# On Unconstrained Ear Recognition For Privacy-Preserving Authentication

Vishesh Kumar, Akshay Agarwal

IISER Bhopal, India

#### Abstract

Biometrics recognition-based person authentication is an effective medium to stop illegal access to the system; however, the concern of privacy has raised serious issues about its use. In this research, we have conducted an extensive study for privacy-preserving person authentication using ear modality. For that, we have proposed a novel triplet loss-based convolutional network architecture. To demonstrate its effectiveness, we have compared our proposed algorithm with several convolutional networks including vision transformers (ViT). Interestingly, it is also observed that the existing ear recognition datasets majorly cover non-Indian identities. Therefore, to handle the dearth of Indian ethnicity ear recognition datasets and to advance the research concerning Indians, *we have collected the first-ever unconstrained Indian ear recognition dataset namely I-Ear*. On the proposed dataset, the performance of the proposed triplet loss network is at least 5% better than baseline networks including ViT.

#### Keywords

Ear Recognition, Indian Ear Dataset, Privacy-Preserving Authentication

# 1. Introduction

The increasing demand for secure and automated identity systems has led to extensive research in the fields of computer vision and intelligent systems. Biometrics, which offer invariance over time, ease of acquisition, and uniqueness for each individual, have become the preferred choice for human identification systems. Among various biometric characteristics, the human face has received significant attention in research and development. However, facial recognition systems possess several challenges of which privacy is the most critical one. In contrast, the human ear has emerged as a promising biometric modality with unique advantages. The distinct shape and features of the ear, such as the contours, ridges, and earlobes, along with non-intrusive acquisition provide valuable information for identification purposes [1]. Further, ear image recognition can solve the privacy issue to some extent. Ongoing research and development in ear recognition technology aim to exploit its capabilities and improve its performance in various applications, which propelled us to research unconstrained ear datasets [2, 3, 4].

In this research, we have utilized ear datasets that comprise both controlled and uncontrolled environments, including ear images acquired from the web. However, the datasets contain a limited number of samples per subject, making it impractical to build a deep-learning model from

D 0000-0001-7362-4752 (A. Agarwal)

WAMWB'23: Workshop on Advances of Mobile and Wearable Biometric, September 26, 2023, Athens, Greece

Svishesh22@iiserb.ac.in (V. Kumar); akagarwal@iiserb.ac.in (A. Agarwal)

https://sites.google.com/iiitd.ac.in/agarwalakshay/home (A. Agarwal)

<sup>© 2023</sup> Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

scratch. To address this limitation, we have employed transfer learning techniques to achieve person recognition using ear images. In our study, we initially utilized the feature extraction capability of deep networks directly for ear recognition. We employed a non-trainable method that involved extracting the features from the deep networks and then computing dissimilarity metrics to compare and match ear images. This approach allowed us to leverage the discriminative power of the deep networks' learned representations. To improve the recognition rate further, we performed a score fusion of the best-obtained models. Score fusion involves combining the dissimilarity scores from multiple models to generate a final decision score. By fusing the scores, we aimed to capture the complementary information and enhance the overall recognition accuracy.

Later, we proposed a triplet loss model for ear recognition utilizing the best-performing pretrained model. The triplet loss model aims to learn a discriminative embedding space where the distance between ear images of the same identity is minimized and the distance between dissimilar identities' ear images is maximized. On top of that, we also performed decision fusion using the score obtained from the multiple classifiers for the triplet loss model and proposed a robust ear biometric recognition technique. Overall, this research exhibits the following contributions:

- We propose an ear recognition network based on the amalgamation of multiple pre-trained deep networks and a proposed triplet loss model. The pre-trained model brings the computational advantage while performing recognition; whereas, the triplet model accurately captures the ear-specific features using an effective base deep neural network;
- Extensive experiments are conducted to evaluate and compare different learning methodologies in our study. The objective also aims to thoroughly analyze the performance of multiple state-of-the-art deep networks, providing a comprehensive understanding of their effectiveness for ear recognition;
- We also present a novel ear recognition dataset namely 'I-Ear', which will be made publicly available to the research community.

