Unconstrained Ear Recognition through state-of-the-art Machine Learning: A Survey

Marwin B. Alejo (2020-20221)

Date of Submission: Jan 12, 2022

ABSTRACT

This paper presented a preliminary result of the performance of selected known and new state-of-the-art machine learning networks on ear biometrics tasks. Provided a straightforward deep learning pipeline, this paper determined the performance of ResNets, ResNeXt, ViT and its known new variants, and ResMLP and MLP-Mixer for ear recognition. Results show that ResNets achieve optimal performance while Transformer-based, particularly the vanilla Vision Transformer, achieve greater performance over its variants and MLP-centric networks on ear recognition or biometrics.

Keywords

Deep Learning, Neural Network, Vision Transformer, MLP, Ear Biometrics.

1. INTRODUCTION

Biometric recognition is a technology that allows the identification of individuals through their unique behavioral or physical traits. These traits are the fingerprint, face, iris, voice, gait, iris, retina, and ears [10,22]. As a means of biometry, the ears allow an individual to be identified through their ear's geometric, color, and texture features from afar in either constrained or unconstrained environments [1,14,16,39]. This advantage of ears for biometry over other means earned a momentum of interest in current computational methods and system development research.

Modern approaches of previous ear biometric studies utilize image processing techniques with statistical methods, and one of these papers is the work of Kavipriya et al. [23]. Their paper uses the canny edge detection algorithm and contour tracking method to realize ear biometrics for personal identification. The works of Mangayarkasi et al. [28] propose a similar approach for ear biometrics but without using a contour tracking method. The paper of Zarachoff et al. [44] proposed using a 2D Wavelet-based Multi-band PCA (2DWMBPCA) for ear biometrics and recognition. The works of Sajadi et al. [31] use the genetic algorithm for extracting local and global features of ear images for ear biometrics. While these ear biometrics methods produced exemplary results, recent studies suggested the use of machine learning for ear biometrics due to its higher performance over traditional methods [15].

Deep learning approaches are the most prevalent methods used in the computational studies of ear biometrics. The works of Khaldi et al. [24] proposed the use of deep unsupervised active learning with a generative adversarial network (GAN) for ear biometrics using various ear databases from the Mathematical Image Analysis (AMI), University of Science and Technology Beijing 2 (USTB2), and Annotated Web Ears (AWE). Their method achieved recognition rates of 100.00%, 98.33%, and 51.25% accordingly on the used datasets. The paper of Lei et al. [26] uses a single-shot MobileNet CNN model on USTB datasets for ear biometrics and achieved recognition accuracy of 99%. The works of Ying et al. [43] handcrafted a deep convolutional neural network architecture (DCNN) called ear-recognition-Net for ear biometrics. Their proposed method achieved a recognition rate of 95% to 98%. Chowdhury et al. [9] also proposed a handcrafted neural network algorithm for robust ear biometrics and achieved recognition accuracy of 98.2%. Instead of using handmade models, the paper of Alshazly et al. [4] proposed using pre-trained AlexNet, VGGNet, Inception, ResNet, and ResNeXt for ear biometrics with transfer learning and fine-tuning on the EarVN1.0 dataset. Their proposed method determined that ResNeXt achieved the best recognition over other models with a recognition rate of 95.85%. Similarly, the works of Alejo and Hate [2] uses pre-trained AlexNet, GoogLeNet, Inception-v3, Inception-ResNet, ResNet, SqueezeNet, ShuffleNet, and MobileNet models with transfer learning and fine-tuning for ear biometrics and determined that ResNet models achieve better results or 100% recognition rate over other CNN models. The paper of Alejo [3] presented the use of recently published state-of-the-art transformer models for unconstrained ear recognition and achieved am recognition accuracy of 93% to 96%. Furthermore, as machine learning continuously grows, recently presented state-of-the-art models like Vision Transformer and Transformer-inspired networks like ResMLP and MLP-mixer remain unexplored for biometrics thus, an open opportunity.

Inspired by the above-discussed information, this paper aimed to determine the performance of selected Transformer-inspired models for ear biometrics in terms of their recognition accuracy and memory utilization. This paper also aimed to determine whether the ResNets, the novel variants of ViT, or the novel variants of MLP which are Transformer inspired are better than ViT for Ear Biometrics. Furthermore, this paper (1) provided a generic deep learning pipeline for ear recognition regardless of the used network models; and (2) compared the recognition accuracy and memory utilization of the Transformer-inspired models to other selected state-of-the-art Transformer-based and CNN models.

The structure of the rest of this paper is as follows: section 2 provides a brief insight of the networks used in modeling the ear biometric models of this paper, section 3 discusses the methods and undertaken of this paper, section 4 briefly discusses the results of the experiments performed in this paper, and section 5 shows the objective-aligned conclusion of this paper.

2. STATE-OF-THE-ART MACHINE LEARNING MODELS

This paper considered state-of-the-art machine learning models for the implementation of ear biometrics on machine learning models. These are the ResNet and ResNeXt for the convolutional neural network, Vision Transformer (ViT), Data-efficient image Transformer (DeiT), Class-attention image Transformer (CaiT), Convolutional Vision Transformer (ConViT), Cross-attention Vision Transformer (CrossViT), Pooling-based imaget Transformer (PiT), Cross-Covariance image Transformer (XCiT), and Swin Transformer for the Transformer-based networks, and ResMLP and MLP-Mixer for the Transformer-inspired networks. Furthermore, this paper selected only the ResNet and ResNeXt as the state-of-the-art models of CNN due to their relative performance results for ear recognition over other CNN models as shown in the published works of Alejo and Hate [2] and Alshazly et. Al. [4] and its comparative performance over Transformers as shown in the paper of Raghu et. Al. [30].

