# ViT and DeiT on Unconstrained Ear Recognition

Marwin B. Alejo
2020-20221
marwin.alejo@eee.upd.edu.ph

## I. Brief Introduction and Motivation

- Ear recognition is a subset of object detection with ear images containing a unique human identity signature as the subject for detection.
- Unconstrained ear recognition as a subset of ear recognition aims to recognize human identities through ear images taken from the wild. It had become a budding and open sector of biometric research in recent years, imploring computer vision methods.
- Past studies utilize image processing algorithms for the fabrication of an unconstrained ear recognition system. However, it is tedious and requires a large number of computing resources. [1, 2, 3]
- Recent studies suggested using different hand-crafted CNN architectures and algorithms to fabricate an unconstrained ear recognition system. However, the limited and small number of unconstrained ear datasets hinders the efficient implementation of these CNN algorithms as it requires a large number of ear images. [1, 2, 3]
- As a resort to the limitations of the recent studies, the concept of transfer learning is introduced to unconstrained ear recognition to extend CNN with only a limited number of unconstrained ear images without relying on hand-crafted and complex algorithms for practical and efficient system development. [1, 2, 3, 4]
- Given the above limitations and outcomes of former unconstrained ear recognition studies and the field itself, this mini-project generally aims to extend a pure Transformer Network to the task of unconstrained ear recognition. Transformer in deep learning is a nascent technology used initially in NLP but recently attracts researchers to explore and extend it in Computer Vision applications.

## II. Objectives

This mini-project aims to extend vanilla Transformer Network to Unconstrained Ear Recognition as a computer vision task and a subset of object detection with the following specific goals:

1. Measure the transformer-based unconstrained ear recognition network's performance in terms of validation loss and accuracy; and
2. Compare the validation accuracy as performance metric of the Transformer-based network with CNN-based through transfer learning network for unconstrained ear recognition.

## III. Limitations of the Mini-project

As a mini-project, this study is limited only to the exploration of Transformer Network, specifically Vision-Transformer (ViT) without the inclusion of hybrid networks and Data-efficient image Transformer (DeiT), with unconstrained ear recognition by extension and compare its performance with the performance of the published results of the same task but uses a CNN-based approach. Furthermore, considering constraints such as time and computing resources (Google Colab Free) and since this mini-project does not aim to develop an unconstrained ear recognition prototype or an end-to-end system, all transformer networks are trained in up to 20-30 epochs only while CNN-based networks are trained beyond 100 epochs.

## IV. Review of Related Works

Transformers first appeared as a simple and scalable solution to attain state-of-the-art results in NLP and is currently being extended in Computer Vision. Like transfer learning in CNN, Transformer Networks may be trained over large datasets and fine-tuned to learn on small datasets. Among the most recently developed transformer architectures for computer vision are Detection-Transformer (DeTr) [5], Vision-Transformer (ViT) [6] and Data-efficient image Transformer (DeiT) [7]. Although DeTr shows a comparable result over CNN-based detection models, its architecture does not suit directly with the recognition task of computer vision through which ViT and DeiT are evenly suited.

There have been no published studies relevant to any transformer-based recognition task at the moment of this writing except for Vision-Transformer (ViT) and Data-efficient image Transformer (DeiT). These Transformer networks achieved a state-of-the-art accuracy after being trained on ImageNet dataset. Most literature works in this area of computer vision and deep learning are subjected to the development of complex hand-crafted Transformer Networks for detection tasks and the architecture alone. Hence, in this mini-project, unconstrained ear recognition is extended to transformer-based networks, specifically the Vision-Transformer (ViT) and Data-efficient image Transformer (DeiT).

### Vision-Transformer (ViT)

Vision-Transformer (ViT) employs a modified transformer network architecture as such it may directly operate on images instead of text. ViT divides an image into grid square patches. Each patch is flattened into a single vector by joining all the channels of pixels in a patch and linearly inject it into the desired dimension. For it to learn, a learnable position embedding is added to each patch and allows the Transformer Network to learn the image's positional patch. Figure 1 shows the ViT architecture as shown in its original paper.