# 2. Related Work

Ear recognition, while gaining attention, is not as widely studied or popular as other biometric modalities such as face, iris, and fingerprint recognition. Consequently, the availability of large-scale datasets specifically designed for ear recognition is still limited. To address the challenge posed by the relatively small size of ear recognition datasets, Emersic et al. [5] proposed a deep learning-based averaging system. This approach aims to mitigate the risk of overfitting, which can arise when training deep learning models on small datasets. Zhang et al. [6] focused on addressing the challenges of ear verification under uncontrolled conditions using convolutional neural networks (CNNs). Zhang et al.[7] focused on addressing the challenges of few-shot learning in the context of ear recognition. Few-shot learning refers to the ability of a model to recognize new classes or categories with only a limited number of training examples. Sinha et al. [8] proposed a framework that localizes the outer ear image using a histogram of oriented gradients (HOG) and support vector machine (SVM) and then uses CNNs to perform ear recognition. Hansley et al. [9] proposed a fusion approach that combines both convolutional neural networks (CNNs) and handcrafted features for ear recognition. They found that this fusion method outperformed

other state-of-the-art CNN-based approaches and concluded that handcrafted features could complement deep learning methods in the context of ear recognition. The research by Štepec et al. [10] contributes to the advancement of ear recognition by introducing a dual-path CNN model that integrates global and local information and incorporates a constellation-based approach for accurate ear images. The framework provides insights into effective strategies for encoding and utilizing different aspects of ear characteristics, leading to improved recognition performance.

# 3. Proposed Ear Recognition Algorithm

This paper delves into the utilization of state-of-the-art deep learning architectures for encoding the distinctive features of ear images. To accomplish this, we conducted a thorough evaluation of various image classification convolutional neural networks (CNNs) and a vision transformer (ViT). Based on the effectiveness of an individual classification network, the top-performing classifiers are then employed for score fusion to enhance the overall recognition accuracy. By selecting the best models and optimizing their computational requirements, we strive to achieve a balance between accuracy and efficiency in ear recognition. To achieve this objective, we explore two distinct recognition algorithms for representation learning, which will be discussed in the subsequent subsections. We aim to identify the most effective representation method, enabling us to build ensemble models that enhance overall recognition performance. The Pseudocode of the proposed ear recognition algorithms is shown in Algorithm 1 and Algorithm 2 and is described separately in different subsections as follows:

## 3.1. Algorithm 1: Transfer Learning

- 1. Split each ear dataset into two sets: (i) gallery and (ii) probe. Match each probe image with the gallery images of different identities for person identification. Following the standard biometrics recognition setting, a single image of each identity is used as a gallery image;
- 2. To compute the biometric representation of the subject, passed each gallery image through Z number of CNN and Transformer. We get a k dimensional feature vector of each subject called a biometric template.
- 3. Repeat step 2 for computing the biometric representation of probe images at the times of testing.
- 4. Compute the Dissimilarity scores between the feature representation obtained using a particular CNN or, Transformer for a probe image and *S* gallery images;
- 5. After comparing the feature representations of different CNNs, we obtain identity dissimilarity scores using different distance metrics. These scores are then fused, employing the average fusion technique, to generate a final decision score vector.
- 6. Once the fusion process is complete, an identity is selected based on the minimum distance value corresponding to the label assigned to the probe image. If the predicted label matches the ground truth label, the correct identification score is incremented by one. After accumulating the correct identification scores, the correct identification rate (CIR)

Algorithm 1: Proposed Ear Recognition Algorithm based on Dissimilarity Metric

**Input:** Gallery Images of S different identities, P probe images of S identities, Z number of CNNs and Transformer (ViT);

Using each CNNs and Transformer, compute the Feature Representation of S gallery images such as  $F_z = \phi_z(S)$ , and we get a matrix of size  $S \times k$  where k is the length of the feature vector for each pre-trained CNN and Transformer;