2.1 Convolutional Neural Network

Convolutional Neural Network (CNN) or ConvNet is a deep neural network originally designed for image analysis and vision tasks that become popular in several application domains since the development of AlexNet [41]. While new CNN architectures gradually arise from paper to paper, their fundamental structure mainly consists of convolutional and pooling layers that operate on the complex dimensions of images. Furthermore, this paper presented only the ResNet and ResNeXt as the state-of-the-art models of CNN for ear biometrics due to their comparative performance over Transformers which are not applicable to other state-of-the-art models of CNN.

2.1.1 ResNet

Residual Networks or ResNet is one of the earlier CNN models that become popular after AlexNet. It is a brainchild of He et. Al. [18] with the philosophy of residual block which solves the issue of vanishing gradients of deep learning as the network becomes deeper. Moreover, this paper determined the performance of the ResNet model for Ear Biometrics using the ResNet18, ResNet50, and ResNet152 variants. Figure 1 shows the architecture summary and structure of these ResNet variants as used in this paper.

2.1.2 ResNeXt

ResNeXt is a CNN model developed by Xie et. Al. [40] as an extension of the networks of residual block model and with the idea of being simple yet a highly modularized architecture for image classification. It substitutes the normal residual block with one that uses the Inception models' "split-transform-merge" method as shown in figure 2. Simply said, rather than convolutions over the whole input feature map, the block's input is projected into several smaller

		ResNet			*ResNeXt*
	Output Size	18-layer	50-layer	152-layer	50-layer
Conv1	112x112	$7 \times 7, 64, stride\ 2$			
Conv2	56x56	$3 \times 3, \max pool, stride\ 2$ $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128, C = 32 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3	28x28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 8$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256, C = 32 \\ 1 \times 1, 512 \end{bmatrix} \times 3$
Conv4	14x14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 36$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512, C = 32 \\ 1 \times 1, 1024 \end{bmatrix} \times 3$
Conv5	7x7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 1024 \\ 3 \times 3, 1024, C = 32 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1x1	$average\ pool, 20\ identities, fc, softmax$			

C = group convolution

Figure 1: ResNet18, ResNet50, ResNet152, and ResNeXt50 Architecture

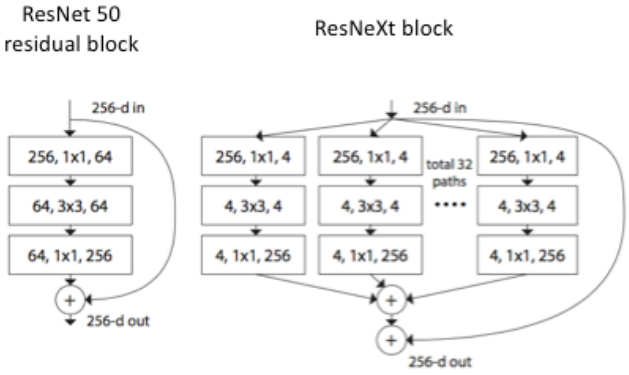

Figure 2: ResNet block vs. ResNeXt block

(channel) dimensional representations, each of which is subjected to a set of convolutional filters before being merged [5]. Moreover, figure 1 above also show the architecture and structure of ResNeXt-50 as used in this paper.

2.2 Transformer-based and Transformer-inspired Models

Transformers as a Neural Network is an innovative deep learning approach developed by Vaswani et. Al. [36]. It is a simple and scalable approach that outperforms the state-of-the-art results of Residual Neural Network (RNN) and CNN on Natural Language Processing (NLP) tasks. While vanilla Transformers are optimal for NLP tasks, several recent studies have been working on it on Computer Vision (CV) tasks [17,25]. Among of these are Detection Transformer (DeTr) [6] and Deformable DeTr [45] for object detection, Axial-DeepLab [37] and Cross-Modal Self-Attention [42] for image segmentation, Image Transformer [29], Image GPT [8], TransGAN [21], and SceneFormer [38] for image generation, and Vision Transformer (ViT) [12], Data-efficient image Transformer (DeiT) [34], Class-attention image Transformer (CaiT) [35], Convolutional Vision Transformer (ConViT) [11], Cross-attention Vision Transformer (CrossViT) [7], Pooling-based Vision Transformer (PiT) [19] and Cross-covariance image Transformer (XCiT) [13] for image recognition. These successes of Transformers in CV also

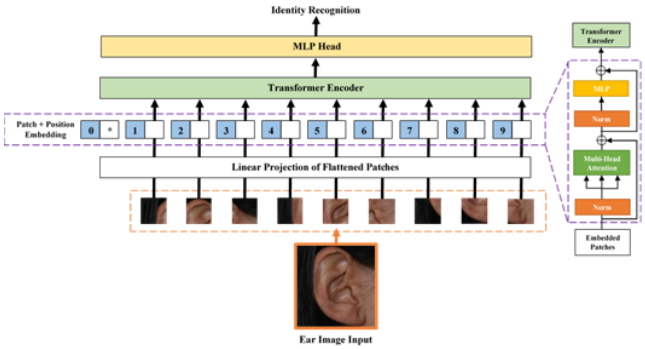

Figure 3: Vanilla ViT Architecture

inspire other studies to extend its features onto Transformer-inspired models and among these are the ResMLP [33] and MLP-Mixer [32]. Moreover, this paper focuses only on the implementation of the above-mentioned Vision Transformers and Transformer-inspired models due to computational constraints. The following subsections below briefly discusses each of these Transformer-based and Transformer-inspired models.

2.2.1 Vanilla Vision Transformer (ViT)

The vanilla Vision Transformer that most research are using today is a state-of-the-art algorithm of Dosovitsky et. Al. [12] that uses a modified Transformer network to suffice directly on images instead of NLP data. ViT chunks the input image into square patches and flattened into a single vector by joining all channels of pixels in a patch and linearly injecting each into the desired dimension. It also uses a positional embedding which allow the Transformer to learn the patches position of the input image. Figure 3 shows the architecture and structure of the used vanilla ViT of this paper. Moreover, this paper implemented the ViT on the standard 16x16 patches.