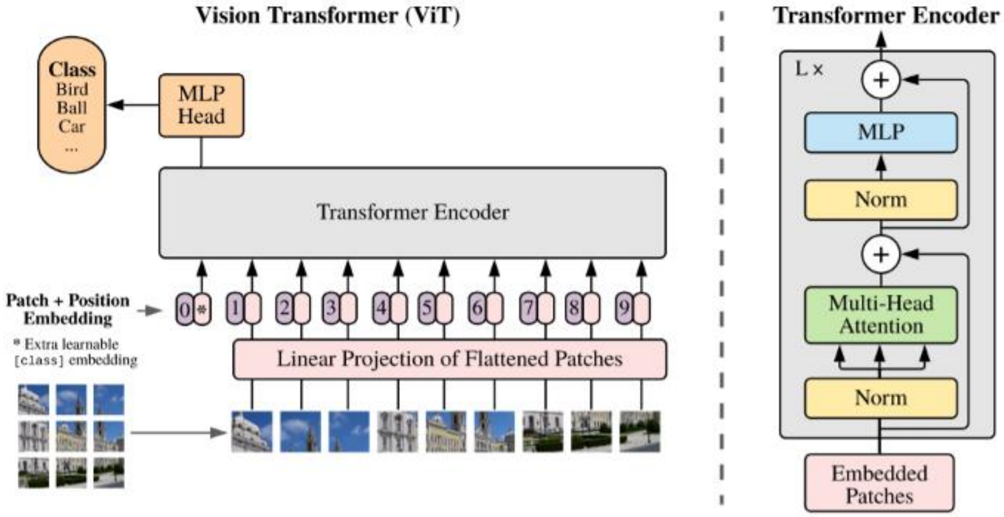

**Figure 1 Vision-Transformer (ViT) Architecture** *[6]*

### Data-efficient image Transformer (DeiT)

DeiT is one of the Transformers networks developed for the recognition task of computer vision without relying to convolutional layer. The training strategy of DeiT is similar to those of convolutional training and learning features from datasets. The unique part which made this architecture comparable to those of CNN architecture is the addition of distillation which aids the learning process of the machine towards recognition modeling. Shown in figure 2 is the architecture of DeiT.

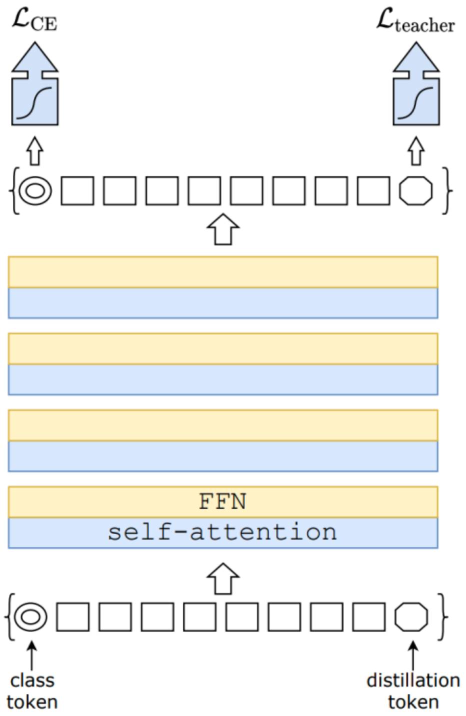

**Figure 2 DeiT Architecture** *[7]*

## V. Methodological Approach

Figure 3 shows the method used in this mini-project in chronological order.