#### for every $p \in P$ do

Compute the features representation  $f_p$  of p using the pre-trained CNNs and Transformer; for each  $f_z(S) \in F_z$  do Euclidean Dissimilarity;  $\mathcal{D}(f_p, f_z(S)) = \sqrt{\sum(f_p - f_z(S))^2}$ OR, Cosine Dissimilarity;  $\mathcal{D}(f_p, f_z(S)) = 1 - \frac{f_p \cdot f_z(S)}{||f_p||||f_z(S)||}$ For each representation of CNN and Transformer, the aforementioned computation is repeated; Perform the dissimilarity score fusion or, Late fusion;  $\mathcal{D} = \bigoplus(\mathcal{D}(f_p, f_{z_i}(S))), i = 1, 2, 3, \dots, Z$ ID=  $min(\mathcal{D})$ ; end

can be computed as follows

$$CIR\% = \left(\frac{\sum_{predicted==groundtruth}^{1}}{total \ samples}\right) \times 100$$

The CIR is determined by dividing the total number of correctly identified samples by the total number of samples in the dataset. This metric provides insight into the accuracy of the identification process and serves as a measure of the system's performance. By calculating the CIR, we gain valuable information about the system's effectiveness in correctly identifying individuals based on their features.

## 3.2. Algorithm 2: Triplet Ear Recognition

- 1. Split each ear dataset into two sets: (i) gallery and (ii) probe. We have used two images from each identity for gallery images and the remaining images i.e., 2, 2, and 13 images for probe images from I-Ear, AWE, and KinEar datasets, respectively.
- 2. Triplet Selection:
  - For each identity in the dataset, randomly select one image as the anchor.
  - Randomly select another image from the same identity as the positive example.



**Figure 1:** The schematic diagram illustrates the proposed ear recognition algorithm 1, which utilizes a late fusion strategy to combine decision probabilities. These probabilities are derived from the dissimilarity (distance) scores between the probe and gallery images belonging to S distinct identities. k and r represent the dimensions of the feature vector obtained by different networks.

- Create anchor-positive pairs by pairing the anchor image with its corresponding positive image.
- Randomly select an image from a different identity as the negative example. Ensure the negative image is dissimilar to the anchor image to create a challenging contrast.
- Repeat the dataset's anchor-positive pair generation and negative example selection process for multiple identities.
- Store the generated triplets, including the anchor, positive, and negative images, for training your triplet loss model.
- 3. Pass selected batch of triplets through each CNNs and Transformer (ViT) and then use the classifier to predict the probability distribution of each class for each triplet. In intermediate steps between the feature extractor and classifier, we aim to minimize the distance between the anchor and positive image meanwhile maximize the distance between the anchor and negative images.
- 4. Then compute the triplet-loss and cross-entropy loss using the triplet function and crossentropy function defined in the pseudocode.
- 5. Compute the total loss and update the parameter of the feature extractor (CNNs and Transformer).
- 6. Backward pass and update model and classifier parameters:
  - Compute the gradients of the triplet loss and cross-entropy loss concerning the model and classifier parameters, respectively.
  - Use an optimization algorithm (e.g., stochastic gradient descent) to update the model and classifier parameters based on the computed gradients.
- 7. Repeat steps 3-6 for multiple iterations or epochs:



**Figure 2:** A schematic diagram of the proposed ear recognition algorithm 2 that incorporates a late fusion strategy of the decision probabilities. The decision probabilities, in this case, are obtained by passing the probe images from the trained **Triplet loss model**.  $\theta$  represents the Deep Network,  $\theta'_1$  and  $\theta'_2$  represent the updated deep network of the trained model, *k* represents the dimension of the feature vector of the deep network, and *S* represents the total number of different identities.

- Iterate over the training dataset in batches and perform forward and backward passes to update the model parameters.
- 8. Monitor the training progress:
  - Evaluate the model periodically using a validation dataset to track performance and detect overfitting.
  - Adjust hyperparameters (e.g., learning rate, margin) if necessary.
- 9. Once the model is trained, pass the probe images through the model and compute the probability score.
- 10. Probability scores of different feature extractors are fused (average fusion) to generate a final decision score vector.