2.2.2 Data-efficient image Transformer (DeiT)

The Data-efficient image Transformer is one of the fork of Vision Transformer particularly developed for object recognition tasks by Touvron et. Al. [34]. It uses a Transformer-centric teacher-student strategy to train the model of the input image. It also relies on distillation token to ensure that the student learn from the teacher through attention. Figure 4 shows the DeiT architecture and structure as used in this paper. Moreover, the DeiT on this paper uses a patch size of 16.

2.2.3 Class-attention image Transformer (CaiT)

The Class-attention image Transformer (CaiT) is a modified Vision Transformer model of Touvron et. Al. [35] which alleviates the training difficulty of ViT at greater depths. Among of the notable features of this model is the multiplication of the output of the residual block to the per channel of the input image and that the patches to attend onto one another and only allow the class token to attend to the patches in the last layers. Figure 5 shows the architecture of CaiT as shown in its original paper. Moreover, this paper implemented the CaiT on a model with 24 blocks.

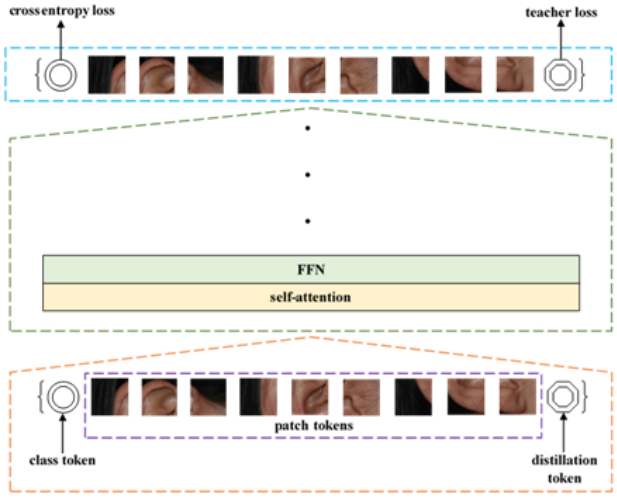

Figure 4: DeiT Architecture

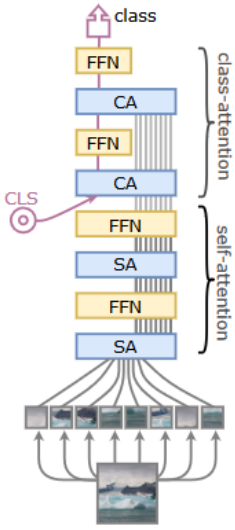

Figure 5: CaiT Architecture

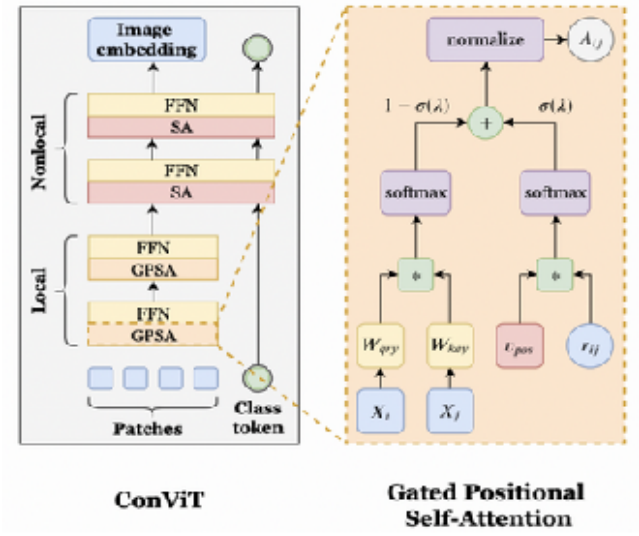

Figure 6: ConViT Architecture

2.2.4 *Vision Transformer with Soft Convolution (ConViT)*

Convolutional Vision Transformer (ConViT) is a modified ViT developed by d'Ascoli et. Al. [11] with the behavior of CNN in mind. It introduces the concept of Gated Positional Self-Attention (GPSA) that act as a soft inductive bias and allow the Transformer network to model itself from the input. The GPSA also allows the network to flexibly learn from the input image by having control over how much standard attention is will consume rather than the standard position-based embeddings of the vanilla ViT. Figure 6 shows that standard architecture and structure of ConViT with GPSA highlighted.

2.2.5 *Cross-Attention image Transformer (CrossViT)*

The Cross-attention image Transformer (CrossViT) is a modified ViT developed by Chen et. Al. [7] that can extract multi-scale representation of images using a dual branch as architecture as shown in figure 7. It mixes picture patches (transformer tokens) of various sizes to provide a stronger visual characteristic. It uses two independent branches with varying computing complexities to process tiny and big patch tokens and fuse each together many times to complement each other. The model achieves the fusion in linear time of the two branches by having each transformer construct a non-patch token as an agent to communicate information with each other through the attention mechanism of Transformers.

2.2.6 *Pooling-based image Transformer (PiT)*

The Pooling-based image Transformer (PiT) is a modified ViT developed by Heo et. Al. [19] with the idea of minimizing the token parameters of ViT through depth-wise convolutions or pooling. Its concept adapted from the downsampling procedure of ResNets which minimizes the features from input to output and applying it onto the core of ViT as shown in figure 8.