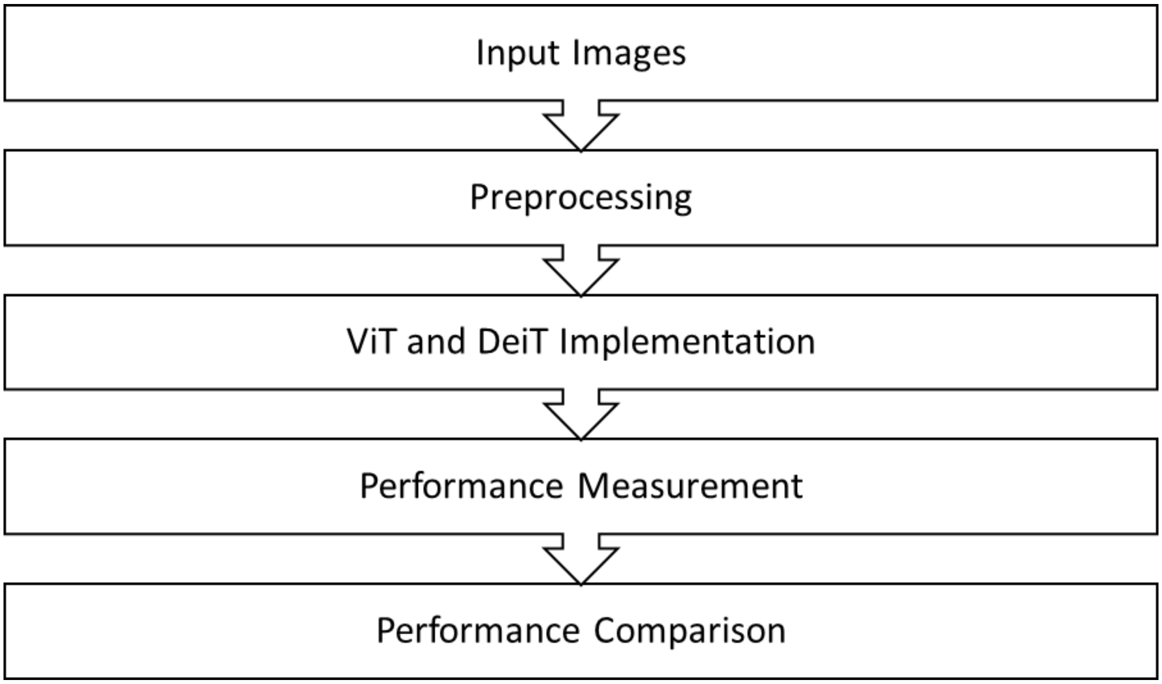

**Figure 3 Methods used in the mini-project.**

### Input Images

EarVN1.0 [8] is the used dataset in this mini-project. This dataset consists of unprocessed unconstrained ear images of 164 people, with each having ~180 images, for a total of 28,412 ear images. However, in consideration of the computing resources constraint for the unconstrained ear recognition task, only the first 20 classes of the EarVN dataset is used for a total of ~4000 images. Figure 03 shows a sample of the used ear images in this mini-project.

Images from each classes is split into training set with 80% of the total images and testing/validation set with 20% of the total images.

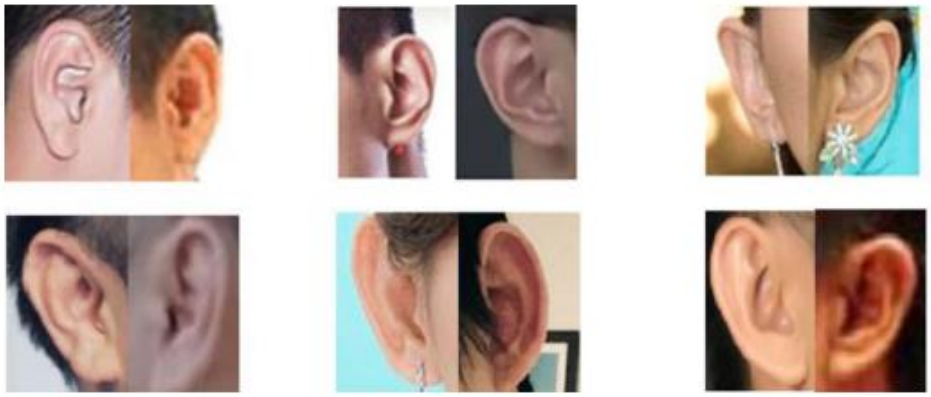

**Figure 4 Sample ear images of EarVN1.0**

### Preprocessing

Preprocessing used in this mini-project includes image resizing to 224 square pixels, horizontal and vertical flipping of an image to 30 degrees, and the standard ImageNet normalization values. These preprocessing steps are applied to the training set while only the resizing and image normalization for testing/validation images. The preprocessed images are both applied on ViT and DeiT.