Algorithm 2: Proposed Ear Recognition Algorithm based on Triplet Loss

**Input:** *M* Gallery Images of *S* different identities, *T* probe images of *S* identities,  $\theta$  represent the feature extractor or, CNNs and Transformer (ViT),  $\psi$  represent the classifier;

Divide Gallery images into Anchor (A), Positive (P), and Negative (N) sets of images. Such that  $A \cup P \cup N = M$  and for each triplet  $A \cap P \cap N = \phi$ ;

Compute the feature representation of A, P and N using the feature extractor  $\theta$  as  $f_{\theta}(A), f_{\theta}(P)$  and  $f_{\theta}(N)$  these are matrices of size  $A \times k, P \times k$  and  $N \times k$  respectively.

Where k is the length of the feature vector for each feature extractor;

Use classifier  $\psi$  to predict the probability belonging to each class as  $f_{\psi}(f_{\theta}(A)), f_{\psi}(f_{\theta}(P))$ and  $f_{\psi}(f_{\theta}(N))$  these all are of size  $1 \times S$ 

for every  $x \in T$  do

for each epoch do **Triplet Loss =** $L_1$ ;  $l_i(f_{\theta}(A), f_{\theta}(P), f_{\theta}(N)) = max \{ \mathcal{D}^+(f_{\theta}(A), f_{\theta}(P)) - \mathcal{D}^-(f_{\theta}(A), f_{\theta}(N)) + \alpha, 0 \}$ Where,  $\mathcal{D}^+(f_{\theta}(A), f_{\theta}(P)) = \sqrt{\sum (f_{\theta}(A) - f_{\theta}(P))^2}$  $\mathcal{D}^-(f_{\theta}(A), f_{\theta}(N)) = \sqrt{\sum (f_{\theta}(A) - f_{\theta}(N))^2}$ and  $\alpha$  is a margin between positive and negative pairs, So;  $L_1 = \sum_{i=1}^n l_i(f_{\theta}(A), f_{\theta}(P), f_{\theta}(N))$  where n is the number of triplets. **Cross Entropy Loss =**  $L_2$ ;  $l_i = \varphi(f(A), f_{\psi}(f_{\theta}(A))) + \varphi(f(P), f_{\psi}(f_{\theta}(P))) + \varphi(f(N), f_{\psi}(f_{\theta}(N)))$ , where f(A), f(P), f(N) are true probability distribution of A, P, N respectively and  $\varphi(f(P), f_{\psi}(f_{\theta}(P))) = -f(P)\log(f_{\psi}(f_{\theta}(P)))$  $L_2 = \sum_{j=1}^n l_j$ Total Loss = L;  $L = L_1 + L_2$ Updating the Parameter;  $\begin{aligned} \theta^{\text{new}} &= \theta^{\text{old}} - l.r \times \frac{\partial L}{\partial \theta} \\ \psi^{\text{new}} &= \psi^{\text{old}} - l.r \times \frac{\partial L}{\partial \psi} \end{aligned}$ Repeat the above computation for each epoch, l.r = learning rate ; end Compute  $f_{\theta^{\text{new}}}(x)$  and  $f_{\psi^{\text{new}}}(f_{\theta^{\text{new}}}(x)))$ Perform the score fusion or, Late fusion;  $P(x) = \bigoplus (f_{\psi_i^{\text{new}}}(f_{\theta_i^{\text{new}}}(x)));$ **ID**=  $argmax\{P(x)\};$ end

11. An identity is selected corresponding to the maximum probability value to label the probe image.

The schematic diagrams of both algorithms are shown in Figure 1 and Figure 2.

# 4. Experimental Setup

This section describes the various components associated with the ear recognition experiments conducted in this research. We begin by providing an overview of the datasets used to perform ear recognition. Moving forward, we discuss the deep neural networks used in our experiments. These networks are chosen based on their proven efficacy in various computer vision tasks including image classification. In the end, we present the analysis concerning different dissimilarity metrics and the Triplet Loss Model used for evaluation.