2.2.7 *Cross-Covariance image Transformer (XCiT)*

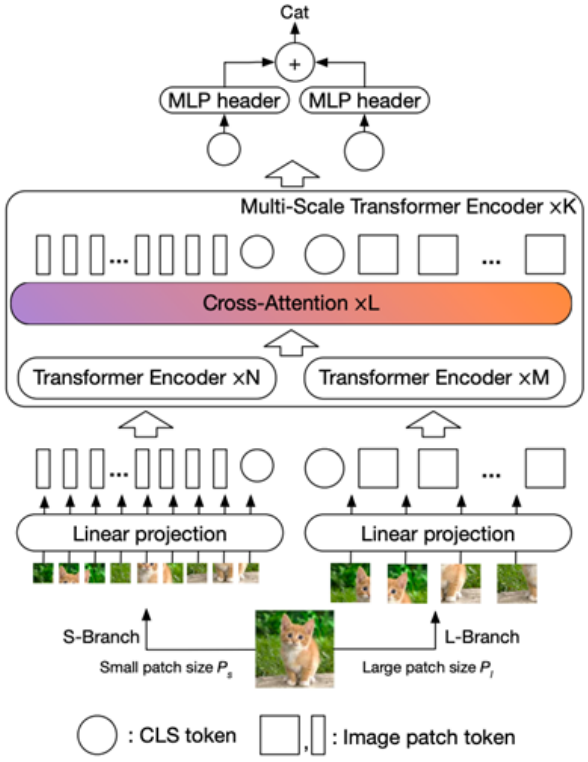

Figure 7: CrossViT Architecture

Cross-covariance image Transformer (XCiT), as developed by El-Nouby et. Al. [13], is a modified ViT with the idea of transposing self-attention and Cross-Covariance Attention (XCA) operation at its core which allow the ViT to be linearly flexible and scalable like ConvNets without compromising the performance over data of various sizes. Figure 9 shows the XCiT architecture highlighting the transposed self-attention layer or XCA. Moreover, this paper implemented the XCiT with 16x16 patches

2.2.8 *Swin Transformer*

Swin Transformer is a novel Vision Transformer model developed by Liu et. al. [27]. It introduces the operation of "shifted windowing" which bring efficient computation on high-resolution of image pixels and limits self-attention computation to non-overlapping local windows while allowing cross-window connection. It also has computational complexity with respect to image size which allow it to be versatile for several vision tasks at various scales. Figure 10 below shows the architecture of Swin Transformer as adapted from its original paper.

2.2.9 *ResMLP*

ResMLP is a Transformer-inspired model developed by Touvron et. Al. with an architecture idea of pure Multi-Layer Perceptron (MLP) for image recognition. It consisted of basic residual networks substituted linear layers through which image patches interact independently and identically across channels. It also contains a two-layer feed-forward network in which channels interact independently per patch. Figure

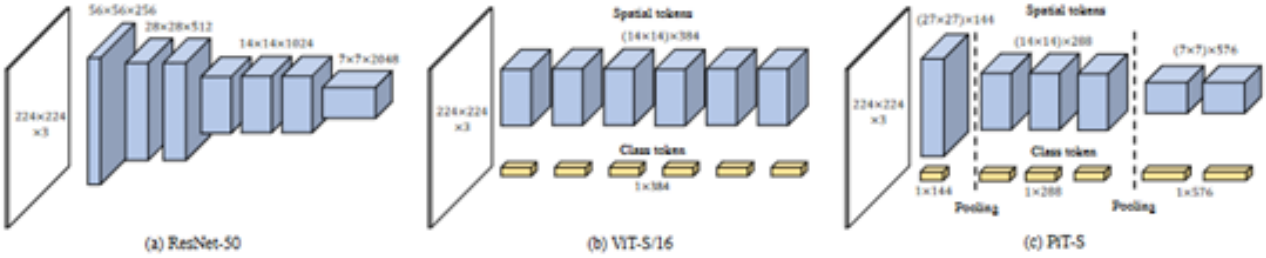

Figure 8: PiT Architecture

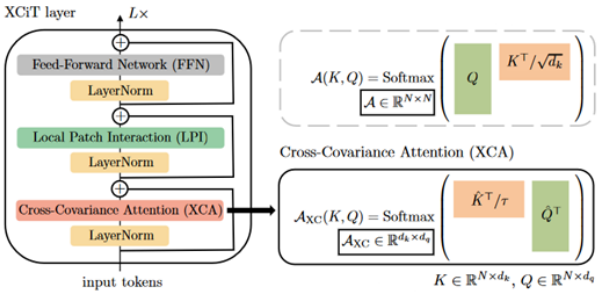

Figure 9: XCiT Architecture

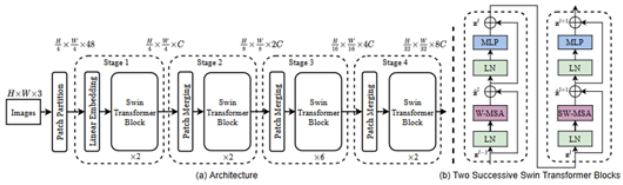

Figure 10: Swin Transformer Architecture

11 shows the architecture and structure of ResMLP. Moreover, this paper implemented this model on three different configurations of residual block of linear layers: ResMLP-12, ResMLP-24, and ResMLP-36.

2.2.10 MLP-Mixer

MLP-Mixer is a Transformer-inspired model developed by Tolstikhin et. Al. [32] that exclusively use MLP as its core for image recognition tasks and do not employ any convolutions and attention-based layers. It consisted of two types of MLP layers: one MLP layer that applied independently to image patches for location embeddings mixing and another one MLP applied across patches for spatial information mixing. Figure 12 below shows the MLP-Mixer architecture and structure. Moreover, this paper implemented this model on one configuration only.

