References

[1] A. Abaza, C. Hebert, and M. A. F. Harrison. Fast learning ear detection for real-time surveillance. In *2010 fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2010.

[2] R. Ahila Priyadharshini, S. Arivazhagan, and M. Arun. A deep learning approach for person identification using ear biometrics. *Applied intelligence*, 51(4):2161–2172, 2021.

[3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[4] M. B. Alejo. Unconstrained ear recognition using transformers. *Jordanian Journal of Computers and Information Technology*, 7(4):326–336, Dec. 2021.

[5] H. Alshazly, C. Linse, E. Barth, and T. Martinetz. Deep convolutional neural networks for unconstrained ear recognition. *IEEE Access*, 8:170295–170310, 2020.

[6] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[7] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 464–473. IEEE, 2017.

[8] A. Bertillon. *La photographie judiciaire: avec un appendice sur la classification et l'identification anthropométriques*. París: Gauthier-Villars, 1890.

[9] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[10] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 768–783. Springer, 2014.

[11] P. Chidananda, P. Srinivas, K. Manikantan, and S. Ramachandran. Entropy-cum-Hough-transform-based ear detection using ellipsoid particle swarm optimization. *Machine Vision and Applications*, 26(2):185–203, 2015.

[12] C. Cintas, M. Quinto-Sánchez, V. Acuña, C. Paschetta, S. De Azevedo, C. Cesar Silva de Cerqueira, V. Ramallo, C. Gallo, G. Poletti, M. C. Bortolini, et al. Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks. *IET Biometrics*, 6(3):211–223, 2017.

[13] D. Deb, D. Aggarwal, and A. K. Jain. Identifying missing children: Face age-progression via deep feature aging. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10540–10547. IEEE, 2021.

[14] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[15] J. Deng, Y. Zhou, and S. Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 60–68, 2017.

[16] Ž. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer. Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation. *IET Biometrics*, 7(3):175–184, 2018.

[17] Ž. Emeršič, D. Štepec, V. Štruc, and P. Peer. Training convolutional neural networks with limited training data for ear recognition in the wild. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 987–994.

[18] Ž. Emeršič, V. Štruc, and P. Peer. Ear recognition: More than a survey. *Neurocomputing*, 255:26–39, 2017.

[19] Ž. Emeršič, A. K. SV, B. Harish, W. Gutfeter, J. Khiarak, A. Pacut, E. Hansley, M. P. Segundo, S. Sarkar, H. Park, et al. The unconstrained ear recognition challenge 2019. In *2019 International Conference on Biometrics (ICB)*, pages 1–15. IEEE, 2019.

[20] I. I. Ganapathi, S. Prakash, I. R. Dave, and S. Bakshi. Unconstrained ear detection using ensemble-based convolutional neural network model. *Concurrency and Computation: Practice and Experience*, 32(1):e5197, 2020.

[21] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2872–2879, 2013.

[22] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.

[23] A. Halawani and H. Li. Human ear localization: A template-based approach. In *Proceedings of the International Workshop on Pattern Recognition (ICOPR 2015), Dubai, UAE*, pages 4–5, 2015.

[24] E. E. Hansley, M. P. Segundo, and S. Sarkar. Employing fusion of learned and handcrafted features for unconstrained ear recognition. *IET Biometrics*, 7(3):215–223, 2018.

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[26] V. T. Hoang. EarVN1. 0: A new large-scale ear images dataset in the wild. *Data in brief*, 27, 2019.

[27] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[28] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[29] K. Joshi and N. Chauhan. Edge detection and template matching approaches for human ear detection. In *International Conference on Intelligent Systems and Data Processing (ICISD)*, pages 50–55. Citeseer, 2011.

[30] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] A. Kumar and C. Wu. Automated human identification using ear imaging. *Pattern Recognition*, 45(3):956–968, 2012.

[32] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[33] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *IEEE Transactions on Information Forensics and Security*, 6(3):1028–1037, 2011.

[34] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 11(4):467–476, 2002.

[35] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[36] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

[37] A. Pflug, A. Winterstein, and C. Busch. Robust localization of ears by feature level fusion and context information. In *2013 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2013.

[38] S. Prakash and P. Gupta. An efficient ear localization technique. *Image and Vision Computing*, 30(1):38–50, 2012.

[39] X. Qi and L. Zhang. Face recognition via centralized coordinate learning. *arXiv preprint arXiv:1801.05678*, 2018.

[40] R. Raposo, E. Hoyle, A. Peixinho, and H. Proença. Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions. In *2011 IEEE workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pages 84–90. IEEE, 2011.

[41] W. Raveane, P. L. Galdámez, and M. A. González Arrieta. Ear detection and localization with convolutional neural networks in natural images and videos. *Processes*, 7(7):457, 2019.

[42] K. Resmi and G. Raju. A novel approach to automatic ear detection using banana wavelets and circular Hough transform. In *2019 International Conference on Data Science and Communication (IconDSC)*, pages 1–5. IEEE, 2019.

[43] K. R. Resmi and G. Raju. Ear Recognition Using Pretrained Convolutional Neural Networks. In M. Singh, V. Tyagi, P. K. Gupta, J. Flusser, T. Ören, and V. R. Sonawane, editors, *Advances in Computing and Data Sciences*, pages 720–728, Cham, 2021. Springer International Publishing.

[44] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006.

[45] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[46] A. Simukanga, M. Kobayashi, L. Etter, W. Qin, R. Pieciak, D. Albuquerque, Y.-J. Chen, M. Betke, W. MacLeod, J. Phiri, et al. The impact of ear growth on identification rates using an ear biometric system in young infants. *Gates Open Research*, 5:179, 2021.

[47] D. Štepec, Ž. Emeršič, P. Peer, and V. Štruc. Constellation-based deep ear recognition. In *Deep biometrics*, pages 161–190. Springer, 2020.

[48] E. G. Sánchez. Análisis biométrico de la orejas, 2008.

[49] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[50] L. Tian and Z. Mu. Ear recognition based on deep convolutional network. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 437–441. IEEE, 2016.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, page 11 pp. Curran Associates, Inc., 2017.

[53] N. K. A. Wahab, E. E. Hemayed, and M. B. Fayek. Heard: An automatic human ear detection technique. In *2012 International Conference on Engineering and Technology (ICET)*, pages 1–7. IEEE, 2012.

[54] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[55] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[56] C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, J. Zhang, Y. Sun, and B. Zheng. Age-invariant face recognition by multi-feature fusionand decomposition with self-attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s):1–18, 2022.

[57] P. Yan and K. W. Bowyer. Biometric recognition using 3d ear shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1297–1308, 2007.

[58] Z. Zhai and J. Zhai. Identity-preserving conditional generative adversarial network. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–5. IEEE, 2018.

[59] Y. Zhang, Z. Mu, L. Yuan, and C. Yu. Ear verification under uncontrolled conditions with convolutional neural networks. *IET Biometrics*, 7(3):185–198, 2018.

[60] Y. Zhang, Z. Mu, L. Yuan, C. Yu, and Q. Liu. USTB-Helloear: A large database of ear images photographed under uncontrolled conditions. In *International Conference on Image and Graphics*, pages 405–416. Springer, 2017.

[61] Y. Zheng, D. K. Pal, and M. Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5097, 2018.