## 4.1. Datasets

To ensure an unbiased study, we have conducted our experiments using several datasets that are collected in uncontrolled environments. This approach allows us to evaluate the performance of the identification system under real-world conditions, where various factors such as lighting conditions, pose variations, and occlusions can significantly impact recognition accuracy. The datasets used in this research are briefly described below:

- 1. **KinEar dataset:** The KinEar dataset [2] contains ear images of 19 families with a total of 76 identities [11];
- 2. **I-Ear dataset:** As part of our research efforts, we have prepared a first-ever Indian ear recognition dataset captured in several unconstrained settings using mobile devices. In total, the dataset contains ear images of 52 identities. The dataset can be accessed by using the following link [I-Ear].
- 3. **AWE dataset:** Annotated Web Ears (AWE) dataset [3, 12, 4], contains ear images collected from the web, ensuring a wide range of variability derived from unconstrained environments. This dataset comprises images of 100 subjects, including some of the most famous people from diverse ethnicities, genders, and age groups.

The characteristics of each dataset are given in Table 1 and a few samples are shown in Figure 3.

## Table 1

Characteristic of the ear datasets used in this research along with the algorithm used for ear recognition.

Algorithm	Dataset	Subject	Total Images	Gallery	Probe
Dissimilarity Metrics	KinEar	76	1064	76	988
Triplet Loss	KinEar	76	1140	152	988
<b>Dissimilarity Metrics</b>	I-Ear	52	208	52	156
Triplet Loss	I-Ear	52	208	104	104
Triplet Loss	AWE	100	400	200	200

# 4.2. CNNs and ViT

Various convolutional neural networks (CNNs) and a vision Transformer are explored to find an effective architecture for the proposed ear recognition algorithm. These networks encompass a



Figure 3: Samples of images from the I-Ear (first three from left), AWE (three in the middle), and KinEar (last three) datasets.

wide spectrum, including sequentially connected networks like VGG [13], residual connected networks such as ResNet [14], compact CNNs like MobileNet [15], EfficientNet [16], and Xception [17]. Additionally, we have ventured into novel network architecture search models like NASMobileNet [18] for feature extraction. We have also utilized a recent and state-of-the-art deep network model namely vision transformer (ViT) which is currently less explored for ear recognition. Contrary to CNNs, the Vision Transformer (ViT) does not incorporate any convolutional layers in its architecture. Instead, ViT adopts a different approach by splitting the input image into smaller patches and directly feeding them into the ViT network along with the attention module to find the most distinctive features in the local region of images.

To obtain image representations, we extract features from the final dense layers of these models. The feature dimensions (K) for different networks are as follows: (1) VGG16: 4096, (2) EfficientNet: 1280,(3) Xception: 2048, (4) MobileNetV2: 1280, (5) NasNetMobile: 1056, (6) ResNet50: 2048 and, (7) ViT: 768. These feature representations serve as robust biometric descriptors that capture the distinctive characteristics of the input images. To perform the matching for ear recognition, we have used two methods:

- 1. **Non-Trainable Method:** Multiple dissimilarity metrics, namely Cosine and Euclidean are used to measure the distance between the image representation belonging to same and different identities. These metrics offer a non-training property, enabling computational efficiency and fast identification of identities.
- 2. **Trainable Method:** As described in Algorithm 2, in this method, we have trained the triplet loss model for ear recognition.

## 4.3. Evaluation Metrics

To evaluate the performance of our ear identification experiments, we employed various performance curves and metrics. These evaluation metrics are listed below:

- **CIR%:** The CIR score, as defined earlier, represents the rate of correctly classified ear images by the ear recognition systems.
- **ROC:** ROC stands for receiver operating characteristic curve and is useful for person verification scenarios. The verification is a 1:1 matching scenario where along with the ear image a person also provides the identity. The image is then matched with all the images in the gallery and the score corresponding to the true identity is considered as genuine score. If the genuine score is greater than the threshold, we considered a match else discard the ear images.
- CMC: CMC stands for cumulative matching characteristics curve and is used to measure the performance of ear recognition systems at different positions (rank) of matching. If

the true dissimilarity score is found at the first position, then it is referred to as a rank-1 matching. Else we keep checking the different positions until the true identity is found but in turn, this increases the search space.

# 5. Experimental Results and Analysis

In this section, we present the results of ear identification experiments conducted on multiple challenging datasets including the proposed dataset using different models and strategies. The analysis can be divided into multiple categories based on the ingredients used to perform the matching such as (i) analysis based on the performance of a CNN, (ii) analysis on the effectiveness of ViT, (iii) analysis concerning the utilization of non-trainable matching method or impact of training the proposed triplet loss model. The analysis on CNN reflects the effectiveness of different CNN in performing ear recognition, later, these best-performing models are selected as a base model in the triplet loss model. Finally, we used Grad-CAM visualization [19], to understand and explain the working of CNN and ViT implemented for ear recognition. The visualization helps in understanding the importance of different ear regions in identifying individuals.