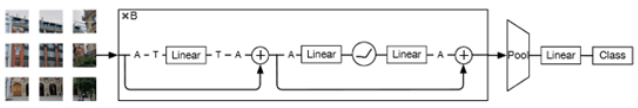

Figure 11: ResMLP Architecture

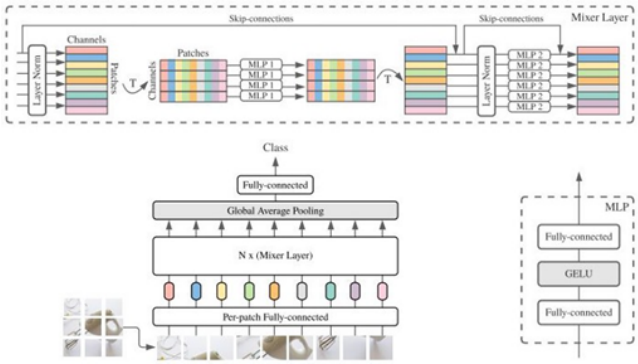

Figure 12: MLP-Mixer Architecture

While these models show exemplary performance on general object recognition task of computer vision, no paper at present shows the efficiency of these models on the context of ear biometrics in terms of recognition performance and memory utilization. Furthermore, this paper might be significant to biometric researchers that seeks to utilize the

3. EXPERIMENT

The experimental procedure of this paper consisted of four stages: (1) Dataset and Input Data, (2) Data Preprocessing, (3) Modeling through Transfer Learning, and (4) Classification and Output. Figure 13 shows the step-by-step flow of this procedure.

3.1 Ear Dataset and Input Data

The dataset utilized in this investigation is EarVN1.0 [20]. It is the world's largest ear picture collection, mostly gathered for identification purposes. It comprised of 164 individuals' raw ear photos taken in the wild (unconstrained), each with a total of 180 images for a total of 28,412 ear images. Due to a lack of computing resources, this study only looked at the first 20 classes of the EarVN1.0 dataset, resulting in a total of 4000 ear pictures. Furthermore, the reduced dataset was divided into two parts: 80 percent training and 20 percent testing/validation. A collection of training and testing/validation datasets is the result of this process. Figure 14 depicts selected ear pictures from the EarVN1.0 dataset.

3.2 Data Preprocessing

The partitioned training dataset was preprocessed by scaling each ear picture to 224 square pixels, flipping horizontally

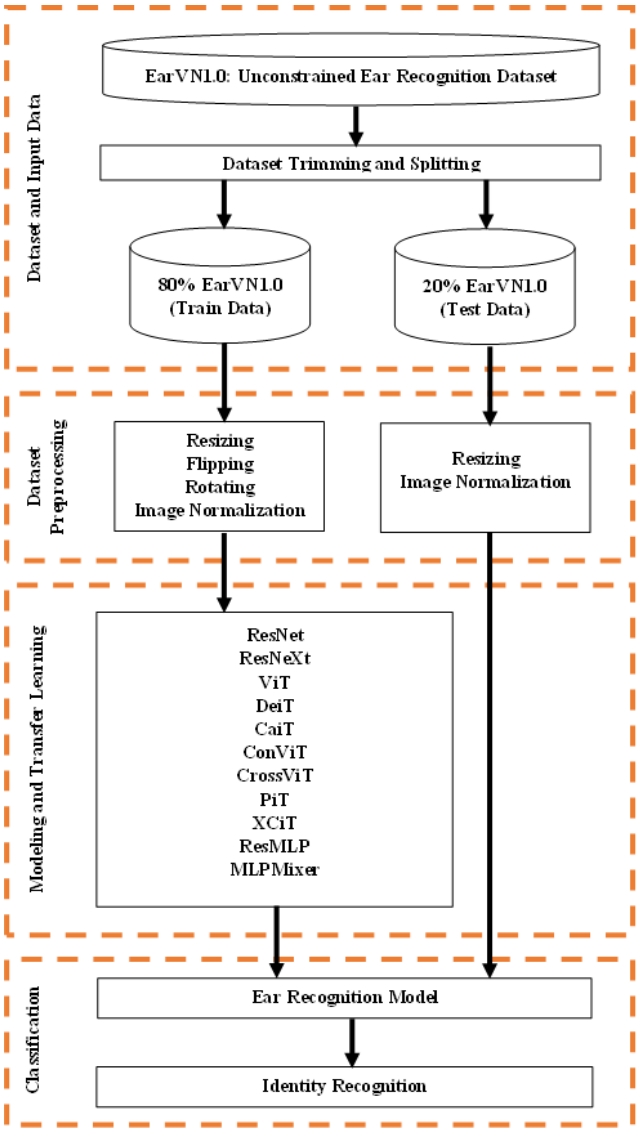

Figure 13: Deep Learning Pipeline

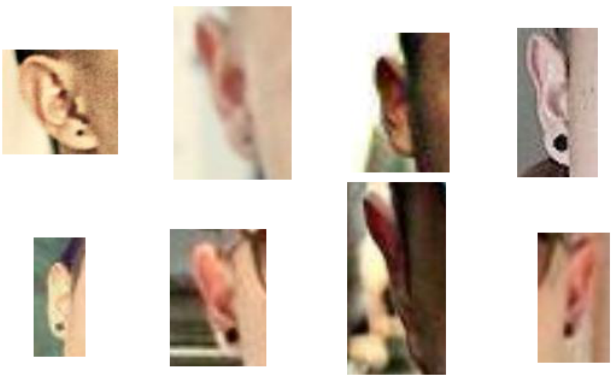

Figure 14: Selected ear images from EarVN1.0 dataset.

Table 1: Pretrained Models and Modeling Configuration

Pretrained Models	Configuration
ResNet18	Batch size: 32 Learning rate: 0.00002 Optimizer: Adam Epoch: 20
ResNet50	
ResNet152	
ResNeXt50	
ViT	
DeiT	
CaiT	
ConViT	
CrossViT	
PiT	
XCiT	
Swin Transformer	
ResMLP12	
ResMLP24	
ResMLP36	
MLP-Mixer	

and vertically, rotating by 30 degrees, and normalizing using conventional ImageNet normalization parameters. The testing/validation dataset was also preprocessed by resining 224 square pixels and normalizing each resized ear picture using conventional ImageNet normalization parameters. A series of preprocessed training and testing/validation ear pictures is the result of this phase.