### ViT and DeiT Implementation

Following the ViT and DeiT description, Both ViT and DeiT for unconstrained ear recognition are implemented with the pre-trained ImageNet-21k model and fine-tuned to classify or recognize 20 people. Vision-Transformer is implemented in this mini-project using Pytorch with XLA on Google Colab TPU while DeiT is implemented with PyTorch on Google Colab GPU. The ViT model is trained on 20 epochs with the following configuration as shown in table 1 while DeiT is trained on 20-40 epochs provided that its setup is machine-efficient (see the attached notebook for codes). Overfitting occurs on models trained beyond 20 epochs on ViT architecture.

**Table 1 ViT Implementation configuration**

| Training Parameter | ViT | DeiT |
|---|---|---|
| Batch size | 8 | 32 |
| Learning Rate | 0.00002 on Adam | 0.001 on Adam |
| Gamma | 0.7 | - |
| Epochs | 20<br>(overfitting occurs beyond this number) | 20 - 50 |

**ViT Performance and Results**

Vision-Transformer on Unconstrained Ear Recognition achieves a validation accuracy of 95.31% with a loss of 26.36%. DeiT on the other hand achieves an accuracy of 88.33% with 1.4% loss on 20 epochs, 93.33% with 1.02% loss on 30 epochs, 96.11% with 0.88% on 40 epochs, and 96.11% with 0.69% loss @ 50 epochs. These outcomes are closely comparable to the CNN-based transfer learning results on the same task [4]. Table 2 shows the comparative results of ViT and DeiT on Unconstrained Ear Recognition and the accuracy results of state-of-the-art CNN architectures on the same task with transfer learning *(from my previous study).* Although it is not included in the objectives, DeiT is more efficient (faster and accurate) than ViT for unconstrained ear recognition and that overfitting does not occur instantly with DeiT.

| Architecture | Validation Accuracy (%) |
|---|---|
| **This mini-project (ViT-Ear)** | **95.31** |
| **This mini-project (DeiT-Ear)** | **88.33 @ 20 epochs,**<br>**93.33 @ 30 epochs**<br>**96.11 @ 40 epochs** |
| ResNet18 (Transfer Learning) | 100.00 |
| AlexNet (Transfer Learning) | 97.30 |
| ResNet50 (Transfer Learning) | 96.70 |
| Inception-ResNet (Transfer Learning) | 94.70 |
| Inception-v3 (Transfer Learning) | 96.70 |

| | |
|---|---|
| GoogLeNet (Transfer Learning) | 93.30 |
| SqueezeNet (Transfer Learning) | 87.30 |
| ShuffleNet (Transfer Learning) | 86.70 |
| MobileNetv2 (Transfer Learning) | 81.30 |

## VI. Conclusion

In this mini-project, Transformer Network is extended to unconstrained ear recognition with Vision-Transformer and Data-efficient image Transformer (DeiT). Both of these transformers achieve an accuracy result that are closely comparable to the state-of-the-art CNN-models' accuracy results on the same task and may serve as an alternative and practical approach to the development of an end-to-end ear recognition system and other classification tasks of computer vision.

## VII. Recommendations

Since this is a mini-study and shortly time-bounded, it is recommended to extend this into an end-to-end system incorporating Detection-Transformer (DeTr) and Vision-Transformer (ViT) as one. Also, it is recommended to use another fabricated unconstrained ear dataset similar to the make of Annotated Web Ears (AWE) of UERC and other available large datasets.

It is also recommended to include inferential codes on both transformer networks.

## VIII. References

[1] Ž. Emeršič, V. Štruc and P. Peer, "Ear Recognition: More Than a Survey," *CoRR,* vol. abs/1611.06203, 2016.

[2] Ž. Emeršič, D. Štepec, V. Štruc and P. Peer, "Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, 2017.

[3] E. Z. e. Al., "The Unconstrained Ear Recognition Challenge 2019," *CoRR,* 2019.

[4] M. Alejo and C. P. Hate, "Unconstrained Ear Recognition through Domain Adaptive Deep Learning Models of Convolutional Neural Network," *International Journal or Recent Technology and Engineering,* vol. 8, no. 2, pp. 3143-3150, 2019.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-End Object Detection with Transformers," *Lecture Notes in Computer Science,* p. 213–229, May 2020.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR,* October 2020.

[7] M. C. Hugo Touvron, M. Douze, F. Massa, A. Sablayrolles and H. Jégou, "Training data-efficient image transformers & distillation through attention," *CoRR,* December 2020.

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