### Table 2

CIR score range in the interval [0-1] of the different pre-trained deep networks used for **KinEar** biometrics recognition along with the proposed dissimilarity fusion algorithm.

Model	Euclidean	Cosine
VGG16	0.36	0.39
EfficientNet	0.30	0.30
Xception	0.20	0.20
MobileNetV2	0.22	0.22
NasNetMobile	0.10	0.10
ResNet50	0.41	0.44
ViT	0.24	0.25
Proposed	0.41	0.42

## 5.1. Experimental Results and Analysis on KinEar Dataset

We first present the analysis of the KinEar dataset using both the evaluation metrics namely dissimilarity metrics and triplet loss. In this research, we have performed the ear identification, which is a 1 : S matching scenario, where S is the total identities. The results of person identification using the KinEar dataset are reported in Table 2. It is observed that the feature embedding obtained using the network architecture search (NAS) network (NASNet) performs worst on the ear identification task. We want to mention that similar to other CNNs, the NASNet model is also pre-trained on the ImageNet dataset and only used as the image encoder. Out of all the CNNs, the ResNet50 model performs the best for ear recognition on the KinEar dataset. The prime reason might be the deep nature of the architecture in comparison to efficient architectures such as EfficientNet and MobileNetV2. While in most cases, both distance metrics perform the

#### Table 3

CIR score using proposed decision probability fusion algorithm in the triplet loss model on the **KinEar** dataset.

Model	ResNet50	VGG16	EfficientNet	ViT	Proposed
CIR	0.48	0.37	0.52	0.52	0.57

#### Table 4

Ear identification performance on the **I-Ear** dataset in terms of the CIR score [0-1]. Here dissimilarity matching has been used for ear recognition.

Modality	Dissimilarity	VGG16	EfficientNet	Xception	MobileNetV2	NasNetMobile	ResNet50	ViT	Proposed
Ear(Left)	Euclidean	0.69	0.67	0.67	0.68	0.49	0.69	0.73	0.78
	Cosine	0.72	0.69	0.67	0.70	0.50	0.72	0.73	0.78
Ear(Right)	Euclidean	0.58	0.55	0.57	0.62	0.42	0.68	0.67	0.73
	Cosine	0.61	0.57	0.61	0.62	0.40	0.70	0.67	0.75

same, on ResNet embeddings Cosine dissimilarity metric outperforms the Euclidean distance by a significant margin. In other words, cosine dissimilarity improves the ear recognition performance by 3% as compared to the Euclidean metric on the KinEar dataset when the ResNet and VGG16 encoders are used for feature extraction. In the first case of improving the performance of ear recognition, we have performed the average dissimilarity score fusion of a couple of best-performing models. In this research, based on the performance of the models, we have performed the fusion of ResNet50 and VGG16. It is seen that this dissimilarity fusion does not improve the recognition performance for Euclidean and with cosine slight reduction in the accuracy is observed.

In the second attempt of increasing the performance of ear recognition, we have trained the triplet loss model using CNNs and ViT. While training the triplet loss model increases the computational complexity of the ear recognition system but it shows a tremendous improvement in the recognition performance. The results reported in Table 3 show the performance of the best performing ResNet model boost to 48% in comparison to 44% obtained using the cosine dissimilarity metric. In comparison to other models, ViT and EfficientNet show a significant jump in performance. For example, the performance of ViT improves more than 2 times using the triplet model as compared to the dissimilarity metric-based matching. Overall, the proposed fusion using the triplet model improves the best matching performance on the KinEar dataset and is 15% better than the fusion algorithm where an average of dissimilarity scores are used for matching.

## 5.2. Experimental Results and Analysis on the Proposed I-Ear Dataset

We have acquired the first-ever Indian recognition dataset using mobile devices in unconstrained settings. We have captured the ear images of both the left and right ear and hence individually both ears are used for matching. Similar to the experiments on the KinEar dataset, we have performed the ear recognition on the I-Ear dataset using dissimilarity metric and triplet loss models.