3.3 Modeling and Transfer Learning

Considering the descriptions of the above-discussed models in their papers and the number of used training data, this paper implemented these models through transfer learning on their pretrained ImageNet models, as provided in Table 1, and fine-tuned each models' final layers to only recognize 20 people. To provide a fair configuration of these models, this paper implemented these models using PyTorch and PyTorch Image Models (TIMM) on Google Colab GPU with a batch size of 32, a learning rate of 0.00002, and Adam optimizer on 20 epochs. This paper only considered 20 epochs for all models due to the memory restrictions with the used computing node. Furthermore, Table 1 summarizes the used modeling configuration of this paper.

3.4 Classification and Output

This stage determines the performance of the trained ear biometrics models using the partitioned test dataset of this paper. The output of this stage is a comparative analysis of the recognition performance as well as the memory utilization of the models discussed above on the context of ear biometrics.

4. ANALYSIS AND DISCUSSION OF RESULTS

This paper trained the ear biometrics models on selected ResNet, ResNeXt, Transformer-based, and Transformer-inspired

networks on 20 epochs with a batch size of 32, a learning rate of 0.00002, and the Adam optimizer. Considering the modeling constraints of this study, the fine-tuned ResNets of this paper took 16 minutes to 1.5 hours to develop an ear biometric model out of the EarVN1.0 dataset and achieved recognition rates of 79.57% for ResNet18, 81% for ResNet50, and 61.42% for ResNet152, and with a memory utilization of up to 2.40GB as shown in Table 2. It is observable from these ResNet results that ResNet50 is the most optimal ResNet network depth for the currently used ear biometrics dataset and configuration. It also implies that as the depth of ResNet becomes deeper than ResNet50, the network suffers from generalization loss thus, lowering the accuracy. Furthermore, the ResNeXt50 model also achieved a comparable result to ResNet18 in terms of recognition accuracy and ResNet50 in terms of memory utilization, implying that the inception-like block of ResNeXt allows the model to generalize further even on deeper network depth as compared with the standard ResNet block.

The fine-tuned Transformer-based models of this paper achieved greater recognition accuracy and lower memory utilization as compared with those of ResNets. The fine-tuned ViT of this paper achieved recognition accuracy of 97.36% with a memory utilization of 2.36GB although it took four hours of modeling. The DeiT, ConViT, and PiT achieved recognition accuracy of 95% to 97% with memory utilization of up to three gigabytes and took of up to four hours of modeling time. The XCiT, CaiT, Swin Transformer, and CrossViT achieved recognition accuracy of 75% to 89% with memory utilization of up to 2.8GB. Moreover, notable of these Transformer-based networks is the modeling time of XCiT, due to its XCA module that works like ConvNet, is on par with the performance of CNN-based models. Table 3 below shows the summary of the performance of these Transformer-based models.

The fine-tuned Transformer-inspired models of this paper achieved lesser recognition accuracy as compared with the ResNets and the Transformer-based models. The fine-tuned ResMLP12, ResMLP24, and ResMLP36 achieved recognition accuracy of up to 52% on 20 epochs with memory utilization of up to 2.3GB which is of equivalent to the memory used by XCiT, CaiT, Swin Transformer, and ViT. The fine-tuned MLP-Mixer model of this study achieved recognition accuracy of up to 40% with a memory utilization of 2.18GB. These results imply that these transformer-inspired networks suffer greatly in the task of Ear Biometrics considering the present configuration of the experiment. These results also mean that the statistics of the currently used ear datasets, as well as the data augmentation processes, do not suffice these models to generalize and generate on par results to Transformer-based models and ResNets. Table 4 shows the summary results of these Transformer-inspired models on Ear biometrics.

5. CONCLUSION

This paper determined the performance of recently published and known state-of-the-art machine learning models for ear biometrics in terms of their recognition accuracy and memory utilization. Among these networks are the ResNets and ResNeXt, ViT and its known variants, and MLP-centric models like ResMLP and MLP-Mixer. This paper also provided a straightforward deep learning pipeline that employed these models. The modeling process of this paper considered using Transfer Learning and Fine-tuning which replaces the final layer of the pre-trained models of these networks on ImageNet as such it can only recognize the selected 20 classes or identities of the used EarVN1.0 dataset. Moreover, this paper determined that ResNets and ResNeXt achieved recognition accuracy of up to 81% and of up to 2.40GB memory utilization, selected Transformer-based networks achieved recognition accuracy of up to 97.36% on up to 3GB memory utilization and that MLP-centric network could achieve recognition accuracy of up to 50% on up to 2.3GB memory utilization.

This paper determined that Transformers particularly the ViT performed better in the context of Ear Biometrics over ResNets and novel MLP-centric networks. Regardless of the novel features of the variants of Vision Transformer, the vanilla ViT still outperforms its variants in terms of recognition accuracy. This paper also realized that novel MLP-centric networks could achieve better accuracy on ear biometrics tasks, although not in the configuration of this paper when it considered training on longer epochs and greater dataset statistics of augmentation processes. Overall, ViT and its variants that have greater parameter size could achieve better performance on the ear recognition task and could outperform state-of-the-art CNN models like ResNets and its variants and the novel ResMLPs and MLP-Mixers.

Considering that this paper presented only preliminary results of the performance of selected known state-of-the-art machine learning models, interested parties and individuals are encouraged to conduct further relevant studies, execution, and development of the topic of this paper prior to journal realization, presentation, and publication.

6. REFERENCES

[1] AbazaAyman, RossArun, HebertChristina, HarrisonMary Ann F., and NixonMark S. 2013. A survey on ear biometrics. ACM Computing Surveys (CSUR) 45, 2 (March 2013). DOI:https://doi.org/10.1145/2431211.2431221

[2] M. Alejo and C.P.G. Hate. 2019. Unconstrained ear recognition through domain adaptive deep learning models of convolutional neural network. International Journal of Recent Technology and Engineering 8, 2 (2019). DOI:https://doi.org/10.35940/ijrte.B2865.078219

[3] Marwin Alejo. 2021. UNCONSTRAINED EAR RECOGNITION USING TRANSFORMERS. Jordanian Journal of Computers and Information Technology 0 (2021), 1. DOI:https://doi.org/10.5455/JJCIT.71-1627981530

[4] Hammam Alshazly, Christoph Linse, Erhardt Barth, and Thomas Martinetz. 2020. Deep convolutional neural networks for unconstrained ear recognition. IEEE Access 8, (2020), 170295–170310. DOI:https://doi.org/10.1109/ACCESS.2020.3024116

[5] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. 2016. An Analysis of Deep Neural Network

Table 2: Summary Results of ResNets

Model	Modeling Time (hours)	Recognition Accuracy (%)	Memory Utilization (GB)	Parameter Size (MB)
ResNet18	0.25	79.57	2.15	11.19
ResNet50	0.3	81	2.25	23.55
ResNet152	1.30	61.42	2.38	58.21
ResNeXt50	0.75	78.25	2.30	23.02

Table 3: Summary Results of Transformer-based Models

Model	Modeling Time (hours)	Recognition Accuracy (%)	Memory Utilization (GB)	Parameter Size (MB)
ViT	4	97.36	2.36	85.81
DeiT	3.5	95.43	2.82	85.81
ConViT	3.25	96.76	2.77	85.79
PiT	3	95.73	2.27	72.76
XCiT	0.75	75.24	2.34	2.92
CaiT	3.5	80.77	2.37	11.77
Swin Transformer	3	86.66	2.32	86.77
CrossViT	3.25	88.94	2.82	103.89

Table 4: Summary Results of Transformer-inspired Models

Model	Modeling Time (hours)	Recognition Accuracy (%)	Memory Utilization (GB)	Parameter Size (MB)
ResMLP12	3	50	1.87	14.97
ResMLP24	3	50.04	2.18	14.97
ResMLP36	3	52.16	2.24	44.31
MLP-Mixer	3	40.39	2.18	20.72

Models for Practical Applications. (May 2016). Retrieved January 10, 2022 from https://arxiv.org/abs/1605.07678v4

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12346 LNCS, (May 2020), 213–229. Retrieved August 2, 2021 from https://arxiv.org/abs/2005.12872v3

[7] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. (March 2021). Retrieved January 11, 2022 from https://arxiv.org/abs/2103.14899v2

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative Pretraining From Pixels. In Proceedings of the 37th International Conference on Machine Learning, PMLR, 1691–1703. Retrieved August 2, 2021 from http://proceedings.mlr.press/v119/chen20s.html

[9] Mozammel Chowdhury, Rafiqul Islam, and Junbin Gao. 2018. Robust ear biometric recognition using neural network. Proceedings of the 2017 12th IEEE Conference on Industrial Electronics and Applications, ICIEA 2017 2018-February, (February 2018), 1855–1859. DOI:https://doi.org/10.1109/ICIEA.2017.8283140

[10] Shaveta Dargan and Munish Kumar. 2020. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. Expert Systems with Applications 143, (April 2020), 113114. DOI:https://doi.org/10.1016/J.ESWA.2019.113114

[11] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. 2021. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. (March 2021), 139. Retrieved January 11, 2022 from https://arxiv.org/abs/2103.10697v2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (October 2020). Retrieved August 2, 2021 from http://arxiv.org/abs/2010.11929

[13] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jegou. 2021. XCiT: Cross-Covariance Image Transformers. (June 2021). Retrieved January 11, 2022 from https://arxiv.org/abs/2106.09681v2

[14] Žiga Emeršič, Vitomir Štruc, and Peter Peer. 2017. Ear recognition: More than a survey. Neurocomputing 255, (September 2017), 26–39. DOI:https://doi.org/10.1016/J.NEUCOM.2016.08.139

[15] Žiga Emeršič, Aruna Kumar S. V., B. S. Harish, Weronika Gutfeter, Jalil Nourmohammadi Khiarak, Andrzej Pacut, Earnest Hansley, Mauricio Pamplona Segundo, Sudeep Sarkar, Hyeonjung Park, Gi Pyo Nam, Ig-Jae Kim, Sagar G. Sangodkar, Ümit Kaçar, Murvet Kirci, Li Yuan, Jishou Yuan, Haonan Zhao, Fei Lu, Junying Mao, Xiaoshuang Zhang, Dogucan Yaman, Fevziye Irem Eyiokur, Kadir Bulut Özler, Hazım Kemal Ekenel, Debbrota Paul Chowdhury, Sambit Bakshi, Pankaj K. Sa, Banshidhar Majhi, Peter Peer, and Vitomir Štruc. 2019. The Unconstrained Ear Recognition Challenge 2019 - ArXiv Version With Appendix. (March 2019). Retrieved September 5, 2021 from https://arxiv.org/abs/1903.04143v3

[16] Lauren P. Etter, Elizabeth J. Ragan, Rachael Campion, David Martinez, and Christopher J. Gill. 2019. Ear biometrics for patient identification in global health: a field study to test the effectiveness of an image stabilization device in improving identification accuracy. BMC Medical Informatics and Decision Making 2019 19:1 19, 1 (June 2019), 1–9. DOI:https://doi.org/10.1186/S12911-019-0833-9

[17] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. 2020. A Survey on Vision Transformer. (December 2020). Retrieved January 11, 2022 from https://arxiv.org/abs/2012.12556v4