#### Table 5

	ResNet50	EfficientNet	VGG16	ViT	Proposed
Ear(Left)	0.82	0.76	0.64	0.75	0.83

Ear identification performance on the **I-Ear** dataset in terms of the CIR score [0-1] using the trained triplet loss models.

#### Table 6

Ear identification performance using trained triplet loss models on the **AWE** dataset in terms of the CIR score [0-1].

	ResNet50	EfficientNet	VGG16	ViT	Proposed
Ear(Left)	0.32	0.31	0.20	0.28	0.41
	0.24	0.29	0.24	0.20	0.30

In comparison to the KinEar dataset, the NASNet architectures show better performance on the I-Ear dataset. Interestingly, the pre-trained ViT model outperformed the CNN models for left ear recognition on the proposed *I-Ear* dataset. Whereas, the ResNet50 model performs the best when right-ear images are used for matching. The average dissimilarity score fusion of best-performing models shows a significant boost in recognition performance. Again overall, the cosine similarity with score fusion performs the best and for the right ear, it yields at -least 5% better performance than the best-performing model. While the simple dissimilarity fusion shows a jump in the recognition performance, the accuracy improves further with a fusion of triplet loss-trained models. For left ear recognition, the triplet loss model fusion yields 5% better performance than the dissimilarity score fusion. An improvement of 7% is also observed for right ear recognition when triplet loss model fusion is implemented contrary to the dissimilarity fusion.

## 5.3. Experimental Results and Analysis on the AWE Dataset

Since the triplet loss model shows significant improvements on both the earlier datasets namely KinEar and I-Ear datasets, we have further evaluated it for unconstrained ear recognition to evaluate its robustness in handling real-world factors. For that, we have used the AWE dataset collected directly from the web containing the ear images captured in-the-wild settings. The proposed fusion of the triplet loss model shows a significant jump in the recognition performance even on this in-the-wild dataset as compared to the individual CNN and ViT. It is observed that through the experiments on both I-Ear and AWE datasets the left ear yields higher recognition performance than the right ear. We want to highlight while the ResNet model consistently outperforms other networks, for right recognition, on the AWE dataset, the EfficientNet model shows the highest accuracy. The results on the AWE dataset are reported in Table 6.

The above-reported ear recognition performances are reported at rank-1, where it is assumed that the true match always occurs at the first spot. In other words, in rank-1 identification, it is assumed that only the true identity yields the lowest dissimilarity score. However, that might not be the case due to several reasons including environmental factors in which the images



**Figure 4:** The cumulative match characteristic (CMC) performance summarizing the recognition performance of the proposed triplet loss model under the proposed fusion or score fusion of the deep networks on the test set of the Datasets used in this research.



Figure 5: Grad CAM visualization results for correctly classified and misclassified images of each dataset.



Figure 6: Ear Recognition performance of the proposed triplet loss fused models in terms of ROCs on each dataset.

are captured which decreases the inter-class variation but increases the intra-class variations. To tackle this, we have also performed the ear recognition at different ranks and to report the performance at different ranks CMC values are used. As the rank increase, the search space to find the identity also increases; therefore, keeping in mind, we restrict the ear recognition till rank-10. Figure 4 shows the ear recognition on different datasets using the proposed triplet loss fusion network. As expected as soon as the rank increases, the matching performance increases drastically. As shown in Table 6, the AWE dataset yields only 38% accuracy for the right ear, and 41% accuracy for the left ear show a jump of at least 22% when the rank of identification increases from 1 to 5. Due to the significantly high performance of the I-Ear dataset at rank-1 itself, the improvement with increasing rank is lower than the other datasets.

The proposed triplet loss model-based ear recognition shows improved performance as compared to the individual CNN and ViT models; however, it is not clear which feature or region led to this performance. Therefore, to present a responsible ear recognition framework, we performed the Grad CAM visualization to better understand the region of interest of the ear where these models are focusing while extracting the features. We choose the best-performing networks obtained through the extensive experiments performed on the multiple ear recognition datasets including the proposed I-Ear dataset. Some examples of correctly classified and misclassified images taken from each dataset are shown in Figure 5. It is observed when the models took the correct decision, they are focusing on the ear region; whereas, the misclassification due to undesired distortion of the ear share, blurriness of the images, and when the models are focusing on the entire image where the background is higher than the foreground.