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem, (2016), 770–778. DOI:https://doi.org/10.1109/CVPR.2016.90

[19] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. 2021. Rethinking Spatial Dimensions of Vision Transformers. (March 2021). Retrieved January 11, 2022 from https://arxiv.org/abs/2103.16302v2

[20] Vinh Truong Hoang. 2019. EarVN1.0: A new large-scale ear images dataset in the wild. Data in Brief 27, (December 2019), 104630. DOI:https://doi.org/10.1016/J.DIB.2019.104630

[21] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. 2021. TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. (February 2021). Retrieved August 2, 2021 from http://arxiv.org/abs/2102.07074

[22] Aman Kamboj, Rajneesh Rani, and Aditya Nigam. 2021. A comprehensive survey and deep learning-based approach for human recognition using ear biometric. Visual Computer (April 2021), 1–34. DOI:https://doi.org/10.1007/S00371-021-02119-0/TABLES/12

[23] P. Kavipriya, M.R. Ebenezar Jebarani, T. Vino, and G. Jegan. 2021. Ear biometric for personal identification using canny edge detection algorithm and contour tracking method. Materials Today: Proceedings (April 2021). DOI:https://doi.org/10.1016/J.MATPR.2021.03.351

[24] Yacine Khaldi, Amir Benzaoui, Abdeldjalil Ouahabi, Sebastien Jacques, and Abdelmalik Taleb-Ahmed. 2021. Ear Recognition Based on Deep Unsupervised Active Learning. IEEE Sensors Journal (2021), 1–1. DOI:https://doi.org/10.1109/JSEN.2021.3100151

[25] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Transformers in Vision: A Survey. (January 2021). Retrieved August 2, 2021 from http://arxiv.org/abs/2101.01169

[26] Yanmin Lei, Baowei Du, Junru Qian, and Zhibin Feng. 2020. Research on Ear Recognition Based on SSD-MobileNet-v1 Network. Proceedings - 2020 Chinese Automation Congress, CAC 2020 (November 2020), 4371–4376. DOI:https://doi.org/10.1109/CAC51589.2020.9326541

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. (March 2021). Retrieved January 12, 2022 from https://arxiv.org/abs/2103.14030v2

[28] N. Mangayarkarasi, G. Raghuraman, and A. Nasreen. 2019. Contour Detection based Ear Recognition for Biometric Applications. Procedia Computer Science 165, (January 2019), 751–758. DOI:https://doi.org/10.1016/J.PROCS.2020.01.016

[29] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image Transformer. 35th International Conference on Machine Learning, ICML 2018 9, (February 2018), 6453–6462. Retrieved August 2, 2021 from https://arxiv.org/abs/1802.05751v3

[30] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do Vision Transformers See Like Convolutional Neural Networks? (August 2021). Retrieved January 9, 2022 from https://arxiv.org/abs/2108.08810v1

[31] Shabbou Sajadi and Abdolhossein Fathi. 2020. Genetic algorithm based local and global spectral features extraction for ear recognition. Expert Systems with Applications 159, (November 2020), 113639. DOI:https://doi.org/10.1016/J.ESWA.2020.113639

[32] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP Architecture for Vision. (May 2021). Retrieved January 11, 2022 from https://arxiv.org/abs/2105.01601v4

[33] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, Hervé Jégou, and Facebook Ai. 2021. ResMLP: Feedforward networks for image classification with data-efficient training. (May 2021). Retrieved November 24, 2021 from https://arxiv.org/abs/2105.03404v2

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2020. Training data-efficient image transformers amp; distillation through attention. (December 2020). Retrieved August 2, 2021 from https://arxiv.org/abs/2012.12877

[35] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, Hervé Jégou, and Facebook Ai. 2021. Going deeper with Image Transformers. (March 2021). Retrieved January 11, 2022 from https://arxiv.org/abs/2103.17239v2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems 2017-December, (June 2017), 5999–6009. Retrieved August 2, 2021 from https://arxiv.org/abs/1706.03762v5

[37] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. 2020. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12349 LNCS, (March 2020), 108–126. Retrieved August 2, 2021 from https://arxiv.org/abs/2003.07853v2

[38] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. 2020. SceneFormer: Indoor Scene Generation with Transformers. (December 2020). Retrieved August 2, 2021 from https://arxiv.org/abs/2012.09793

[39] Zhaobin Wang, Jing Yang, and Ying Zhu. 2019. Review of Ear Biometrics. Archives of Computational Methods in Engineering 2019 28:1 28, 1 (November 2019), 149–180. DOI:https://doi.org/10.1007/S11831-019-09376-2

[40] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-January, (November 2016), 5987–5995. DOI:https://doi.org/10.1109/CVPR.2017.634

[41] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. 2018. Convolutional neural networks: an overview and application in radiology. Insights into Imaging 9, 4 (August 2018), 611–629. DOI:https://doi.org/10.1007/S13244-018-0639-9/FIGURES/15

[42] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-Modal Self-Attention Network for Referring Image Segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June, (April 2019), 10494–10503. Retrieved August 2, 2021 from https://arxiv.org/abs/1904.04745v1

[43] Tian Ying, Wang Shining, and Li Wanxiang. 2018. Human ear recognition based on deep convolutional neural network. Proceedings of the 30th Chinese Control and Decision Conference, CCDC 2018 (July 2018), 1830–1835. DOI:https://doi.org/10.1109/CCDC.2018.8407424

[44] Matthew Zarachoff, Akbar Sheikh-Akbari, and Dorothy Monekosso. 2019. Single image ear recognition using wavelet-based multi-band PCA. European Signal Processing Conference 2019-September, (September 2019). DOI:https://doi.org/10.23919/EUSIPCO.2019.8903090

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. (October 2020). Retrieved August 2, 2021 from https://arxiv.org/abs/2010.04159v4