Figure 6 shows the ROC curve when the proposed triplet loss algorithm is used for evaluation using the score fusion of best-performing networks. On the AWE dataset, at  $10^{-3}$  false accept rate, the proposed model yield close to 41% true accept rate for right ear recognition. The left ear recognition shows a 6% higher true accept rate at the same false accept rate.

# 6. Conclusion

While biometrics recognition is an important and accurate medium of person identification, an issue of privacy can limit its deployment. Therefore, it is necessary not only to use the accurate biometric modality but also can preserve the privacy of individuals. It is observed that ear recognition has the potential to identify individuals while maintaining privacy; however, not received significant attention. To advance, the research in privacy-preserving identity recognition, first, we have benchmarked the effectiveness of various state-of-the-art deep networks including a vision transformer for unconstrained ear recognition. Later, based on the effectiveness of different deep neural networks, we proposed a triplet model for ear recognition. We have also presented a first-ever Indian ethnicity ear recognition stan conventional CNNs. In the future, we aim to develop a novel CNN inspired by the success of our proposed amalgamated model and increase the size of our dataset to cover the wide demographic variation of Indians.

# References

- H. Alshazly, C. Linse, E. Barth, T. Martinetz, Ensembles of deep learning models and transfer learning for ear recognition, Sensors 19 (2019) 4139.
- [2] G. Dvoršak, A. Diwedi, V. Štruc, P. Peer, Ž. Emeršič, Kinship verification from ear images: An explorative study with deep learning models, International Workshop on Biometrics and Forensics (2022).
- [3] Ž. Emeršič, V. Štruc, P. Peer, Ear recognition: More than a survey, Neurocomputing 255 (2017) 26–39.
- [4] Ž. Emeršič, B. Meden, P. Peer, V. Štruc, Evaluation and analysis of ear recognition models: performance, complexity and resource requirements, Neural computing and applications (2018) 1–16.
- [5] Ž. Emeršič, D. Štepec, V. Štruc, P. Peer, Training convolutional neural networks with limited training data for ear recognition in the wild, arXiv preprint arXiv:1711.09952 (2017).
- [6] Y. Zhang, Z. Mu, L. Yuan, C. Yu, Ear verification under uncontrolled conditions with convolutional neural networks, IET Biometrics 7 (2018) 185–198.
- [7] J. Zhang, W. Yu, X. Yang, F. Deng, Few-shot learning for ear recognition, in: Proceedings of the 2019 international conference on image, video and signal processing, 2019, pp. 50–54.
- [8] H. Sinha, R. Manekar, Y. Sinha, P. K. Ajmera, Convolutional neural network-based human identification using outer ear images, in: Soft Computing for Problem Solving: SocProS 2017, Volume 2, Springer, 2019, pp. 707–719.
- [9] E. E. Hansley, M. P. Segundo, S. Sarkar, Employing fusion of learned and handcrafted features for unconstrained ear recognition, IET Biometrics 7 (2018) 215–223.
- [10] D. Štepec, Ž. Emeršič, P. Peer, V. Štruc, Constellation-based deep ear recognition, Deep biometrics (2020) 161–190.
- [11] G. Dvoršak, A. Dwivedi, V. Štruc, P. Peer, Ž. Emeršič, Kinship verification from ear images: An explorative study with deep learning models, in: 2022 International Workshop on Biometrics and Forensics (IWBF), IEEE, 2022, pp. 1–6.
- [12] Ž. Emeršič, L. L. Gabriel, V. Štruc, P. Peer, Convolutional encoder-decoder networks for pixel-wise ear detection and segmentation, IET Biometrics 7 (2018) 175–184.
- [13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
- [14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
- [16] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: International conference on machine learning, PMLR, 2021, pp. 10096–10106.
- [17] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
- [18] B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le, Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8697–8710.
- [19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.